ANALISIS DE CALIDAD DE DATOS DE CLIENTES A PARTIR APRENDIZAJE NO SUPERVISADO EN EL SECTOR DE CONSUMO

OSCAR MARIO GUTIERREZ MONTES

INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO
FACULTAD DE INGENIERIA
ESPECIALIZACIÓN EN BIG DATA
MEDELLÍN
2024

ANALISIS DE CALIDAD DE DATOS DE CLIENTES A PARTIR APRENDIZAJE NO SUPERVISADO EN EL SECTOR DE CONSUMO

OSCAR MARIO GUTIERREZ MONTES

Trabajo de grado para optar al título de Especialista en Big Data

Asesores

Ingrid Natalia Gomez Miranda

Juan Carlos Briñez de León

Rubén Darío Fonnegra Tarazona

Sebastián Camilo Roldan Colorado

INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO

FACULTAD DE INGENIERIA

ESPECIALIZACIÓN EN BIG DATA

MEDELLÍN

2024

Dedicatoria

Dedico este trabajo con todo mi corazón a quienes forman mi hogar, el pilar de mi vida.

A mi esposa, **Johana**, quien, a pesar de no estar completamente de acuerdo con mi decisión de estudiar por el tiempo que pasaríamos separados, tuvo la paciencia y fortaleza para apoyarme hasta el final. Este logro también es suyo, porque su sacrificio fue tan grande como el mío.

A mis hijos, **Sara**, **Sofia**, **Miguel**, y **Mía**, quienes son mi mayor inspiración y la razón de cada esfuerzo que realizo desde el momento en que me levanto. A ellos, les debo una disculpa, porque este año no he estado tan presente como merecen. Sin embargo, con toda seguridad, les digo que este sacrificio también fue por ustedes, y ahora mi tiempo y mi vida les pertenecen por completo.

Agradecimientos

Ante todo, mi agradecimiento eterno a **Dios**, quien cada día me brinda una nueva oportunidad al regalarme el don de la vida y la fortaleza para enfrentar cada desafío.

A mi madre, **Rosa Adela Montes González**, por ser mi mayor inspiración y mi impulso constante. Su amor, sacrificio y ejemplo me motivan a seguir aprendiendo y a trabajar cada día para convertirme en un mejor profesional.

A mis **profesores**, quienes dedicaron su tiempo y esfuerzo para brindarme las herramientas necesarias para afrontar este importante reto académico. Su guía y enseñanza han sido fundamentales en mi formación.

A mi colega de trabajo, **Deyner Elías López**, por mostrar siempre interés en ayudarme y complementarme con sus conocimientos, siendo un apoyo invaluable en este camino.

A mis mejores amigos, **Jorge Andrés Zapata**, **Jaiber González**, y **Cristian Méndez**, quienes siempre me motivaron con sus palabras llenas de confianza, recordándome que soy capaz de lograr cualquier objetivo.

A todos, mi más sincero agradecimiento.

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Lista de anexos

Resumen

ANALISIS DE CALIDAD DE DATOS DE CLIENTES A PARTIR APRENDIZAJE NO SUPERVISADO EN EL SECTOR DE CONSUMO

OSCAR MARIO GUTIERREZ MONTES

Este trabajo presenta un análisis de la calidad de los datos de clientes en una importante empresa del sector de consumo, utilizando un modelo de aprendizaje no supervisado basado en técnicas de clustering. El análisis se desarrolla en cuatro etapas principales, comenzando con la evaluación inicial de la calidad de los datos mediante la herramienta **SAP Information Steward**, que sirve como fuente de datos para la implementación del modelo.

El estudio se centra en un problema crítico para las empresas: la mala calidad de los datos, la cual genera pérdidas económicas significativas, afecta la credibilidad organizacional y aumenta la insatisfacción de los clientes. Con el objetivo de abordar esta problemática, se propone una metodología para diagnosticar la calidad de los datos de clientes, permitiendo identificar deficiencias y establecer las bases para la implementación de un sistema de gestión de calidad de datos.

La metodología presentada combina herramientas estadísticas y modelo de Machine Learning para evaluar y clasificar los datos en función de su calidad. Este enfoque no solo proporciona un diagnóstico claro, sino que también facilita la toma de decisiones estratégicas para optimizar la gestión de los datos y mejorar la efectividad de las operaciones empresariales.

Palabras claves: Calidad de Datos, Dimensiones de Calidad, Datos Maestros Clientes, Clustering

Abstract

QUALITY ANALYSIS OF CUSTOMER DATA FROM UNSUPERVISED LEARNING IN THE CONSUMER SECTOR

OSCAR MARIO GUTIERREZ MONTES

This work presents an analysis of the quality of customer data in an important company in the consumer sector, using an unsupervised learning model based on clustering techniques. The analysis is carried out in four main stages, starting with the initial assessment of data quality using the SAP Information Steward tool, which serves as the data source for the model implementation.

The study focuses on a critical problem for companies: poor data quality, which generates significant economic losses, affects organizational credibility and increases customer dissatisfaction. With the aim of addressing this problem, a methodology is proposed to diagnose the quality of customer data, allowing deficiencies to be identified and the foundations to be established for the implementation of a data quality management system.

The methodology presented combines statistical tools and a Machine Learning model to evaluate and classify data based on its quality. This approach not only provides a clear diagnosis, but also facilitates strategic decision making to optimize data management and improve the effectiveness of business operations.

Keywords: Data Quality, Quality Dimensions, Customer Master Data, Clustering

Glosario

Calidad: Se refiere a la medida en que los datos son adecuados, precisos, completos y confiables para su propósito específico de uso.
Completitud: Indica el grado de información completa en los datos.
Conformidad: cuánto cumple el dato con los formatos o reglas predefinidas, como el cumplimiento de formatos de fechas o estándares específicos.
Clúster: Es una técnica de aprendizaje no supervisado que consiste en agrupar un conjunto de datos en subgrupos o clúster.
Duplicidad: Se refiere a la presencia de registros repetidos o redundantes en una base de datos.
Exactitud: Se refiere a la precisión de los datos en relación con una referencia o estándar.
Integridad: Evalúa si los datos están completos y conectados correctamente en todas las tablas o registros.
KeyMapping: Correcta relación de claves primarias entre los diferentes sistemas.
Kunnr: Identificador único del cliente.
SAP Data Services: Herramienta que hace referencia a los procesos de ETL (extracción, transformación y carga de los datos).
SAP Information Steward: Es una herramienta para analizar la calidad de los datos empresariales
SAP Master Data Governance: Es una solución para la gestión y administración de los datos maestros.

Introducción

La calidad de los datos es uno de los mayores desafíos para las empresas y los responsables de sistemas de información, ya que representa uno de los problemas "ocultos" más persistentes y graves dentro de las organizaciones. La falta de calidad en los datos afecta negativamente la eficiencia y eficacia de los procesos, comprometiendo la capacidad de las empresas para tomar decisiones fundamentadas en información confiable.

En el contexto empresarial, una buena calidad de datos es un activo estratégico invaluable. Los datos son el pilar de la mayoría de las decisiones organizacionales, ya sea a nivel operacional, de dirección o estratégico. En un entorno donde la intuición ya no es suficiente, se vuelve fundamental tomar decisiones basadas en datos confiables y de alta calidad. Sin embargo, cuando los datos son defectuosos, pueden surgir problemas que impactan directamente en la operación y los resultados de la organización.

La calidad de los datos se mide a través de un conjunto de aspectos o atributos, denominados **dimensiones de calidad**, que definen si un conjunto de datos cumple con especificaciones claras y refleja adecuadamente su propósito original. Para evaluar estas dimensiones, se emplean herramientas diagnósticas que permiten identificar el nivel actual de calidad, destacando las deficiencias y proponiendo mejoras específicas.

En este trabajo, se realiza un análisis de calidad de datos enfocado en los **datos maestros de clientes**, utilizando técnicas de **aprendizaje no supervisado**, como el análisis de clústeres. Estas técnicas permiten identificar patrones y grupos dentro de los datos que reflejan diferentes niveles de calidad, proporcionando una visión clara de las áreas más críticas y las oportunidades de mejora.

El presente proyecto aborda estos desafíos mediante el uso de herramientas de análisis de datos y generación de informes automatizados. El objetivo principal es identificar las causas de los errores en los datos de clientes, agrupar a los clientes según su perfil de calidad y determinar las áreas críticas donde se deben implementar mejoras. Adicionalmente, se busca automatizar la generación de informes que faciliten la visualización de resultados, promoviendo así acciones concretas para mejorar la calidad de los datos.

En particular, este estudio se enfoca en las dimensiones de calidad de datos, como **exactitud**, **completitud** y **conformidad**, proporcionando un marco de referencia para optimizar la gestión de la información. Con este enfoque, se busca no solo mejorar la calidad de los datos, sino también potenciar la efectividad de las estrategias comerciales y la satisfacción del cliente, contribuyendo al fortalecimiento de la competitividad de la organización.

1. Planteamiento del problema

1.1 Descripción

En la actualidad, la compañía cuenta con un modelo de gobierno para el manejo de datos maestros. Sin embargo, este modelo no está alineado con las diferentes dimensiones que sustentan la calidad de los datos, lo que limita su eficacia. Aunque existen algunos procesos documentados, estos no abarcan el ciclo de vida completo de los datos maestros de clientes, lo que dificulta garantizar la gobernabilidad, integridad y unicidad de la información.

Además, el seguimiento de estos procesos se realiza principalmente de forma local en cada planta, generando múltiples excepciones específicas para cada una de ellas. Esta descentralización provoca inconsistencias en la gestión de los datos maestros y dificulta la implementación de un enfoque estándar en toda la organización.

Los problemas asociados a esta situación tienen implicaciones significativas, como:

- Toma de decisiones basadas en información incorrecta o incompleta.
- Incremento de los costos operativos debido a la necesidad de rectificar errores.
- Generación de insatisfacción en los clientes por fallas en los servicios.
- Aumento de casos en las mesas de servicio relacionados con problemas de datos.
- Deterioro de la imagen corporativa y pérdida de confianza en la calidad de los datos.

Un factor crítico de esta problemática radica en la **captura de datos**, que constituye el punto de entrada y el origen más frecuente de errores que afectan la calidad. Entre los errores comunes se incluyen:

- Ortografía incorrecta y uso de abreviaturas no estándar.
- Inclusión de caracteres especiales no permitidos.
- Información ingresada en campos incorrectos.
- Datos que exceden el tamaño permitido para un campo.
- Campos vacíos o incompletos.

Estas deficiencias se han intensificado a medida que la compañía expande sus canales de venta y adquiere nuevos clientes. La falta de controles adecuados y estándares en la captura y gestión de datos ha incrementado la frecuencia y gravedad de estos problemas.

Además, se hace necesario **culturizar a los analistas de cada planta** para que adopten y apliquen buenas prácticas en la captura y manejo de datos. Esto implica fomentar su sensibilización sobre la importancia de la calidad de los datos y su impacto en los procesos operativos y estratégicos de la compañía. Sin un cambio cultural y un enfoque disciplinado, será difícil mitigar los problemas existentes y prevenir su recurrencia en el futuro.

1.2 Formulación

El presente trabajo de grado tiene como propósito analizar la calidad de los datos maestros de la organización utilizando técnicas de **aprendizaje no supervisado**, con el fin de identificar patrones, inconsistencias y errores que afectan dimensiones críticas como la gobernabilidad, integridad y unicidad de la información. El enfoque principal es la aplicación de un modelo de **clustering** que permita clasificar y agrupar los datos según sus características específicas, evaluando su nivel de calidad y detectando deficiencias a lo largo de su ciclo de vida.

Para ello, se plantea el uso del algoritmo **K-Means**, que facilitará el análisis y permitirá relacionar los resultados del clustering con dimensiones fundamentales de calidad, tales como **exactitud, consistencia, completitud y validez**. Esto proporcionará un diagnóstico integral y permitirá identificar las áreas críticas que requieren intervención para garantizar la mejora continua de los datos maestros.

Como complemento, el trabajo propone desarrollar un **plan de acción estratégico** basado en los resultados del análisis, enfocado en abordar las deficiencias detectadas. Este plan estará orientado a:

- Mejorar los indicadores de la mesa de servicio.
- Reducir las novedades represadas en el gestor de procesos.
- Custodiar la calidad de los datos para mantener un valor superior al 95%.

Adicionalmente, se plantea la generación de **informes automatizados de calidad**, que servirán como herramientas clave para la toma de decisiones. Estos informes se compartirán con los analistas de planta y otros actores involucrados, fomentando la mejora continua mediante la retroalimentación sobre los resultados obtenidos en el análisis del clustering.

Este enfoque permitirá no solo identificar y mitigar problemas actuales, sino también construir un marco sostenible que asegure la calidad de los datos maestros a futuro.

2. Justificación

La calidad de los datos maestros es un pilar fundamental para el éxito de cualquier organización, ya que afecta directamente la precisión de las decisiones estratégicas, la eficiencia operativa y la satisfacción del cliente.

En este contexto, la implementación de técnicas de **aprendizaje no supervisado**, específicamente el modelo de clustering K-Means, resulta altamente relevante. Este enfoque no solo permite analizar grandes volúmenes de datos de manera eficiente, sino también identificar patrones, inconsistencias y errores que no son evidentes mediante métodos tradicionales.

El trabajo es también relevante porque propone desarrollar un **plan de acción estratégico** basado en los resultados del análisis. Este plan no solo abordará las deficiencias detectadas, sino que estará alineado con los objetivos estratégicos de la organización, como la mejora de los indicadores de la mesa de servicio, la reducción de novedades represadas en el gestor de procesos y el mantenimiento de un nivel de calidad superior al 95%. Estos resultados contribuirán a optimizar la operación, reducir costos y fortalecer la confianza en los datos por parte de los usuarios y clientes.

La generación de **informes automatizados de calidad** proporcionará una herramienta valiosa para la retroalimentación continua de los analistas de planta. Este aspecto es clave para fomentar una cultura de calidad en toda la organización, capacitando a los equipos y sensibilizándolos sobre la importancia de los datos en la operación diaria y la toma de decisiones estratégicas.

En definitiva, este trabajo de grado no solo responde a la necesidad inmediata de mejorar la calidad de los datos maestros, sino que también establece una base metodológica y tecnológica que permitirá a la organización avanzar hacia un manejo más eficiente y sostenible de sus datos en el futuro. La aplicación de estas soluciones tendrá un impacto positivo y duradero en la operación, los procesos estratégicos y la percepción de los clientes, justificando plenamente su realización.

3. Objetivos

3.1 Objetivo general

Realizar un análisis para identificar los problemas y causas de la mala calidad de los datos a partir de un modelo de Clustering y que responda a las siguientes preguntas:

1. ¿Cuáles son los factores que generan deficiencias en la calidad de los datos?
2. ¿Qué patrones existen entre los clientes con baja calidad de datos?
3. ¿Qué variables son críticas para garantizar una alta calidad?

3.2 Objetivos específicos

Los resultados que se logran con este trabajo son:

1. Realizar un análisis descriptivo de las dimensiones de calidad.
2. Desarrollar un modelo de Machine Learning no supervisado aplicando Clustering basados en las dimensiones de calidad.
3. Generar un plan de acción para la mejora de la calidad.
4. Crear informes automatizados con el detalle de la mala calidad, que sean compartidos con los analistas.

4. Marco teórico

4.1 Calidad de los Datos

La calidad de los datos es definida como el grado en que los datos cumplen con los requisitos necesarios para su uso previsto. Diversas dimensiones son esenciales para evaluar la calidad de los datos, entre ellas:

- **Completitud:** Proporción de datos registrados respecto a los datos requeridos.
- **Conformidad:** Cumplen los datos con las reglas definidas para su formato o estructura.
- **Exactitud:** Medida en que los datos reflejan correctamente la realidad.

4.2 Datos Maestros y Gobernanza de Datos

Los datos maestros son las entidades principales dentro de una organización, tales como clientes, productos, proveedores. Se caracterizan por ser información centralizada, reutilizable y compartida a través de diferentes sistemas y procesos. Según Loshin (2020), los datos maestros son la "única fuente de verdad" para las organizaciones, ya que proporcionan una visión unificada y confiable de las entidades clave.

La gobernanza de datos se refiere al conjunto de políticas, procesos y estándares que aseguran la calidad, seguridad y uso ético de los datos dentro de una organización (DAMA-DMBOK, 2017). Este marco establece roles y responsabilidades claras, fomentando la transparencia en la gestión de la información.

4.3 Aprendizaje No Supervisado

El aprendizaje no supervisado es una rama de la inteligencia artificial y del aprendizaje automático que analiza datos sin etiquetas, buscando patrones ocultos o estructuras subyacentes.

- **Clustering:** Una técnica clave dentro del aprendizaje no supervisado, cuyo objetivo es agrupar datos en clusters o grupos según su similitud.
    - **K-Means:** Es uno de los algoritmos más utilizados, conocido por su eficiencia en la segmentación de datos. Divide el conjunto de datos en kkk clusters, minimizando la varianza interna en cada grupo.
    - **DBSCAN:** Clustering basado en densidad, útil para detectar ruido y datos atípicos.
    - **Agglomerative Clustering:** Método jerárquico que permite analizar la estructura de los datos a múltiples niveles de granularidad.

4.4 Aplicación de Modelos de Clustering en Calidad de Datos

El clustering ha demostrado ser una técnica efectiva para abordar problemas de calidad de datos en grandes volúmenes de información. Este método puede:

- Detectar duplicados y datos inconsistentes.
- Identificar patrones de error frecuentes en la captura de datos.
- Agrupar datos según su nivel de calidad, facilitando la priorización de acciones correctivas.

4.5 Herramientas para la Mejora de la Calidad de los Datos

Para garantizar resultados sostenibles en el tiempo, es importante implementar herramientas y procesos complementarios, como:

1. **Tableros de control de calidad de Datos:** Visualización de las evaluaciones de las dimensiones calidad de los datos aplicado reglas de calidad a los campos ya previamente identificados como críticos para la compañía.
2. **Generación de informes automatizados:** Permiten monitorear periódicamente la calidad de los datos y evaluar los avances en su mejora.
3. **Capacitación y sensibilización de los analistas:** Una cultura organizacional enfocada en la calidad de los datos es esencial para minimizar errores y fomentar buenas prácticas.
4. **Planes de acción estratégicos:** Basados en los resultados del análisis, permiten intervenir áreas críticas, reducir costos y mejorar indicadores clave como los casos en mesas de servicio o las novedades represadas.

4.6 Importancia de la Retroalimentación en el Ciclo de Vida de los Datos

La retroalimentación constante es fundamental para asegurar que las mejoras implementadas sean efectivas y sostenibles. La integración de los analistas de planta como actores clave en este proceso garantiza que las soluciones estén alineadas con las necesidades operativas y que los errores se reduzcan progresivamente.

5. Metodología

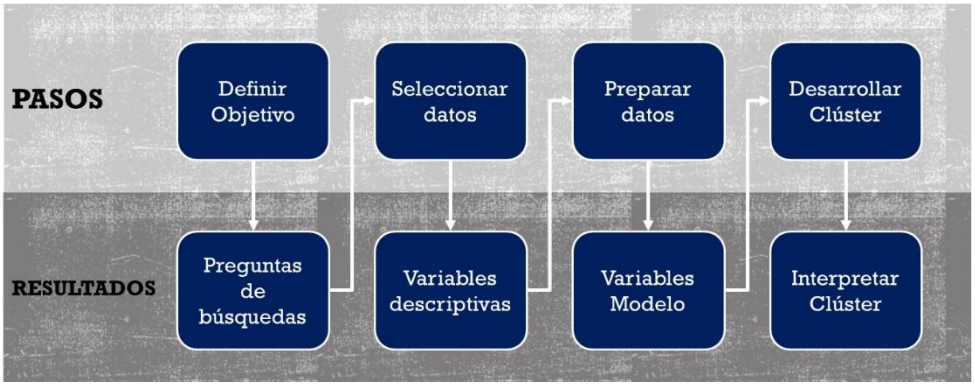
5.1 Tipo de proyecto

Este proyecto se desarrolla en el ámbito de la gestión de calidad de datos y utiliza una técnica de clustering basada en centroides. Su objetivo principal es identificar deficiencias en la calidad de los datos y establecer prioridades para las acciones correctivas, optimizando así los procesos y garantizando una mayor confiabilidad en los sistemas de información.

5.2 Método

La metodología utilizada en este trabajo sigue 4 pasos tal como se muestra en la siguiente figura:

Figura 1. Metodología del proyecto



1. Definir Objetivo

Paso: Identificar claramente el propósito del análisis. En este caso, el objetivo es mejorar la calidad de los datos mediante la identificación de deficiencias y la priorización de acciones correctivas utilizando técnicas de clustering.

Resultado: Se generan **preguntas de búsqueda**, como:

1. ¿Cuáles son los factores que generan deficiencias en la calidad de los datos?
2. ¿Qué patrones existen entre los clientes con baja calidad de datos?
3. ¿Qué variables son críticas para garantizar una alta calidad?

2. Seleccionar Datos

Paso: Elegir las fuentes y variables relevantes para el análisis. En este proyecto, se seleccionaron los Datos Maestros de Clientes en MDG, con un enfoque en clientes activos desde el 2018 y plantas migradas al nuevo sistema.

Resultado: Se identifican las **variables descriptivas**, como:
Dimensiones de calidad: exactitud, completitud, integridad, unicidad y conformidad.
Atributos de clientes: región, fecha de ingreso, sistema origen, entre otros.

3. Preparar Datos

Paso: Realizar tareas de transformación y consolidación de datos. Aquí se incluye la construcción de la tabla **TT_MDG_NIVELACION_SABANA** y el manejo de datos fallidos mediante reglas de calidad.

Resultado: Se generan **variables modelo** listas para el análisis, que incluyen:
Variables estandarizadas o escaladas.
Indicadores de calidad por cliente.

4. Desarrollar Clúster

Paso: Implementar un algoritmo de clustering. En este proyecto, se utilizó KMeans como técnica principal para segmentar clientes con base en la calidad de sus datos.

Resultado: Se identifican grupos de clientes (clústeres) con características comunes, facilitando el análisis de patrones y áreas problemáticas, como:

- Plantas con alta calidad de datos clientes vs. Plantas con datos clientes deficientes.
- Deficiencias específicas en cada clúster que permiten priorizar acciones correctivas.

5.3 Población y muestra

La población de este proyecto se centra en los **Datos Maestros de Clientes**, específicamente aquellos registros utilizados para evaluar la calidad de la información mediante reglas predefinidas. Estas reglas están diseñadas para analizar aspectos críticos en las dimensiones de calidad, tales como **exactitud**, **completitud**, **unicidad**, **integridad**, y **conformidad**, con el objetivo de identificar deficiencias y áreas de mejora.

En el análisis realizado, se evaluaron un total de **286.375 registros** correspondientes a clientes activos. Esta población incluye exclusivamente los datos de las plantas que han sido **migradas al nuevo sistema de datos** dentro del marco del proyecto. Adicionalmente, se consideraron únicamente los clientes activos desde el **año 2018**, garantizando que los datos analizados sean relevantes para los procesos actuales del negocio.

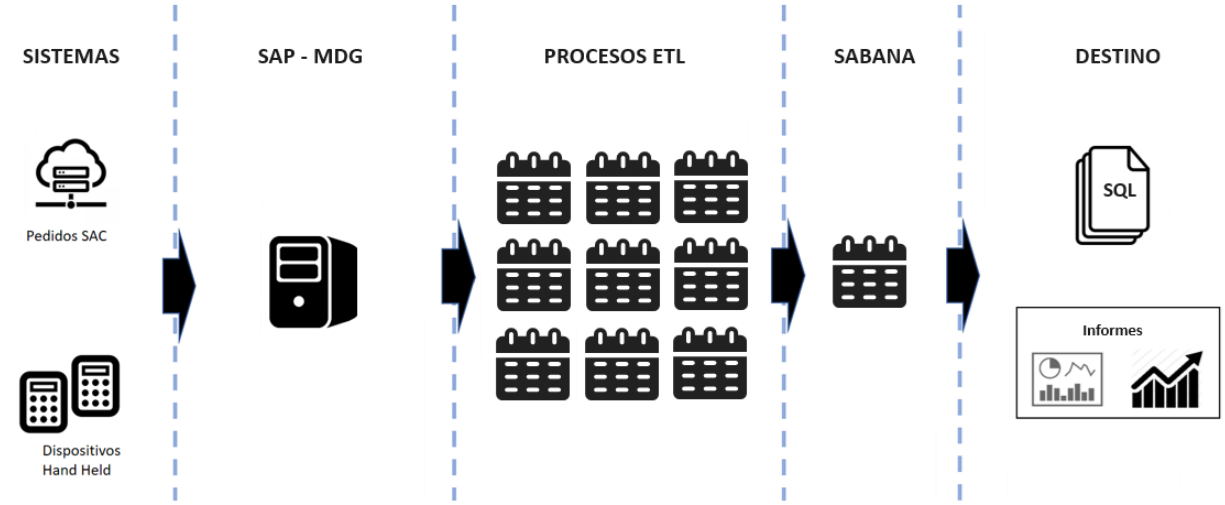
Los registros evaluados se someten a una validación detallada a través de un modelo de calidad que mide el cumplimiento de cada regla en las dimensiones mencionadas. Como resultado, se genera un conteo de datos fallidos, los cuales son almacenados y analizados para identificar patrones, determinar las causas raíz de las inconsistencias y priorizar las acciones correctivas en el sistema.

5.4 Instrumentos de recolección de información

5.4.1 Fuentes primarias.

Las fuentes primarias para este proyecto corresponden a los Datos Maestros de Clientes almacenados en el sistema Master Data Governance (MDG). La creación de una solicitud de cambio (Ingreso, Modificación, Retiro) de Cliente puede ser generada directamente en SAP MDG, así como también puede ser generada por los sistemas PEDIDOS SAC o por un representante de ventas a través de un dispositivo Hand Held. Estos datos son obtenidos directamente del sistema, y representan información cruda esencial para el análisis.

Figura 2. Arquitectura simplificada del proceso



Descripción del Proceso

Extracción:

A través de una herramienta de SAP Data Services (‘ETL’ Extract, Transform, Load), se extraen las 11 tablas que componen el modelo Entidad-Relación de Clientes en el sistema MDG y se llevan a tablas temporales. Estas tablas contienen información clave sobre los clientes, como identificadores, datos demográficos, datos personales, y otros atributos relevantes.

Transformación:

Las tablas extraídas se combinan en un único dataset consolidado denominado **TT_MDG_NIVELACION_SABANA**. Este proceso garantiza que la información esté centralizada y lista para ser utilizada en el análisis.

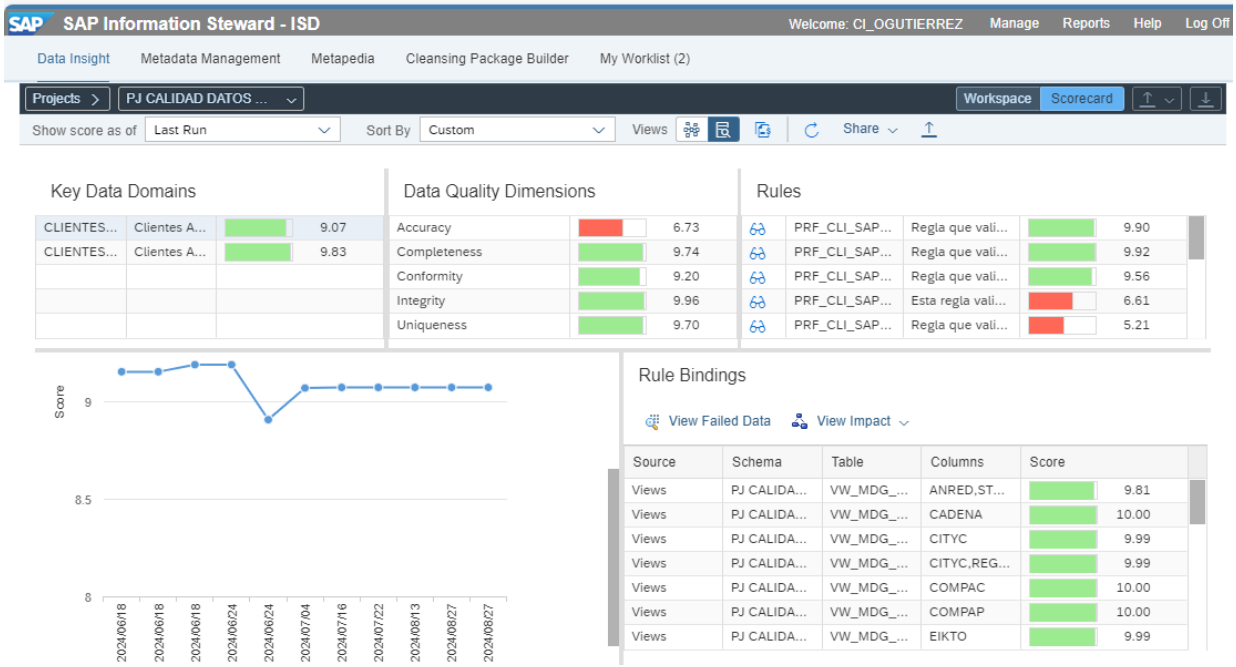
Volumen de Datos:

La tabla **TT_MDG_NIVELACION_SABANA** contiene un total de 1.398.299 registros de clientes, entre ellos, 1.038.253 clientes están clasificados como activos, lo que representa un 74.24% del total de registros.

5.4.2 Fuentes secundarias.

La calidad de los datos se evalúa utilizando la herramienta **SAP Information Steward**, la cual proporciona un tablero de control que facilita la supervisión y análisis de la integridad y confiabilidad de los datos de clientes.

Figura 3. Tablero Calidad datos clientes IS



Este tablero está configurado de la siguiente manera:

1. **Dominios de Clientes**:
    - **Clientes Nuevos**: Evalúa la calidad de los datos ingresados recientemente, garantizando que los registros sean precisos y completos desde el inicio.
    - **Clientes Históricos**: Analiza la calidad de datos almacenados previamente, permitiendo identificar inconsistencias acumuladas y orientar estrategias de limpieza o enriquecimiento.

2. **Dimensiones de Calidad de Datos**:
    - **Exactitud (Accuracy)**: Evalúa la precisión de los datos frente a valores esperados o referencias válidas.
    - **Completitud (Completeness)**: Mide el nivel de llenado de los campos necesarios.

- **Conformidad (Conformity)**: Valida que los datos cumplan con estándares definidos (e.g., formatos o rangos).
- **Integridad (Integrity)**: Verifica que las relaciones entre los datos sean consistentes y completas.
- **Unicidad (Uniqueness)**: Garantiza que no existan duplicados en los registros críticos.

3. **Reglas de Calidad**:
   - Las reglas configuradas permiten validar criterios específicos en los campos identificados como críticos para el negocio, generando puntuaciones individuales para cada una.
   - Cada dominio tiene reglas adaptadas para evaluar las características particulares de los clientes nuevos y los históricos.

4. **Gestión de Datos Fallidos**:
   - Los registros que no cumplen con las reglas de calidad son identificados y almacenados en una tabla detallada.
   - Mediante la herramienta de ETL, estos datos fallidos son extraídos y cargados en una tabla temporal, donde al final, se terminaría cruzando con la TT_MDG_NIVELACION_SABANA.

5. **Visualización y Seguimiento**:
   - El tablero muestra puntuaciones generales para cada dominio de clientes, tendencias temporales y el impacto de las reglas de calidad sobre columnas específicas. Esto ayuda a identificar áreas de mejora y garantizar la alineación con los estándares de calidad.

6. Resultados del proyecto

En esta sección, se listan los pasos que se ejecutaron en la realización del proyecto:

6.1 Fuentes de Información

En la siguiente Tabla, se listan los campos que se seleccionan en la consulta a la fuente de datos.

Tabla 1. *Variables para el análisis descriptivo*

SELECCION CAMPOS	TIPO VARIABLE	DESCRIPCION	VALORES PERMITIDOS
KUNNR	CATEGORICA NOMINAL	Identificador único del cliente	1000007976
SISOGN	CATEGORICA NOMINAL	Sistema de origen por donde se crean los clientes	MDG, GEO, SAC, HH, B2B, etc
FECHA_INGRESO	CATEGORICA ORDINAL	Fecha de ingreso del cliente	16/11/2018
ANIO	CATEGORICA ORDINAL	Año de ingreso del cliente	2018, 2019, ..,2025
CD_PLANTA	CATEGORICA NOMINAL	Código del centro suministrador de productos al cliente	0103, '0215', '0163', etc
NOMBRE_PLANTA	CATEGORICA NOMINAL	Descripción textual de la planta	YUMBO, LUX BOGOT, ITAGUI, etc
REGIONAL	CATEGORICA NOMINAL	Región donde se encuentra la planta	ANTIOQUIA, CENTRO, COSTA, OCCIDENTE, SANTANDER
COMPLETITUD	DISCRETA	Conteo de campos se encuentran llenos	0, 1, 2 hasta 10
EXACTITUD	DISCRETA	Conteo de campos cumplen con los valores permitidos	0, 1 hasta 3
CONFORMIDAD	DISCRETA	Conteo de campos que tienen la forma correcta	0, 1, 2 hasta 5
INTEGRIDAD	DISCRETA	El dato es coherente con otras tablas	0, 1 hasta 2
DUPLICADO	CATEGORICA BINARIA	Existe otro cliente con el mismo número de identificación y dirección	SI, NO
KEYMAPPING	CATEGORICA BINARIA	Tiene definido correctamente sus claves entre los sistemas	SI, NO

Tabla 2. *Variables para aplicar en el modelo*

SELECCION CAMPOS	TIPO VARIABLE	DESCRIPCION	VALORES PERMITIDOS
SISOOGN	NUMERICO CONTINUO	Conversión a númerico de la columna SISOGN	0.9, 0.8, 0.7,.. 0.1
ANIOS_INGRESO	NUMERICO CONTINUO	Años de ingreso del cliente hasta la fecha de ejecución del modelo	5.99, 4.32, 3.05, 2.95
ANTIOQUIA	NUMERICO DISCRETO	Variables dummy para indicar la región especifica	1, 0
CENTRO	NUMERICO DISCRETO	Variables dummy para indicar la región especifica	1, 0
COSTA	NUMERICO DISCRETO	Variables dummy para indicar la región especifica	1, 0
OCCIDENTE	NUMERICO DISCRETO	Variables dummy para indicar la región especifica	1, 0
SANTANDER	NUMERICO DISCRETO	Variables dummy para indicar la región especifica	1, 0
COMPLETITUD	NUMERICO CONTINUO	El conteo se convierte en un score de calidad	1.00, 0.8, 0.33, etc
EXACTITUD	NUMERICO CONTINUO	El conteo se convierte en un score de calidad	1.00, 0.8, 0.33, etc
CONFORMIDAD	NUMERICO CONTINUO	El conteo se convierte en un score de calidad	1.00, 0.8, 0.33, etc
INTEGRIDAD	NUMERICO CONTINUO	El conteo se convierte en un score de calidad	1.00, 0.8, 0.33, etc
DUPLICADO	NUMERICO DISCRETO	Conversión SI en 1 y NO en 0	1, 0
KEYMAPPING	NUMERICO DISCRETO	Conversión SI en 1 y NO en 0	1, 0

6.1.1 Carga de datos

Se inicia con el llamado a las librerías, se crea la conexión con la base de datos Hana Studio, y por último se ejecuta una consulta.

Esta consulta tiene como objetivo consolidar y calcular las métricas de calidad para cada cliente, agrupando errores por dimensiones clave y relacionándolos con el sistema de origen, las plantas y las reglas evaluadas.

Figura 4. Script SQL Hana para consultar los clientes

```
query = """
SELECT
    KUNNR, SISOGN, MDG.ERDAT AS FECHA_INGRESO, SUBSTRING(MDG.ERDAT, 1, 4) AS ANIO, VDM.CODPLA CD_PLANTA,
    REPLACE(REPLACE(REPLACE(UPPER(VDM.NOMBRE), 'POSTOBON ', ''), 'CEDI ', ''), 'GASEOSAS ', '') AS NOMBRE_PLANTA,
    CASE
        WHEN CODPLA IN ('0123', '0112', '0122', '0825', '0826', '0827', '0927', '2858', '0928', '0929', '2038', '0926', '0106', '
        WHEN CODPLA IN ('1027', '0115', '1028', '0126', '1533', '0120', '0127', '1029', '1030', '1534', '1535', '0104', '0101', '
        WHEN CODPLA IN ('1128', '1735', '2971', '2972', '1634', '1637', '1635', '1638', '1636', '1129', '1737', '1738', '1739', '
        WHEN CODPLA IN ('1330', '2240', '0116', '0113', '0114', '0125', '0107', '0121', '1331', '1332', '1836', '0124', '0103', '
        WHEN CODPLA IN ('0423', '0422', '0429', '0428', '1229', '8337', '0224', '0223', '0426', '0430', '0215', '1230', '0227', '
        ELSE 'UNKNOWN'
    END AS REGIONAL,
    COALESCE(SUM(CASE WHEN DIMENSION_NAME = 'Completeness' THEN CONTEO_ERROR END), 0) AS COMPLETITUD,
    COALESCE(SUM(CASE WHEN DIMENSION_NAME = 'Accuracy' THEN CONTEO_ERROR END), 0) AS EXACTITUD,
    COALESCE(SUM(CASE WHEN DIMENSION_NAME = 'Conformity' THEN CONTEO_ERROR END), 0) AS CONFORMIDAD,
    COALESCE(SUM(CASE WHEN DIMENSION_NAME = 'Integrity' THEN CONTEO_ERROR END), 0) AS INTEGRIDAD,
    COALESCE(SUM(CASE WHEN DIMENSION_NAME = 'Uniqueness' THEN CONTEO_ERROR END), 0) AS DUPLICADO,
    KEYMAPPING
FROM (
SELECT MDG.KUNNR, COALESCE(MDG.SISOGN, 'MDG') SISOGN, MDG.VWERK, MDG.ERDAT, RULE.DIMENSION_NAME,
CASE WHEN KEYM.INDKEYMAPPING = 100 THEN 1 ELSE 0 END AS KEYMAPPING, COUNT(DISTINCT FR.RULE_ID) CONTEO_ERROR
FROM STAGE_MDG_H1.TT_MDG_NIVELACION_SABANA MDG
LEFT JOIN STAGE_MDG_H1.TT_MDG_NIVELACION_CLIENTES_KEYMAPPING KEYM ON
MDG.KUNNR = KEYM.KUNNR
LEFT JOIN STAGE_MDG_H1.TT_IS_MDG_CLIENTES_FD FD
ON MDG.KUNNR = FD.KUNNR
LEFT JOIN STAGE_MDG_H1.TT_IS_MDG_CLIENTES_FR FR ON
FD.RUN_ID = FR.RUN_ID AND FD.IS_GEN_ROWID = FR.IS_GEN_ROWID
LEFT JOIN STAGE_MDG_H1.TT_IS_FD_RULES_INFO RULE ON
FR.RUN_ID =RULE.RUN_ID AND FR.RULE_ID = RULE.RULE_ID
WHERE DIMENSION_NAME IS NOT NULL AND MDG.KUNNR IN (
SELECT DISTINCT MDG.KUNNR
FROM STAGE_MDG_H1.TT_MDG_NIVELACION_SABANA MDG
INNER JOIN STAGE_MDG_H1.TT_MDG_NIVELACION_ZBVT027 Z ON
(MDG.VWERK = Z.WERKS )
WHERE MDG.IDRETI = 'A' AND Z.INDDSD = 'X'
AND Z.WERKS IN ('0002','0111','0410','0113','0021','0422',
'0079','0134','0292','0441','0389','0381','0358','0142',
'0090','0145','0382','0118')
)
GROUP BY MDG.KUNNR, COALESCE(MDG.SISOGN, 'MDG'), MDG.VWERK, MDG.ERDAT, RULE.DIMENSION_NAME, KEYM.INDKEYMAPPING) AS MDG
LEFT JOIN STAGE_MDG_H1.TT_AS400_NIVELACION_VDMCENTR VDM ON
VWERK = VDM.CENSUM
--LEFT JOIN STAGE_MDG_H1.TT_MDG_NIVELACION_ZBVT027 Z ON
--VDM.CODPLA = Z.CODPL
WHERE VDM.CODPLA IS NOT NULL
GROUP BY KUNNR, SISOGN, VDM.CODPLA, MDG.ERDAT, VDM.NOMBRE, KEYMAPPING
"""
```

6.1.2 Exploración inicial de los datos

Con el objetivo de conocer los datos, se ejecuta el siguiente código:

Figura 5. Información del DataFrame fuente

```
# Información del DataFrame como Nombre columnas, tipo
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 286375 entries, 0 to 286374
Data columns (total 13 columns):
KUNNR            286375 non-null object
SISOGN           286375 non-null object
FECHA_INGRESO    286375 non-null object
ANIO             286375 non-null object
CD_PLANTA        286375 non-null object
NOMBRE_PLANTA    286375 non-null object
REGIONAL         286375 non-null object
COMPLETITUD      286375 non-null int64
EXACTITUD        286375 non-null int64
CONFORMIDAD      286375 non-null int64
INTEGRIDAD       286375 non-null int64
DUPLICADO        286375 non-null int64
KEYMAPPING       286375 non-null int64
dtypes: int64(6), object(7)
memory usage: 28.4+ MB
```

Figura 6. Descripción estadística básica

```
# Descripción estadística básica
df.describe()
```

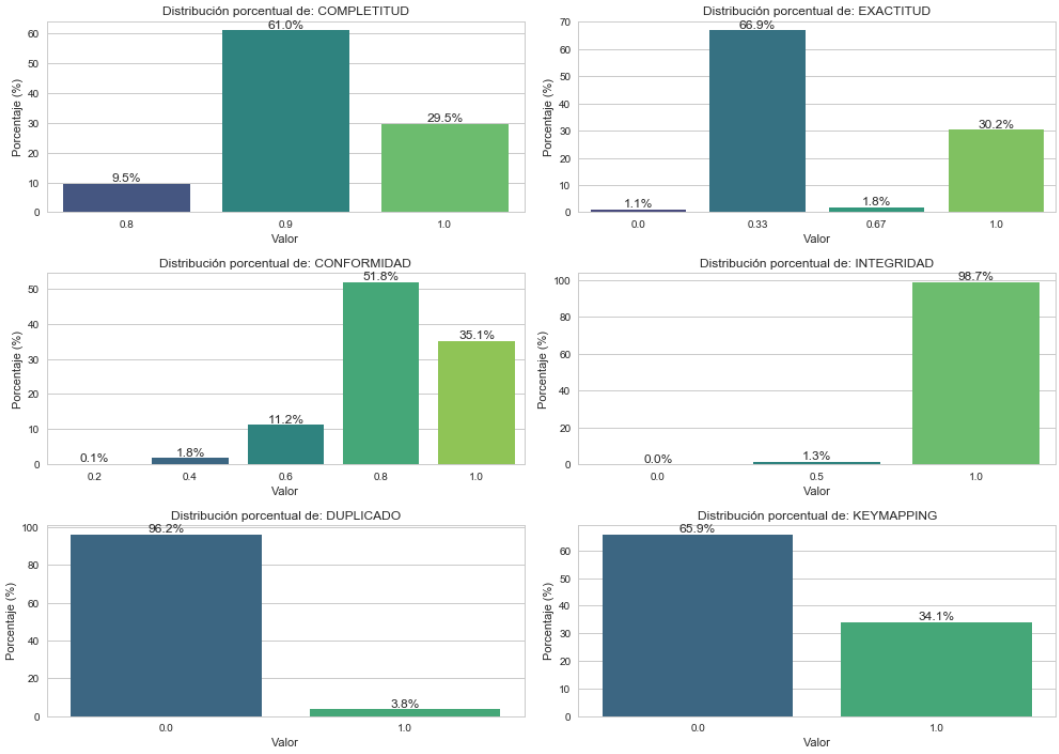
	COMPLETITUD	EXACTITUD	CONFORMIDAD	INTEGRIDAD	DUPLICADO	KEYMAPPING
count	286375.000000	286375.000000	286375.000000	286375.000000	286375.000000	286375.000000
mean	0.799522	1.388169	0.799071	0.012882	0.038338	0.341102
std	0.591730	0.929537	0.707517	0.112919	0.192011	0.474080
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	2.000000	1.000000	0.000000	0.000000	0.000000
75%	1.000000	2.000000	1.000000	0.000000	0.000000	1.000000
max	2.000000	3.000000	4.000000	2.000000	1.000000	1.000000

6.2 Analítica descriptiva

6.2.1 Variables de Calidad

Generar un gráfico para cada dimensión de calidad, con los valores porcentuales. (Con este gráfico, se cumple el objetivo específico 1)

Figura 7. Gráfico dimensiones de Calidad



6.2.2 Transformación de los datos

Este código es parte de la etapa de preparación de datos para el análisis de calidad. Los objetivos principales son:

Estandarización y limpieza de datos: Garantizar que las columnas estén en un formato adecuado para análisis.
Transformación categórica a numérica: Permitir el uso de variables categóricas en modelos.
Normalización de métricas de calidad: Facilitar la comparación y el análisis entre diferentes dimensiones de calidad.

Figura 8. Código transformación de datos

```
# Variable de fecha
today = date.today()

# Convertir la columna de fecha a tipo date
today = pd.to_datetime(date.today())
df['FECHA_INGRESO'] = pd.to_datetime(df['FECHA_INGRESO'])

# Calcular la diferencia de días agregando una nueva columna DIAS_INGRESO
df['DIAS_INGRESO'] = round((((today - df['FECHA_INGRESO']).dt.days) / 365), 2)
```

```
# Convertir a númerico la columna SISOGN
df['SISOGN'] = df['SISOGN'].map({'MDG': 0.9, 'GEO': 0.8, 'SGM': 0.7, 'SAC': 0.6, 'HH': 0.5,
                                 'B2C': 0.4, 'B2B': 0.3, 'CLW': 0.2, 'SAG': 0.1})
```

```
# Pivotear la columna REGIONAL
dummies_regional = pd.get_dummies(df['REGIONAL']).astype(int)

# Unir los dummies al DataFrame
df = pd.concat([df, dummies_regional], axis = 1)
```

```
# El conteo de errores lo convierto en score de calidad
df['COMPLETITUD'] = round(((10 - df['COMPLETITUD'])/10), 2)
df['EXACTITUD'] =   round(((3 - df['EXACTITUD'])/3), 2)
df['CONFORMIDAD'] = round(((5 - df['CONFORMIDAD'])/5), 2)
df['INTEGRIDAD'] =  round(((2 - df['INTEGRIDAD'])/2), 2)
```

6.2.3 Escalamiento

Se le aplica la función MinMaxScaler(), solo a la variable DIA_INGRESO, ya que tiene un valor por fuera del rango 0 y 1.

Figura 9. Código escalamiento

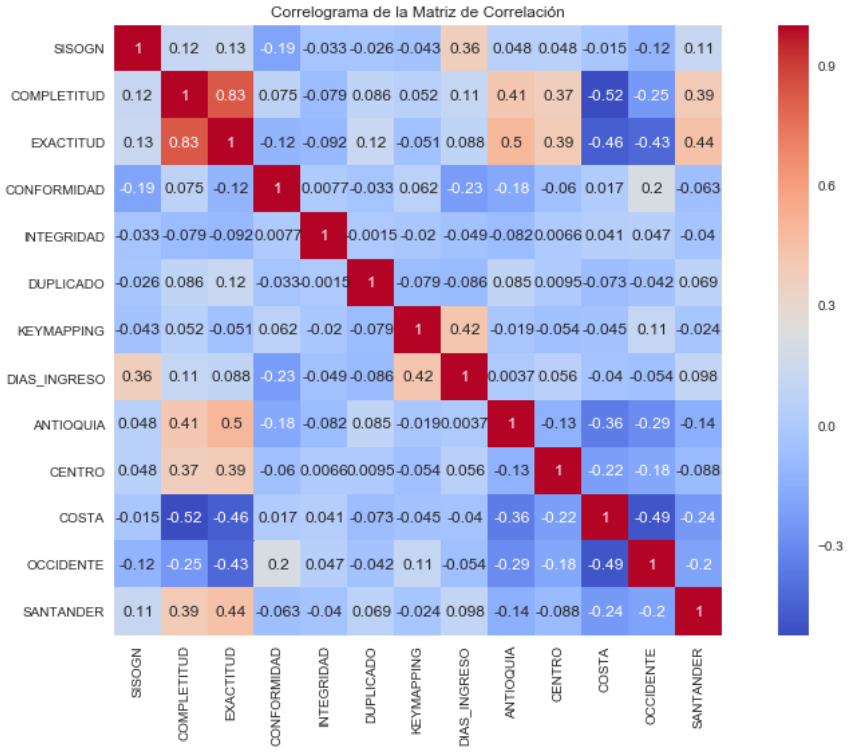
```
# Crear un nuevo DataFrame solo con las columnas numéricas
df_scaler = df[['DIAS_INGRESO' ]].copy()
```

```
scaler = MinMaxScaler().fit(df_scaler.values)

df_scaled = pd.DataFrame(scaler.transform(df_scaler),
                         columns=['DIAS_INGRESO'])

df['DIAS_INGRESO'] = df_scaled['DIAS_INGRESO']
```

6.2.4 Matriz correlación

Para identificar cuáles son las variables más relevantes para el modelo de Clustering, se hace un análisis de correlación, donde se evalúa la importancia de cada variable en relación con las demás.

Figura 10. Matriz de correlación



A partir del correlograma de la matriz de correlación, se puede concluir que:

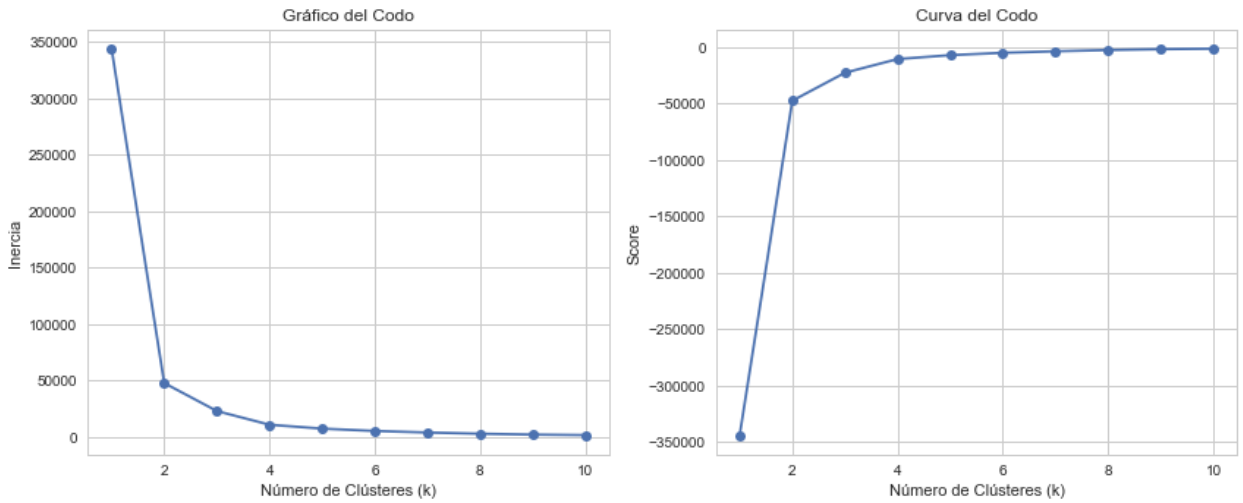
1. Existe una correlación alta (0.83) entre las dimensiones de calidad *COMPLETITUD y EXACTITUD,* lo que puede indicar que estas dimensiones están relacionadas, pero como cada dimensión evalúan campos diferentes, entonces no se eliminan del modelo.
2. La INTEGRIDAD, no tiene correlaciones significativas con otras dimensiones de calidad, esto sugiere que los problemas de integridad no están directamente relacionados con completitud, exactitud o conformidad.
3. DIAS_INGRESO, tiene una correlación moderada con KEYMAPPING (0,42), lo que indica que los registros más antiguos o recientes afectan la coherencia en el mapeo de claves.
4. COSTA y OCCIDENTE son las regiones con los valores de calidad más bajos según la correlación con COMPLETITUD y EXACTITUD.

6.3 Modelamiento

6.3.1 Clúster Óptimos

Como se puede visualizar en ambos gráficos, nos indica un K óptimo con valor de 3 o 4 clústeres.

Figura 11. Gráfico del codo y Curva del codo



Luego de realizar los pasos de preparación de los datos y selección del K-optimo, se procede a desarrollar el modelo de Clúster usando KMeans. (Con este modelo, se cumple el objetivo específico 2)

Figura 12. Código desarrollar Modelo

```
df_scaler = df.select_dtypes(include='number')

# Aplicar K-Means
k_optimo = 3
kmeans = KMeans(n_clusters=k_optimo, random_state=0)
kmeans.fit(df_scaler)

df['Cluster'] = kmeans.labels_
```

```
cluster_labels = kmeans.fit_predict(df_scaler)
```

```
# Contar la distribución de los datos en el campo 'Cluster'
distribucion_cluster = df['Cluster'].value_counts()
```

```
# Mostrar la distribución en la consola
print(distribucion_cluster)
```

```
0    106763
2     97171
1     82441
Name: Cluster, dtype: int64
```

6.3.2 Reducir dimensiones PCA

Para poder representar los clústeres en un gráfico 2D, se hace la reducción de la dimensionalidad usando PCA.

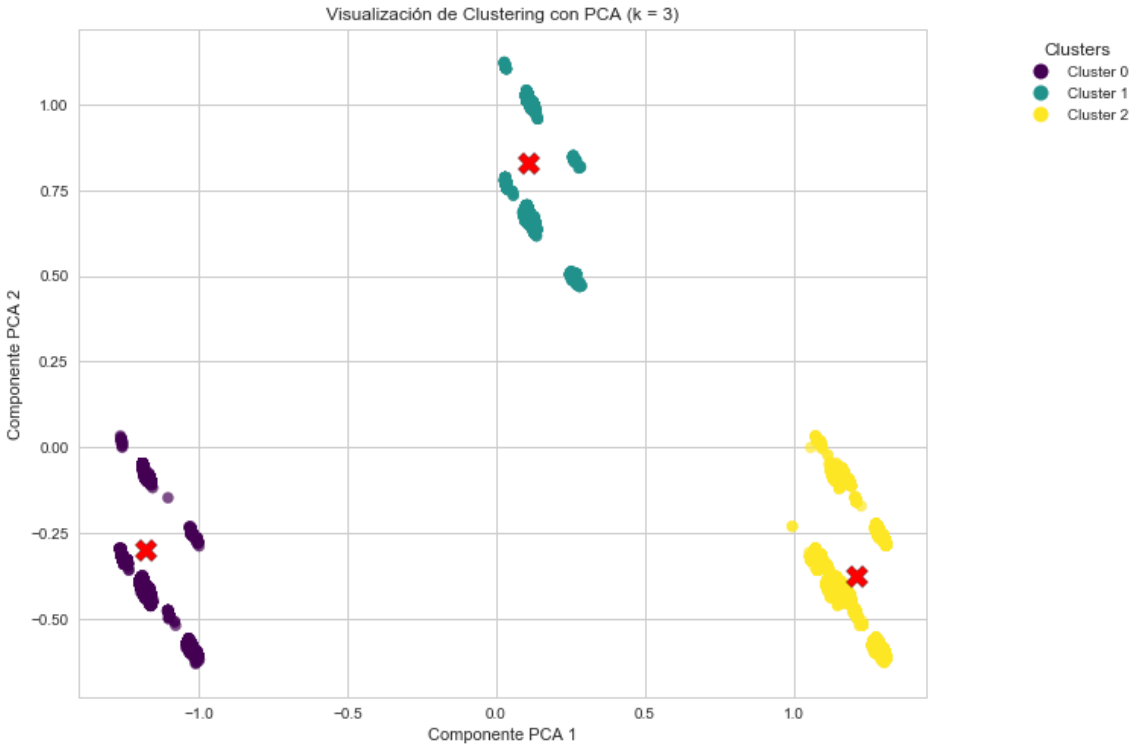
Figura 13. Código Reducir dimensiones PCA

```
# Reducir dimensionalidad para visualización usando PCA (2 componentes para gráfico 2D)
pca = PCA(n_components=2)
df_pca = pca.fit_transform(df_scaler)

# Transformar los centroides al espacio PCA
centroids_pca = pca.transform(centroids)

# Crear el gráfico de clusters
plt.figure(figsize=(10, 8))
scatter = plt.scatter(df_pca[:, 0], df_pca[:, 1], c=cluster_labels, cmap='viridis', s=50, alpha=0.7, label='Datos')

# Agregar los centroides al gráfico
plt.scatter(centroids_pca[:, 0], centroids_pca[:, 1], c='red', marker='X', s=200, edgecolor='black', label='Centroides')

# Añadir barra de color
#plt.colorbar(scatter, label='Etiqueta de Cluster')
plt.xlabel('Componente PCA 1')
plt.ylabel('Componente PCA 2')
plt.title(f'Visualización de Clustering con PCA (k = {k_optimo})')

# Crear una leyenda personalizada para los clusters
unique_clusters = sorted(set(cluster_labels))  # Lista de clusters únicos
colors = [scatter.cmap(scatter.norm(i)) for i in unique_clusters]  # Colores de cada cluster
handles = [plt.Line2D([0], [0], marker='o', color=color, linestyle='None', markersize=10) for color in colors]

# Añadir la leyenda para los clusters
plt.legend(handles, [f'Cluster {i}' for i in unique_clusters], title='Clusters', bbox_to_anchor=(1.1, 1), loc='upper left')
plt.show()
```

6.3.3 Visualizar clustering

La gráfica, resultado del modelo de clustering con KMeans en un espacio de dos componentes principales (PCA), muestra tres clústeres claramente diferenciados (Clúster 0, Clúster 1 y Clúster 2) representados por colores distintos. Los centroides, marcados con "X" rojas, indican el centro promedio de cada grupo en el espacio PCA, mientras que la proximidad de los puntos a estos centroides refleja la homogeneidad dentro de cada clúster.

Figura 14. Visualizar clustering y centroides



La distribución de los clúster formado por los 3 grupos, se encuentra muy uniforme:

Figura 15. Distribución clúster

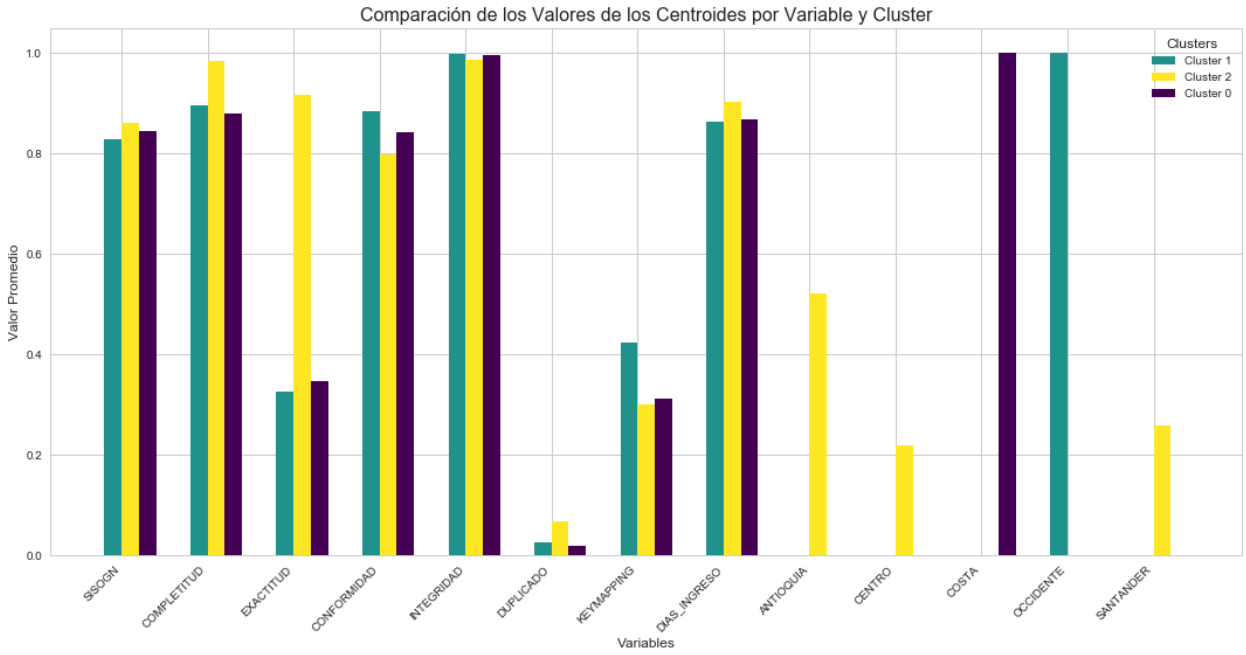
Clúster 0	Clúster 1	Clúster 2
106.763	82.441	97.171

6.4 Visualización resultados

6.4.1 Análisis de centroides

El siguiente gráfico se compara los valores de los centroides para cada variable y clúster.

Figura 16. Análisis de centroides



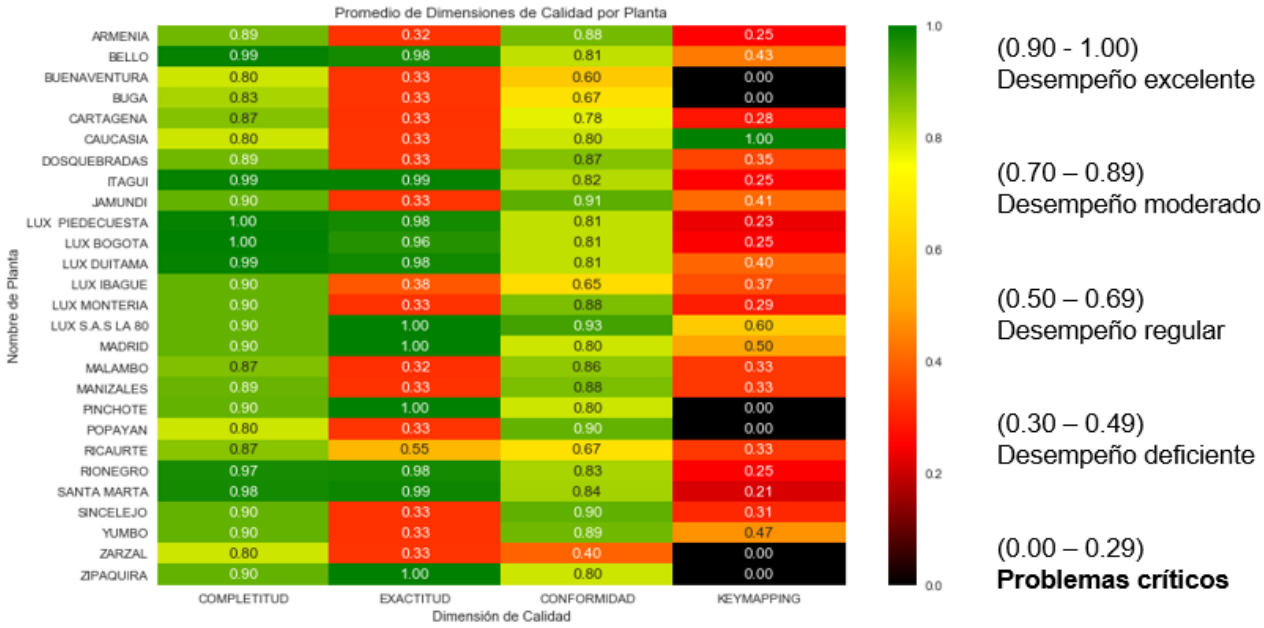
- Este análisis puede usarse para identificar fortalezas y debilidades específicas dentro de cada clúster. Por ejemplo, si se busca mejorar la calidad de datos en un clúster relacionado con una región particular, las métricas del gráfico pueden guiar las estrategias de mejora.

- El gráfico también muestra cómo los clústeres se diferencian claramente en términos regionales, lo que puede ayudar en la segmentación y toma de decisiones estratégicas basadas en las regiones más afectadas.

6.4.2 Análisis Calidad por plantas

El siguiente gráfico representa un mapa de calor, donde se ha definido un esquema de colores que ayudan a identificar las dimensiones débiles y resaltar las plantas con mejores resultados en calidad.

Figura 17. Mapa de calor por planta



Del anterior gráfico se puede concluir:

1. Aprender de las mejores plantas

- **(LUX La 80, Madrid),** reunirnos con los analistas de cada planta y conocer la forma que trabajan y sus buenas prácticas, para replicar esas estrategias en la demás plantas.

2. Monitoreo continuo

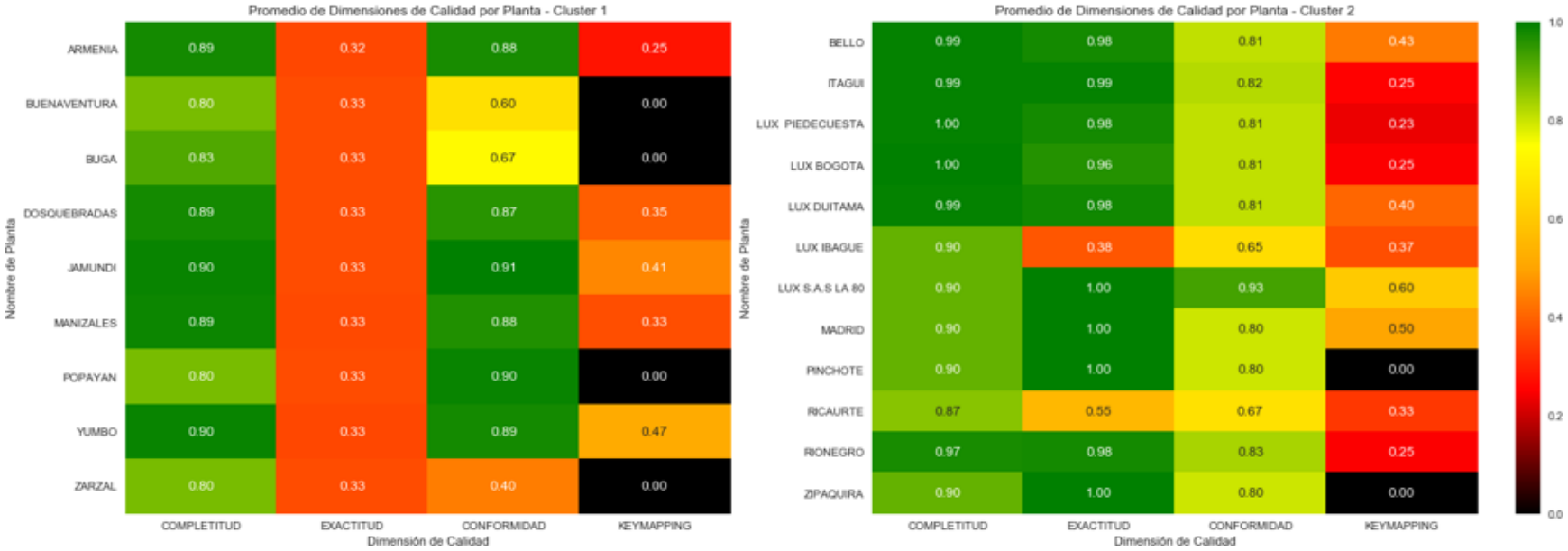
- Implementar revisiones semanales, para asegurar que las plantas se mantengan o mejoren su desempeño en las dimensiones de Calidad

3. Priorizar intervenciones

- **(Buga, Zarzal),** Comenzar por las plantas en rojo y negro para resolver los problemas críticos
- **(Bello, Dosquebradas),** Trabajar con las plantas en amarillo para mejorar las áreas con oportunidades

6.4.3 Comparación Calidad por Clúster

Figura 18. Comparación Calidad por Clúster



En este análisis, comparamos dos gráficos que representan el desempeño en calidad de datos. Identificamos que el clúster con mayores oportunidades de mejora muestra una tendencia generalizada en todas las plantas hacia un bajo desempeño en la dimensión de *exactitud*, así como deficiencias específicas en *conformidad* para algunas plantas particulares.

Por otro lado, el clúster 2, aunque presenta un desempeño general más alto en calidad, destaca áreas de intervención específica en plantas como *Lux Ibagué* y *Ricaurte*, particularmente en las dimensiones de *exactitud* y *conformidad*. Estas observaciones subrayan la necesidad de implementar estrategias enfocadas en estas áreas críticas para acercarnos al objetivo de alcanzar un 100% de calidad óptima.

6.5 Plan de Acción

Este plan de acción permitirá no solo mejorar la calidad de los datos en las plantas identificadas, sino también establecer una cultura organizacional orientada a la excelencia en la gestión de la información.

Figura 19. Plan de Acción

Objetivos estratégicos del plan		Indicador
	01 1. Alcanzar un 90% de cumplimiento en la mesa de Servicio (Actual: 67,7%)	Porcentaje de Cumplimiento de la Mesa de Servicio
	01 2. Asegurar que los tickets abiertos en mesa de Servicio no superen los 15 diarios	Promedio diario de tickets abiertos
	02 3. Disminuir la cantidad de novedades represadas en el gestor de proceso 6.336	Cantidad de novedades represadas en el gestor de proceso
	02 4. Mantener estable el gestor de procesos, donde la cantidad denovedades no supere las 1.000	Cantidad de novedades represadas en el gestor de proceso
	03 5. Custodiar la evaluación de calidad de clientes en un valor superior a los 95%	Porcentaje de Cumplimiento de calidad por cada dominio

Nro	Acciones propuestas	Diciembre				Enero				Febrero				Marzo				Abril				Mayo				Junio				Julio			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
01	Mesa de Servicio																																
1	Revisión de solicitudes e incidentes en mesa de servicio																																
1.1	Asignar de casos Nivel 1, Nivel 2, Equipo																																
1.2	Priorizar los casos																																
1.3	Reunión semanal para evaluar resultados																																
1.4	Registrar los tickets vencidos asignados al grupo																																
2	Establecer normar generales																																
3	Diseñar un sistema de incentivos o penalidades por cumplimientos semanales																																
4	Analizar las causas principales de apertura de tickets, para idear soluciones preventivas																																
5	Diseñar elementos de comunicación para compartir con usuarios de la mesa (Comunicados internos)																																
6	Crear una base de conocimientos																																
7	Generar un informe de resultados																																
02	Gestor de Procesos																																
1	Revisión de gestor de procesos clientes																																
1.1	Monitorear novedades con clientes nuevos																																
1.2	Crear un base de analistas activos por cada planta																																
1.2	Retroalimentación temprana a los analistas de Planta usando imágenes "TAL CUAL"																																
2	Analizar patrones en las novedades para automatizar soluciones con reglas derivadas o cargas masivas																																
3	Implementar reuniones semanales de seguimiento																																
4	Definir responsables para cada tipo de novedad																																
5	Crear un infome para revisar el avance mensual																																
6	Crear bitacora de errores gestor																																
03	Tablero de Calidad en IS																																
1	Revisión del tablero de calidad en Information Steward																																
2	Implementar nuevas reglas de calidad de Datos al tablero en IS																																
3	Activar la funcionalidad de envio de correo cuando el indicador de Calidad baje del 95%																																

Con base en los resultados obtenidos en el modelo, es imprescindible implementar un plan de acción dirigido a las plantas que registraron valores bajos en calidad. Este plan incluiría las siguientes iniciativas, entre otras:

1. **Validaciones diarias**: Realizar revisiones diarias de las nuevas solicitudes para identificar y corregir errores de forma inmediata, complementadas con sesiones de retroalimentación para los analistas responsables.

2. **Informes mensuales de calidad**: Generar y distribuir reportes detallados que destaquen indicadores clave de calidad, identificando tendencias y oportunidades de mejora a lo largo del tiempo.

3. **Presentación de resultados en comités directivos**: Incorporar el análisis de calidad de datos en los comités de dirección para fomentar el compromiso organizacional y definir estrategias conjuntas de mejora.

6.6 Informes Automatizados

Finalmente, el trabajo contempla la creación de informes automatizados en la herramienta **Jupyter Notebooks**, con un enfoque en garantizar la **inmediatez en el acceso a los datos** y facilitar la **comunicación de resultados** con los analistas. Estos informes permiten analizar de manera eficiente los indicadores clave de calidad de datos y priorizar acciones de mejora en tiempo real.

A continuación, se listan los informes desarrollados en el marco de este trabajo:

1. Check List de Clientes DSD
2. Perfilamiento Coordenadas
3. Informe Clientes Duplicados
4. Informe Clientes Segmentación
5. Informe Mesa Servicio
6. Informe del Gestor
7. Informe completo de Calidad
8. Informe puntual por dimensión y planta
9. Informe KeyMapping
10. Informe Clientes desnivelados
11. Informe de IS con Datos Fallidos

Estos informes automatizados no solo reducen el tiempo dedicado a la preparación manual de reportes, sino que también aseguran la consistencia y confiabilidad de los datos, mejorando la toma de decisiones y optimizando los procesos de gestión de calidad.

7. Conclusiones

El modelo aplicado corresponde a una versión inicial, sujeta a cambios, de acuerdo con una discusión posterior con el equipo de trabajo de la empresa. Este proceso permitirá validar su pertinencia y así desarrollar una solución más adecuada para su futura adopción e implementación. Además, será necesario complementar el tablero de calidad incorporando las reglas de validación pendientes por evaluar.

El modelo Kmeans logró responder a las preguntas planteadas en el objetivo, donde a partir de los resultados obtenidos, se identifica que las deficiencias en la calidad de los datos están relacionadas con:

1. **Falta de validaciones previas en migraciones al nuevo sistema:**
   Las migraciones realizadas carecieron de validaciones rigurosas para garantizar la calidad de los datos antes de su migración al nuevo sistema.
2. **Baja calidad de datos en ciertas plantas:**
   A pesar de utilizar las mismas aplicaciones para la captura de datos, existe una disparidad en la calidad, lo que sugiere la necesidad de reforzar la capacitación y culturizar a los analistas en dichas plantas.

La mala calidad de los datos es inevitable, pero se pueden implementar estrategias para reducirla, prevenirla y casi eliminarla. Aplicando un plan estratégico con acciones asignadas a responsables, y monitoreando indicadores periódicamente, se podrán cumplir las metas y corregir desviaciones rápidamente. Es fundamental incorporar retroalimentación continua, capacitar en buenas prácticas de manejo de datos, automatizar procesos y fomentar una cultura de mejora constante. Así, no solo se logran los objetivos establecidos, sino que se construye una base sólida de datos confiables con impacto positivo y duradero en la organización.

8. Recomendaciones

Para complementar este trabajo, sería útil realizar pruebas con otros modelos, como **DBSCAN** o **Agglomerative Clustering**, lo que permitiría comparar sus resultados con los obtenidos mediante el modelo actual. Esto proporcionaría un análisis más completo y podría identificar enfoques alternativos o complementarios.

Además, se recomienda incorporar al modelo un nivel de detalle más granular, que permita analizar específicamente los campos que están contribuyendo a la mala calidad de los datos. Esto facilitaría la identificación de problemas críticos y ayudaría a implementar soluciones más efectivas para mejorar la calidad de los datos.

9. Referencias bibliográficas

1. Factorydea consultoría. “Redacción del plan de acción y ordenación de las acciones en función de su prioridad”. https://factorydea.es/redaccion-plan-accion-ordenacion-acciones-funcion-prioridad/

10. Bibliografía

1. Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia, (Octubre 2019). “Modelo de Uso Calidad de Datos.,”. https://herramientas.datos.gov.co/sites/default/files/Modelo_de_Uso_de_Calidad_de_Datos.pdf
2. A. M. Rangel-Carrillo, G. P. Maestre-Góngora, M. A. Osorio-Sanabria. (2020). Principios, lineamientos, dimensiones y atributos para la evaluación de calidad de Datos Abiertos de Gobierno. Revista de investigación, administración e ingeniería, vol. 8. Pp 54-65.
3. Blog de Datos y Análisis. (2022). Cómo mejorar la calidad de tus datos empresariales. www.datablog.com.