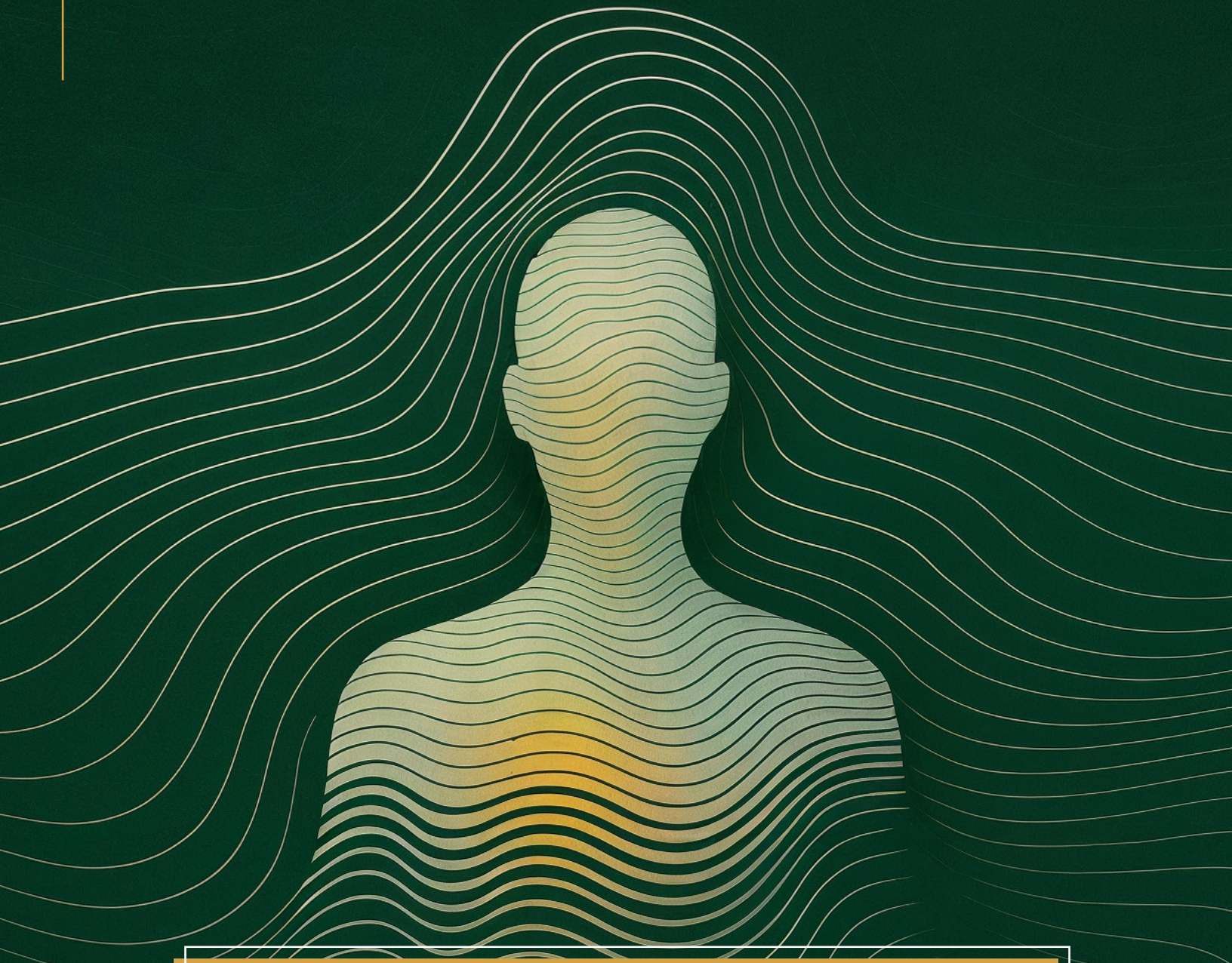
SERVICIO GEOLÓGICO COLOMBIANO



Guía de anonimización de datos personales

a través de software libre para el Servicio Geológico Colombiano

Guía que aborda de manera integral las principales técnicas de anonimización de datos personales, orientadas a su aplicación práctica mediante software libre (KNIME).

Por:

Juan David Acuña García

Jaime Alberto Garzón Barrios

SERVICIO
GEOLÓGICO
COLOMBIANO

Citación: Acuña García, J.D. y Garzón Barrios, J. A. (2025). *Guía de anonimización de datos abiertos de datos personales - a través de software libre para el Servicio Geológico Colombiano* (2025). Bogotá, D. C.: Servicio Geológico Colombiano.

Para acceder a la versión digital de esta colección usar el siguiente enlace web

Agradecimientos

Los autores expresan su sincero agradecimiento al Servicio Geológico Colombiano, cuyo apoyo institucional y disposición permanente hicieron posible el desarrollo de esta guía. En espeical, se reconoce a la Dirección de Gestión de la Información, por su orientación estratégica y por promover la divulgación de este material en coherencia con las mejores prácticas en el tratamiento y protección de datos personales, contribuyendo al fortalecimiento de una cultura institucional orientada a la gestión responsable de la información.

De igual manera, se agradece al equipo de oficialización, liderado por Luis Eduardo Salas Vargas, por su compromiso y rigurosidad en los procesos de validación de la información. Se extiende también un reconocimiento al diseñador Andrés Camilo Romero, por su valioso aporte creativo en la construcción de la propuesta visual de la obra; al ingeniero Ricardo Mancera y a la ingeniera Andrea Neira, por su contribución en la conceptualización de los términos desarrollados; y al ingeniero Samuel Fernando Mesa, por la cuidadosa diagramación de este documento mediante el uso de herramientas de software libre, garantizando tanto la calidad editorial como la alineación con principios tecnológicos abiertos y sostenibles.

Asimismo, se reconoce de manera especial al ingeniero Wildher Sarmiento, quien, a través de la gestión de recursos orientados a la analítica de datos, se logró el desarrollo y fortalecimiento de los componentes técnicos que soportan esta guía, aportando significativamente a la implementación de prácticas modernas y eficientes para el tratamiento seguro de la información.

Finalmente, se reconoce y agradece la colaboración de todas las personas que, de manera directa o indirecta, aportaron su tiempo, conocimiento y esfuerzo para la materialización de esta Guía de anonimización de datos personales, reafirmando el compromiso institucional con la protección de la información, el cumplimiento normativo y la adopción de buenas prácticas en el tratamiento de datos.

Índice de contenidos

Índice de figuras

1. Introducción

El Servicio Geológico Colombiano (SGC), como entidad científica del Estado, tiene entre sus funciones misionales la construcción, administración y custodia de datos estratégicos para el país. En un entorno global donde la protección de la información personal y la privacidad adquieren una relevancia cada vez mayor, resulta indispensable que las organizaciones públicas adopten buenas prácticas para el tratamiento responsable de los datos.

En este contexto, y considerando el volumen creciente de información que se genera y analiza en la entidad, se hace necesario establecer métodos claros y estandarizados de anonimización que permitan preservar la utilidad de los datos para fines científicos, técnicos o administrativos, sin comprometer la identidad ni los derechos de las personas.

La presente Guía de anonimización de datos tiene como propósito divulgar el conocimiento técnico y práctico sobre las principales técnicas de anonimización aplicadas en la herramienta KNIME, alineadas con las políticas del Gobierno Nacional en materia de protección de datos personales. Este documento está dirigido a todos los colaboradores del SGC que realicen actividades de tratamiento o análisis de datos personales o sensibles, promoviendo así una cultura institucional de protección de la información y cumplimiento normativo.

En este contexto, y de acuerdo con la Ley 1581 de 2012, se entiende por dato personal toda información asociada o que pueda asociarse a una persona natural identificada o identificable.

Dentro de esta categoría, los datos sensibles constituyen un tipo especial de dato personal que afecta la intimidad del titular o cuyo uso indebido puede generar discriminación, como aquellos que revelan el origen racial o étnico, la orientación política, las convicciones religiosas o filosóficas, la pertenencia a sindicatos, los datos de salud, la vida sexual o los datos biométricos. Por su nivel de riesgo, estos datos requieren medidas reforzadas de protección y un tratamiento más restrictivo, aspecto que esta guía busca abordar mediante la aplicación de técnicas adecuadas de anonimización.

2. Resumen

El documento establece las pautas y lineamientos del Servicio Geológico Colombiano (SGC) para la anonimización de datos personales y sensibles, en cumplimiento de la Ley 1581 de 2012 y su normativa complementaria. Su propósito es orientar la aplicación de técnicas que permitan conservar la utilidad analítica y científica de la información, garantizando al mismo tiempo la protección de la privacidad y la confidencialidad de los titulares.

La guía aborda el marco legal e institucional que sustenta el tratamiento responsable de los datos, describe el alcance organizacional de la política, y desarrolla un modelo metodológico compuesto por siete fases: definición del equipo de trabajo, determinación del objetivo del tratamiento, análisis de riesgos de reidentificación, pre-anonimización, aplicación de técnicas, documentación y publicación, y auditoría con mejora continua.

Incluye un glosario técnico, una matriz RACI de responsabilidades y un caso práctico en herramienta open-source (KNIME) para la implementación operativa de las técnicas. Estas técnicas comprenden métodos de aleatorización, generalización, seudonimización, k-anonimato, l- diversidad y privacidad diferencial, entre otros.

La guía consolida el compromiso del SGC con la responsabilidad demostrada, la minimización del dato y la trazabilidad institucional, promoviendo una cultura organizacional de cumplimiento normativo, seguridad de la información y respeto por los derechos fundamentales asociados al hábeas data.

Asimismo, busca extender y divulgar el conocimiento sobre anonimización de datos mediante el uso de una herramienta práctica y de fácil adopción, que facilite la comprensión y aplicación de los principios y técnicas presentados, fomentando así la capacitación interna y el fortalecimiento de las capacidades institucionales en materia de protección de datos y análisis responsable de la información.

3. Abstract

This document establishes the guidelines and principles of the Servicio Geológico Colombiano (SGC) for the anonymization of personal and sensitive data, in compliance with Law 1581 of 2012 and its complementary regulations. Its purpose is to guide the application of techniques that preserve the analytical and scientific utility of information while ensuring the protection of individuals' privacy and the confidentiality of data subjects.

The guide addresses the legal and institutional framework that underpins the responsible processing of data, describes the organizational scope of the policy, and develops a methodological model composed of seven phases: definition of the work team, determination of

the processing objective, re-identification risk analysis, pre-anonymization, application of techniques, documentation and publication, and continuous improvement through auditing. It includes a technical glossary, a RACI responsibility matrix, and a practical case implemented with an open-source tool (KNIME) to demonstrate the operational application of the techniques. These techniques include methods such as randomization, generalization, pseudonymization, k- anonymity, l-diversity, and differential privacy, among others.

The guide reinforces the SGC proven responsability, data minimization, and institutional traceability, promoting an organizational culture of regulatory compliance, information security, and respect for the fundamental rights associated with habeas data. Furthermore, it seeks to extend and disseminate knowledge on data anonymization through a practical and user-friendly tool that facilitates understanding and application of the principles and techniques presented, thereby fostering internal training and strengthening institutional capacities in data protection and responsible information analysis.

4. Glosario

Aleatorización:

Técnica que modifica ligeramente los datos originales introduciendo ruido o variaciones aleatorias para impedir coincidencias exactas sin afectar gravemente los análisis globales.

Anonimización:

Proceso mediante el cual los datos personales son transformados de forma irreversible, de modo que no puedan asociarse con una persona natural identificada o identificable, ni siquiera mediante el cruce con otras fuentes de información.

Análisis de riesgos de reidentificación:

Etapa del proceso de anonimización que evalúa la probabilidad e impacto de que un individuo pueda ser identificado nuevamente a partir de datos anonimizados. Permite establecer medidas de mitigación y definir un umbral aceptable de riesgo residual.

Datos personales:

Toda información asociada o que pueda asociarse a una persona natural identificada o identificable, conforme a la Ley 1581 de 2012.

Datos sensibles:

Tipo especial de dato personal que afecta la intimidad del titular o cuyo uso indebido puede generar discriminación. Incluye información sobre origen racial o étnico, orientación política, creencias religiosas o filosóficas, pertenencia a sindicatos, salud, vida sexual y datos biométricos.

Datos estructurados:

Conjuntos de datos organizados en formatos tabulares o bases de datos que permiten su tratamiento automatizado y sistemático, facilitando la aplicación de técnicas de anonimización.

Delegado/a de Protección de Datos (DPD/DPO):

Funcionario encargado de supervisar el cumplimiento legal en materia de protección de datos personales, emitir observaciones sobre riesgos residuales y mantener la trazabilidad documental del proceso.

Equipo técnico de anonimización:

Grupo responsable de ejecutar el análisis de variables, realizar pruebas previas y aplicar las técnicas de anonimización definidas en el plan de tratamiento.

Generalización:

Método que reduce el nivel de detalle de los datos (por ejemplo, reemplazar edades exactas por rangos) para evitar la identificación puntual de personas.

k-anonimato (k-anonymity) :

Técnica de anonimización que garantiza que cada registro en un conjunto de datos sea indistinguible de al menos k–1 registros adicionales, reduciendo el riesgo de reidentificación.

L-diversidad (L-diversity) :

Es una extensión del k-anonimato que garantiza que haya suficiente Variación en un atributo sensible

Low Code :

Enfoque de desarrollo que permite crear aplicaciones con muy poca programación manual, usando principalmente interfaces visuales, componentes preconstruidos y flujos configurables.

Ley 1581 de 2012:

Norma que regula la protección de datos personales en Colombia y establece los principios, derechos y obligaciones relacionados con su tratamiento.

Marco normativo de protección de datos personales:

Conjunto de disposiciones legales y reglamentarias que regulan el tratamiento de datos personales en Colombia, incluyendo la Ley 1581 de 2012, el Decreto 1377 de 2013, el Decreto 1081 de 2015 y las directrices de la Superintendencia de Industria y Comercio.

Minimización del dato:

Principio según el cual solo deben recolectarse, tratarse y conservarse los datos estrictamente necesarios para cumplir la finalidad del tratamiento, reduciendo la exposición y el riesgo de reidentificación.

Privacidad diferencial:

Método matemático de anonimización que introduce ruido controlado en los datos o resultados agregados, de manera que sea imposible inferir información específica sobre un individuo, incluso con acceso a múltiples fuentes.

Riesgo residual:

Nivel de riesgo que permanece luego de aplicar las medidas de anonimización y seguridad, considerado aceptable de acuerdo con los criterios institucionales de tolerancia al riesgo. Seudonimización: Técnica que sustituye identificadores personales por valores codificados o aleatorios, preservando la posibilidad de reversión bajo condiciones controladas. A diferencia de la anonimización, los datos seudonimizados

aún se consideran datos personales. Supresión: Eliminación total o parcial de datos personales o identificadores, utilizada como técnica de anonimización o en procesos de eliminación irreversible de información.

Seudonimización:

Proceso en el que los identificadores directos se reemplazan por códigos o seudónimos, manteniendo la posibilidad de revertir el proceso mediante una clave separada.

T-Cercania (T-closeness)

Técnica que obliga a que la distribución del atributo sensible en cada grupo anonimizado sea similar (a una distancia ≤ t) a la distribución global del conjunto de datos.

Técnicas de anonimización:

Conjunto de métodos aplicados para reducir el riesgo de identificación de personas en un conjunto de datos. Incluyen la aleatorización, generalización, seudonimización, k-anonimato, l-diversidad, t-closeness, entre otras. Tratamiento de datos personales: cualquier operación o conjunto de operaciones sobre datos personales, como la recolección, almacenamiento, uso, circulación o supresión, conforme al artículo 3 de la Ley 1581 de 2012.

Responsabilidad demostrada:

Principio que obliga a las entidades responsables del tratamiento a demostrar de manera verificable el cumplimiento de las normas de protección de datos personales, incluyendo la implementación de medidas técnicas, organizativas y legales.

Reidentificación:

Proceso mediante el cual se intenta asociar nuevamente información anonimizada con una persona específica. Su prevención es el objetivo central de la anonimización. Riesgos de privacidad: amenazas o vulnerabilidades que pueden permitir la identificación, vinculación o inferencia indebida de información sobre una persona a partir de los datos tratados.

Trazabilidad:

Capacidad de registrar y verificar todas las acciones, decisiones y transformaciones realizadas sobre los datos a lo largo del proceso de anonimización, garantizando transparencia y control.

Tokenización:

Técnica que reemplaza datos sensibles por identificadores aleatorios o “tokens”, vinculados mediante una tabla segura que permite su reversión controlada.

Utilidad de los datos:

Grado en que los datos anonimizados conservan su valor analítico o científico tras el proceso de anonimización. Se busca mantener un equilibrio entre privacidad y utilidad.

5. Objetivo

Establecer las pautas y lineamientos para la aplicación de técnicas de anonimización de datos personales y datos sensibles tratados por la entidad, promoviendo la protección de la información conforme al marco normativo vigente y la divulgación del conocimiento técnico entre los colaboradores mediante una guía práctica para la implementación de procesos de anonimización de datos estructurados.

En este sentido, la aplicación de técnicas de anonimización debe contemplar tanto los datos personales en general como, de manera diferenciada, los datos personales sensibles o de naturaleza confidencial tratados por la entidad, garantizando medidas reforzadas de seguridad, control de acceso y minimización del riesgo de reidentificación, en cumplimiento de los principios de legalidad, proporcionalidad y responsabilidad demostrada.

6. Marco legal e institucional

El tratamiento y la protección de los datos personales en Colombia se rigen por un conjunto de normas que garantizan el derecho fundamental al hábeas data y establecen los principios para un manejo responsable de la información.

De acuerdo con la Ley 1581 de 2012, su Decreto Reglamentario 1377 de 2013 y las directrices de la Superintendencia de Industria y Comercio (SIC), toda actividad que involucre datos personales debe realizarse bajo criterios de legalidad, finalidad, libertad, veracidad, seguridad y confidencialidad.

La anonimización se reconoce como una medida técnica y organizativa que permite eliminar la posibilidad de identificar directa o indirectamente a los titulares, convirtiéndose en una herramienta esencial para el cumplimiento del principio de minimización del dato y del deber de protección integral de la información. Una vez que la información ha sido debidamente anonimizada de manera que no sea posible su reidentificación, deja de considerarse “dato personal” conforme a la legislación colombiana.

En el ámbito institucional, la Política de Protección de Datos Personales del Servicio Geológico Colombiano (SGC) consulte la política aquí refuerza este compromiso al establecer los lineamientos bajo los cuales la entidad garantiza el tratamiento adecuado de los datos personales de sus usuarios, empleados, contratistas y terceros.

Esta política, en cumplimiento de la Ley 1581 de 2012, fija principios como la legalidad, finalidad, transparencia, seguridad y confidencialidad, aplicables a todo el ciclo de vida de los datos, desde su recolección hasta su eliminación o anonimización.

El SGC dispone que los procesos de eliminación y supresión de datos personales deben realizarse asegurando su irreversibilidad y protegiendo la confidencialidad y seguridad de la información.

En este contexto, la anonimización se integra como una fase previa y complementaria a la eliminación, cuando se requiere conservar información con fines estadísticos, analíticos o de investigación, sin comprometer la identidad de las personas.

De esta forma, el proceso de anonimización implementado por el SGC se encuentra plenamente respaldado por:

- El marco normativo nacional en protección de datos personales (Ley 1581 de 2012, Decreto 1377 de 2013, Decreto 1081 de 2015).
- Las directrices de la Superintendencia de Industria y Comercio en materia de responsabilidad demostrada.

- Y la Política de Protección de Datos Personales del SGC, que establece las obligaciones de seguridad, confidencialidad y eliminación irreversible de datos personales.

En conjunto, estas disposiciones aseguran que la anonimización se desarrolle bajo criterios de legalidad, trazabilidad y transparencia, garantizando que los datos tratados con fines científicos o de gestión institucional no permitan, en ningún caso, la identificación de las personas titulares de la información.

7. Anonimización

Proceso mediante el cual los datos personales son transformados de forma irreversible, de modo que no puedan asociarse con una persona natural, directa o indirectamente, ni aun mediante el cruce con otras fuentes.

Figura 1: Conceptos de anonimización -MinTIC y GPDR

8. Fases del proceso de anonimización

El siguiente diagrama representa el flujo general del proceso de anonimización de datos, diseñado para asegurar una ejecución controlada, coherente y verificable de cada etapa del tratamiento. Este proceso se orienta a preservar la protección de la privacidad de los titulares sin comprometer la utilidad y valor analítico de la información, integrando prácticas de planeación, evaluación de riesgos, aplicación de técnicas y seguimiento continuo.

De esta manera, se garantiza que la anonimización se desarrolle conforme a los principios de legalidad, trazabilidad, confidencialidad y transparencia establecidos en la normativa vigente y en las política institucional de protección de datos personales.

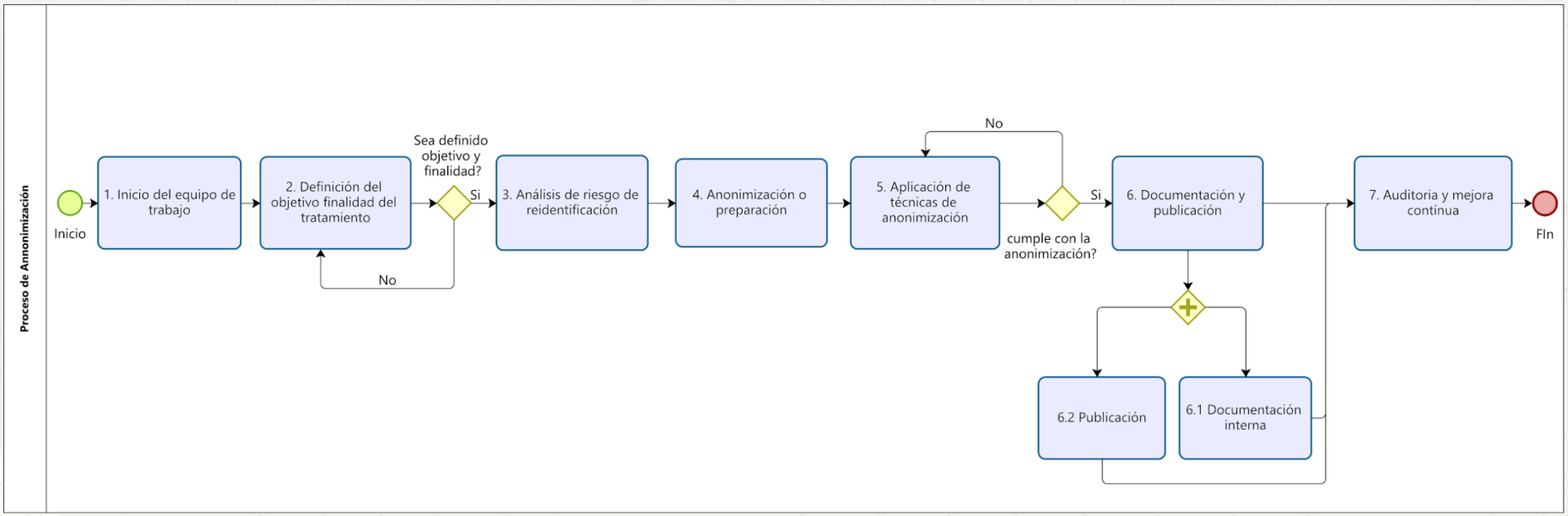


Figura 2: Proceso de anonimización

8.1. Definición del equipo de trabajo

En esta primera fase se trata de organizar los roles, responsabilidades y competencias para llevar a cabo el proceso de anonimización.

Rol / Perfil	Enfoque principal	Funciones principales	Responsabilidades clave
Responsable del tratamiento	Gobernanza y roles / Planeación estratégica	Define el propósito y alcance del proceso, autoriza recursos y establece la finalidad de los datos anonimizados.	Aprueba el inicio del proceso, valida resultados finales y garantiza cumplimiento normativo y ético.
Delegado/a de Protección de Datos (DPD/DPO)	Evaluación contextual / Evidencia y comunicación	Supervisa el cumplimiento legal, revisa las medidas aplicadas y evalúa los riesgos de privacidad.	Asesora al responsable, emite observaciones sobre riesgos residuales y mantiene la trazabilidad documental.
Equipo técnico de anonimización	Diagnóstico técnico / Transformación de datos	Ejecuta el análisis de variables, realiza pruebas previas y aplica las técnicas de anonimización.	Identifica variables sensibles, transforma datos (supresión, generalización, etc.) y documenta cada modificación.
Equipo de evaluación de riesgos o auditoría interna	Evaluación contextual / Vigilancia y evolución	Evalúa la efectividad de la anonimización y mide los riesgos de reidentificación en distintas fases.	Ejecuta pruebas de reidentificación, valida métricas de riesgo residual y propone mejoras continuas.
Responsable de seguridad de la información (CISO)	Diagnóstico técnico / Vigilancia y evolución	Asegurar los entornos y la integridad de los datos durante todo el proceso.	Asegurar la protección de los entornos tecnológicos y la integridad de los datos durante el proceso de anonimización, implementando controles que garanticen la confidencialidad, disponibilidad y trazabilidad de la información tratada.
Comité de seguimiento o coordinación	Gobernanza y roles / Evidencia y comunicación	Coordina la interacción entre áreas, revisa entregables y asegura la coherencia del proceso.	Aprueba fases, valida avances y emite el acta de cierre de cada etapa.

¿Por qué es importante?

Porque un proceso de anonimización no es puramente técnico, sino también organizativo: fallos en la segregación de funciones, conocimientos, documentación, pueden comprometer la "cadena de anonimización" y derivar en riesgo de reidentificación.

8.2. Definición del objetivo y finalidad del tratamiento

En esta fase se clarifica por qué se van a anonimizar los datos y cómo se van a usar después.

Por ejemplo: publicación como datos abiertos, análisis interno de la organización, investigación, etc.

- Esta finalidad condiciona la técnica de anonimización que se usará, el nivel de utilidad que se espera de los datos, el umbral de riesgo aceptable, los destinatarios, las condiciones de uso.
- También implica definir qué variables deben preservarse para cumplir la finalidad, cuáles pueden eliminarse, cuáles deben modificarse.
- Además, se debe definir el contexto de uso: abierto, restringido, con cláusulas de confidencialidad, etc.

8.3. Análisis de riesgos de reidentificación

Es una fase crítica: evaluar los riesgos de que, una vez aplicado el proceso de anonimización, exista la posibilidad de reidentificar a una persona

Esto incluye:

- Identificación de activos implicados (datos, sistemas, variables, procesos) y su categorización (sensibilidad, grado de identificación directa/indirecta)
- Identificar los riesgos: singularización ¿puede individualizarse una persona en el conjunto?, vinculabilidad ¿se puede vincular un registro con otro o con fuentes externas?, inferencia ¿se puede inferir información privada a partir de lo que queda?
- Valorar cada riesgo (probabilidad + impacto) y determinar un umbral aceptable de riesgo residual. Se evalúa el riesgo de reidentificación, considerando dos dimensiones: la probabilidad de que ocurra y el impacto que tendría en caso de materializarse. El resultado permite estimar un nivel de riesgo residual y decidir si este es aceptable o no antes de publicar los datos.

La AEPD (Agencia española de protección de datos), en su guía sobre anonimización, propone una metodología estructurada que inicia con la identificación de amenazas:

1. Reidentificación directa, por presencia de datos únicos o fácilmente combinables.
2. Reidentificación indirecta, mediante inferencias o correlaciones externas.
3. Revelación de información sensible, como datos de ubicación o historial clínico.

Posteriormente, se asigna una probabilidad (en una escala de 1 a 5) y un impacto (también de 1 a 5) a cada amenaza identificada.

Con estos valores se construye una matriz de riesgo, definida como:

Riesgo= Probabilidad × Impacto

Finalmente, se establece un umbral de aceptación, por ejemplo:

Nivel de riesgo	Interpretación
≤ 6	Aceptable
7–12	Requiere revisión o medidas adicionales
≥ 15	No asumible

Este enfoque permite visualizar rápidamente qué atributos o grupos de datos deben tratarse nuevamente mediante técnicas adicionales de anonimización.

- Diseñar medidas de salvaguarda: técnicas, organizativas, de gobernanza para evitar que esos riesgos se materialicen.
- Hacer revisiones periódicas, porque el riesgo de reidentificación puede aumentar con el tiempo, nuevas técnicas, nuevos datos externos

Importante 💡

- El hecho de anonimizar no elimina el 100% riesgo de reidentificación. Siempre habrá un nivel residual.
- La utilidad del dato y la protección de privacidad entran en tensión: mayor anonimización = menos utilidad, y viceversa.

8.4. Pre-anonimización o preparación

Aquí ya se trabaja sobre el conjunto de datos antes de aplicar las técnicas propiamente dichas de anonimización.

Algunas tareas:

- Identificar las variables de identificación directa (por ejemplo nombre, cédula, número de celular , correo electrónico) e indirecta (edad, sexo, código postal, combinación de variables).

- Clasificar las variables por tipo, sensibilidad, función, identificadores directos, identificadores directos, identificadores indirectos, atributos sensibles (salud, orientación, etnia, biometría, etc.).
- Decidir qué variables se eliminarán, cuáles se modificarán, cuáles se mantendrán, en función de la finalidad y el riesgo.
- Verificar la viabilidad técnica y práctica del proceso: ¿es posible anonimizar sin perder la utilidad esperada? ¿Qué coste tiene? ¿Qué técnicas aplicar?
- Realizar un “piloto” o prueba inicial para ver cómo afecta la anonimización a la utilidad de los datos.

8.5. Aplicación de técnicas de anonimización

En esta fase se aplican las técnicas concretas que transforman los datos para reducir el riesgo de reidentificación, al tiempo que se busca preservar la mayor utilidad posible

Técnica principal	**Métodos**	**Descripción general**	**Ventajas**	**Limitaciones o riesgos**	**Ejemplo práctico**
Aleatorización	Adición de ruido.	Se altera ligeramente el valor original (numérico o categórico) para evitar la identificación directa.	Reduce la posibilidad de coincidencias exactas.	Puede distorsionar el análisis estadístico si se aplica en exceso.	Cambiar 35 años → 36 o 34; ingreso $1.500.000 \rightarrow$1.480.000.
	Permutación o intercambio.	Los valores se redistribuyen entre registros, manteniendo la distribución general.	Mantiene coherencia estadística.	Riesgo si el número de registros es pequeño o se combina con datos externos.	Asignar códigos postales de unos registros a otros en el mismo grupo.
	Privacidad diferencial.	Se añade ruido controlado bajo parámetros matemáticos (ε) que limitan la información deducible sobre un individuo.	Método formal y cuantificable de protección.	Requiere conocimiento técnico y afecta la precisión del análisis.	Publicar estadísticas agregadas con ruido calibrado (ej. conteos ±3).
Generalización	Agregación / Binning.	Se agrupan valores dentro de rangos amplios (ej. edades 30–40 en lugar de 34).	Mantiene utilidad analítica.	Puede perder detalle para análisis finos.	Edad exacta 34 → grupo “30–40 años”.
	Supresión parcial o total.	Se eliminan o enmascaran datos identificadores (nombres, coordenadas).	Eficaz y simple.	Si se aplica demasiado, se reduce la utilidad de los datos.	“Juan Pérez” → “J **P**”.

	Reducción de precisión.	Limita la exactitud de valores continuos (truncar decimales, coordenadas).	Evita reidentificación geográfica o temporal.	Puede generar errores si no se documenta.	GPS 4.12345 → 4.12; fecha 2025-10-21 10:32 → “2025-10-21”.
Seudonimización	Cifrado o hash.	Sustituye identificadores por valores codificados irreversibles o recuperables bajo clave.	Permite trazabilidad sin exponer datos reales.	No equivale a anonimización completa; sigue siendo dato personal.	“ID_Usuario_123” → “5f4dcc3...”
	Tokenización.	Reemplaza datos sensibles por tokens aleatorios vinculados mediante una tabla segura.	Reversible de forma controlada.	Si la tabla se filtra, se pierde la protección.	“Tarjeta 4111-2222-3333-4444” → “TKN-9281”.
Combinadas / Avanzadas	k-anonimato.	Garantiza que cada registro sea indistinguible de al menos k–1 otros.	Equilibrio entre privacidad y utilidad.	Vulnerable a ataques por homogeneidad o conocimiento externo.	Si k=5, cada combinación edad–género–zona aparece en ≥5 registros.
	l-diversidad y t--closeness.	Añaden diversidad y distancia en atributos sensibles.	Más robustas ante ataques de inferencia.	Mayor complejidad de implementación.	En grupo k=5 debe haber ≥2 diagnósticos distintos (l-diversidad).

importante 💡

El proceso de anonimización no es puramente técnico, sino también organizativo: fallos en la segregación de funciones, conocimientos, documentación, pueden comprometer la “cadena de anonimización” y derivar en riesgo de reidentificación.

8.6. Documentación y publicación

En esta fase se consolida toda la evidencia del proceso de anonimización y se establecen las condiciones bajo las cuales los datos podrán ser difundidos o compartidos.

Documentación interna :
La documentación del proceso constituye evidencia de cumplimiento y de responsabilidad demostrada frente al tratamiento seguro de los datos personales.

- Se registran las decisiones tomadas, técnicas usadas, variables eliminadas o modificadas.
- Se anexa una evaluación final del riesgo residual.
- Se documentan los roles, fechas, herramientas y validaciones realizadas.
- Este registro se guarda en el expediente del tratamiento.

Públicación :
Las restricciones de acceso podrán clasificarse como uso interno, uso restringido bajo convenio o uso abierto, según la evaluación de riesgo residual y el tipo de dato tratado.

- Si los datos se publican (por ejemplo, como datos abiertos), se añaden metadatos sobre el proceso de anonimización.
- Se definen condiciones de uso, licencias, advertencias y contacto del responsable.
- Se aplican restricciones de acceso si corresponde (p. ej. uso interno, acceso académico).

Una vez completado el proceso de anonimización, se generan dos líneas de acción paralelas: la documentación interna para control y trazabilidad, y la publicación de datos, cuando proceda, para su uso o divulgación autorizada.

Esta fase busca evaluar de forma continua la eficacia de las medidas adoptadas, detectar posibles vulnerabilidades y promover la mejora progresiva del proceso .

El objetivo es mantener un nivel de riesgo de reidentificación aceptable en el tiempo, garantizando la confianza y sostenibilidad del tratamiento.

- Revisión periódica del nivel de riesgo de reidentificación ante nuevas fuentes de datos o tecnologías.

- Auditorías internas o externas para verificar el cumplimiento de políticas y estándares.

- Incorporación de nuevos métodos de anonimización más eficientes o seguros.

- Actualización de documentación, roles y procedimientos.

- Control de uso: detectar si los datos anonimizados han sido mal utilizados o correlacionados externamente.

8.7. Auditoría y mejora continua

Esta fase busca evaluar de forma continua la eficacia de las medidas adoptadas, detectar posibles vulnerabilidades y promover la mejora progresiva del proceso .

El objetivo es mantener un nivel de riesgo de reidentificación aceptable en el tiempo, garantizando la confianza y sostenibilidad del tratamiento.

- Revisión periódica del nivel de riesgo de reidentificación ante nuevas fuentes de datos o tecnologías.
- Auditorías internas o externas para verificar el cumplimiento de políticas y estándares.
- Incorporación de nuevos métodos de anonimización más eficientes o seguros.
- Actualización de documentación, roles y procedimientos.
- Control de uso: detectar si los datos anonimizados han sido mal utilizados o correlacionados externamente.

importante 💡

Esto convierte el proceso en un “ciclo de vida”, no en una acción puntual. La anonimización no termina al publicar: debe mantenerse viva y revisada.

8.8. Matriz RACI para fases de anonimización

Paso	Fase del Proceso	Responsable del Tratamiento	Delegado de Protección de Datos	Equipo Técnico de Anonimización	Equipo de Evaluación de Riesgos	Responsable de Seguridad	Comité de Seguimiento
1	Definición del equipo de trabajo	A	C	I	I	I	R
2	Definición de objetivos y finalidad	A	C	I	I	I	R
3	Análisis de riesgos	C	R	I	A	C	I
4	Pre-anonimización	I	C	R	C	A	I
5	Aplicación técnicas de anonimización	I	C	R	C	A	I
6	Documentación y publicación	A	R	C	C	I	R
7	Auditoria y mejora continua	C	C	I	R	R	A

Leyenda	Responsable	Aprobador	Consultado	Informado

Figura 3: *Matriz RACI para las fases de anonimización - elaboración propia*

9. Ejercicio de anonimización en Knime

KNIME es una plataforma open source orientada a la minería de datos, reconocida por su curva de aprendizaje sencilla gracias a su interfaz visual y enfoque low-code. Su funcionamiento se basa en nodos conectados entre sí, lo que permite mantener una trazabilidad completa a lo largo del flujo de datos. Cada nodo cumple una función específica y se clasifica por colores, de acuerdo con su propósito principal:

Color	Significado	Ejemplos
Amarillo	Transformación / manipulación	Filter, Joiner, GroupBy, String Manipulation
Azul	Visualización	Bar Chart, Scatter Plot
Verde	Machine Learning	Random Forest, SVM
Naranja	Lectura / escritura / Bases de datos	Excel Reader, DB Connector
Morado	Scripts	Python Script, R Snippet
Café	Minería de texto	TF-IDF, Parsing
Gris / Blanco	Componentes	Components, Metanodes
Rojo	Error	Falta de configuración o ejecución fallida

Además, KNIME permite una amplia integración con múltiples fuentes de datos, desde archivos planos como TXT o TSV, hojas de cálculo en Excel, hasta conexiones con bases de datos locales como SQL Server, MySQL u Oracle. También ofrece conectividad con servicios en la nube de grandes plataformas como GCP, AWS, Azure, así como con servicios expuestos mediante API. Como parte de esta guía, se realizó un proceso de anonimización por medio de un caso práctico, donde se explica paso a paso el uso de KNIME y cómo aplicar técnicas como aleatorización, generalización y seudonimización para proteger los datos personales.

9.1. Caso práctico

Contexto :
El área de bienestar de una compañía solicita los datos de 5.000 empleados para la compra de uniformes deportivos , poder revisar su estado en temas de salud ocupacional y poder generar estadísticas de participación por ciudad y genero.

Riesgo :

Si estos datos se entregan al proveedor de uniformes sin anonimizar, existe un riesgo significativo de exposición de datos personales y sensibles, como:

- correo electrónico
- teléfono
- afinidad religiosa
- información de salud
- entre otros

Esto podría constituir una violación de privacidad y de las políticas internas de protección de datos.

Escenario:

- ¿Quién solicita los datos?

El Comité de Bienestar de la empresa.

- ¿Con qué finalidad?

Organizar un programa de bienestar que incluye actividad física y la entrega de uniformes deportivos.

- ¿Qué datos se recopilan?

Nombre, edad, género, talla de camiseta, ciudad, teléfono y correo electrónico.

- ¿Por qué requieren datos a nivel individual?

El proveedor de uniformes necesita identificar qué persona recibe cada talla, ya que las camisetas serán personalizadas y llevarán el logo corporativo. El área de Bienestar requiere enviar correos y notificaciones personalizadas para coordinar entrenamientos según ciudad y horario, así como prever condiciones de salud (validar el estado de salud de cada empleado para determinar qué tipo de ejercicios puede realizar de manera segura).

1. Excel Reader

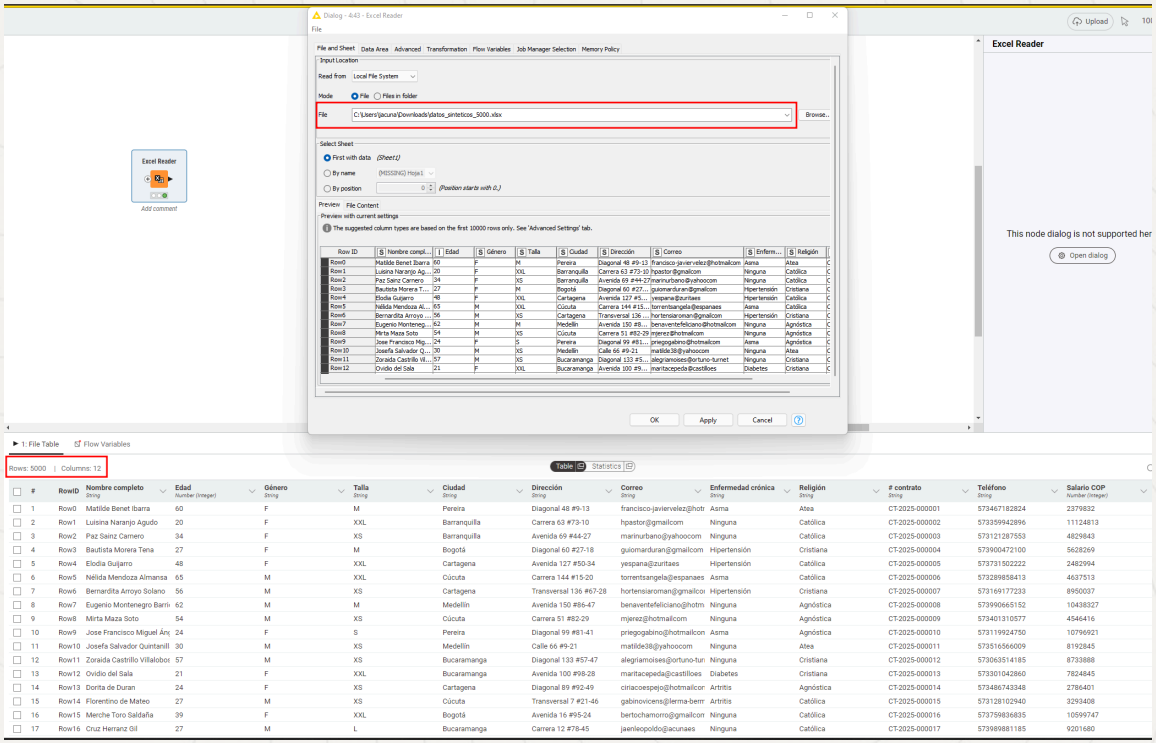


Figura 4: *Integración de datos a tráves del nodo Excel reader*

Objetivo:

Cargar el archivo inicial con los 5000 registros de empleados.

Técnica de anonimización: No aplica (solo lectura de datos).

Configuración: Seleccionar la ruta del archivo Excel → Aplicar y aceptar.

Resultado: Los datos originales quedan disponibles para ser procesados en el flujo.

2. GroupBy

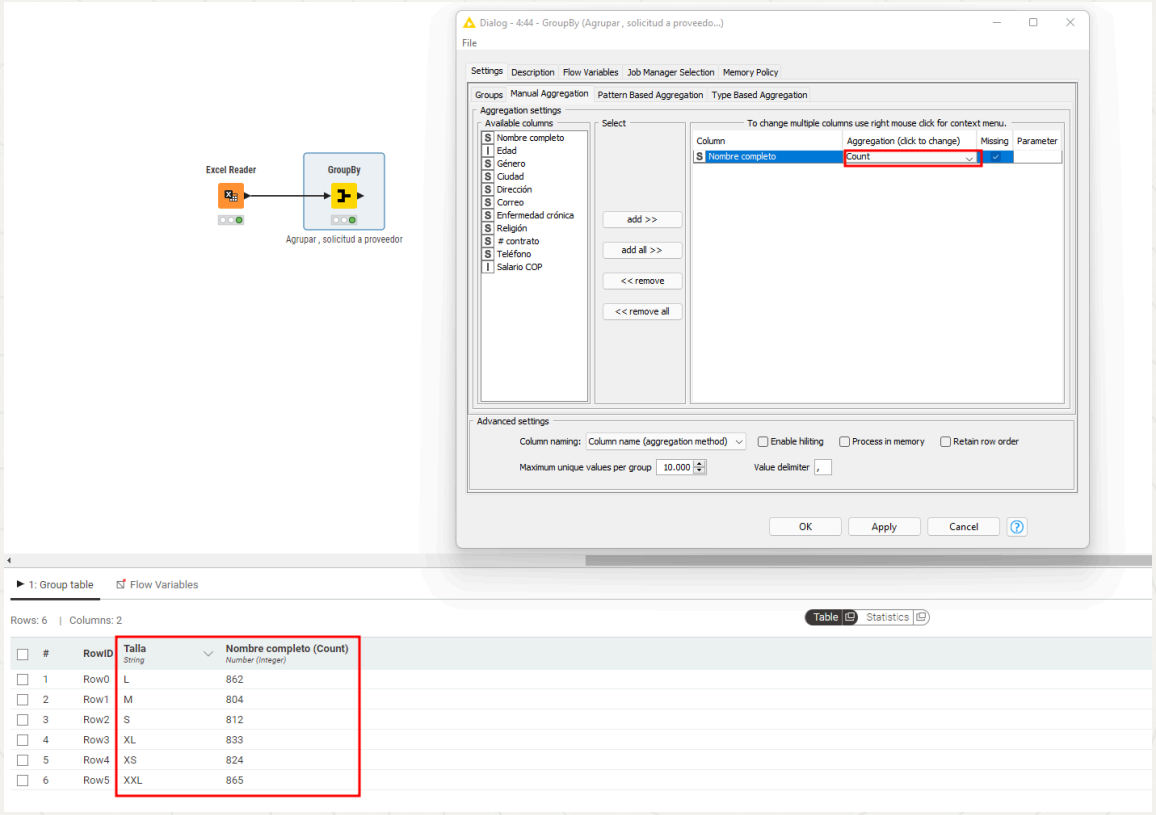


Figura 5: *Resultado nodo groupby*

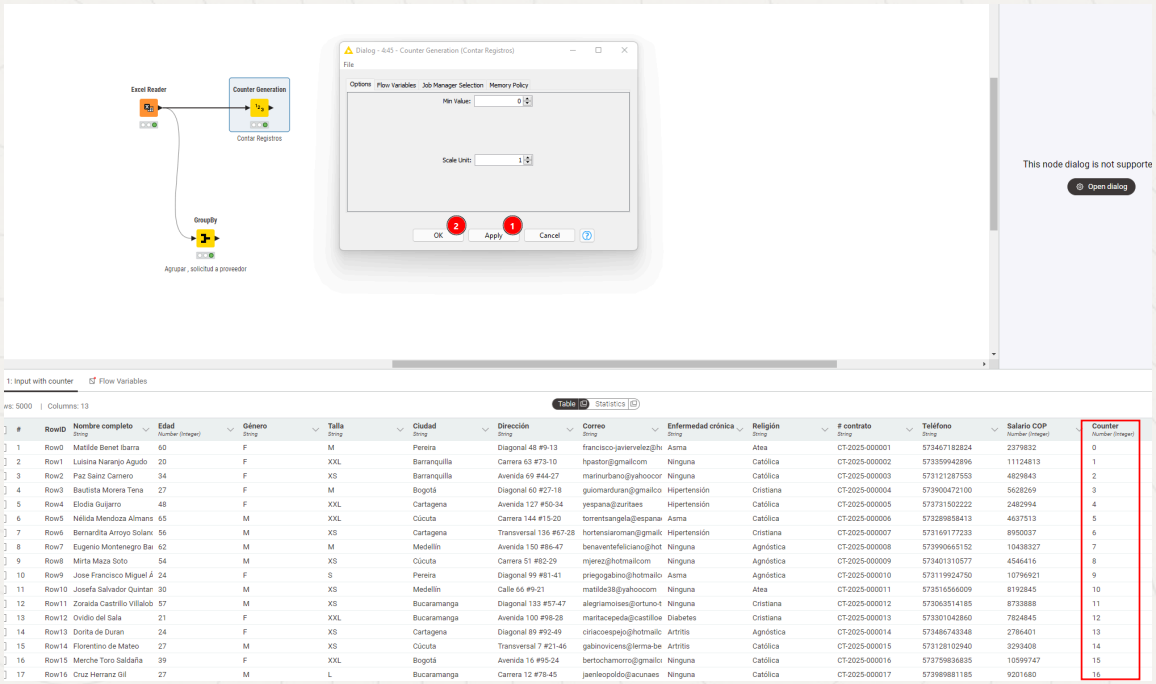
Objetivo: Calcular la cantidad total de camisetas por talla sin exponer información individual.

Técnica de anonimización: Generalización → Agregación (Se agregan los datos a un nivel grupal, eliminando detalle personal.)

Configuración: · Agrupar por Talla. · Crear agregación Count.

Resultado: Un conteo por talla útil para el proveedor sin mostrar datos individuales.

3. Counter Generation



#	RowID	Nombre completo	Edad	Género	Talla	Ciudad	Dirección	Correo	Enfermedad crónica	Religión	# contrato	Teléfono	Salario COP	Counter
1	Row0	Matilde Benet Ibarra	60	F	M	Pereira	Diagonal 48 #9-13	francisco-javiervelez@h	Asma	Atea	CT-2025-000001	573467182824	2379832	0
2	Row1	Luisina Naranjo Agudo	20	F	XXL	Barranquilla	Carrera 63 #73-10	hpastor@gmail.com	Ninguna	Católica	CT-2025-000002	57335994289	11124813	1
3	Row2	Paz Sainz Carnero	34	F	XS	Barranquilla	Avenida 69 #44-27	marinurbano@yahoocor	Ninguna	Católica	CT-2025-000003	573121287553	4829843	2
4	Row3	Bautista Morera Tena	27	F	M	Bogotá	Diagonal 60 #27-18	guiomarduran@gmail.co	Hipertensión	Cristiana	CT-2025-000004	573900472100	5628269	3
5	Row4	Elodia Guijarro	48	F	XXL	Cartagena	Avenida 127 #50-34	yespana@zuritaes	Hipertensión	Católica	CT-2025-000005	573731502222	2482994	4
6	Row5	Nélida Mendoza Almans	65	M	XXL	Cúcuta	Carrera 144 #15-20	torrentsangela@espana	Asma	Católica	CT-2025-000006	573289858413	4637513	5
7	Row6	Bernardita Arroyo Solanc	56	M	XS	Cartagena	Transversal 136 #67-28	hortensiaroman@gmail.c	Hipertensión	Cristiana	CT-2025-000007	573169177233	8950037	6
8	Row7	Eugenio Montenegro Bai	62	M	M	Medellín	Avenida 150 #86-47	benaventefeliciano@hot	Ninguna	Agnóstica	CT-2025-000008	573990665152	10438327	7
9	Row8	Mirta Maza Soto	54	M	XS	Cúcuta	Carrera 51 #82-29	mjerez@hotmail.com	Ninguna	Agnóstica	CT-2025-000009	573401310577	4546416	8
10	Row9	Jose Francisco Miguel Á	24	F	S	Pereira	Diagonal 99 #81-41	priegogabino@hotmail.co	Asma	Agnóstica	CT-2025-000010	573119924750	10796921	9
11	Row10	Josefa Salvador Quintar	30	M	XS	Medellín	Calle 66 #9-21	matilde38@yahoo.com	Ninguna	Atea	CT-2025-000011	573516566009	8192845	10
12	Row11	Zoraida Castrillo Villalob	57	M	XS	Bucaramanga	Diagonal 133 #57-47	alegriamoises@ortuno-t	Ninguna	Cristiana	CT-2025-000012	573063514185	8733888	11
13	Row12	Ovidio del Sala	21	F	XXL	Bucaramanga	Avenida 100 #98-28	maritacepeda@castilloe	Diabetes	Cristiana	CT-2025-000013	573301042860	7824845	12
14	Row13	Dorita de Duran	24	F	XS	Cartagena	Diagonal 89 #92-49	ciriacocepejo@hotmail.c	Artritis	Agnóstica	CT-2025-000014	573486743348	2786401	13
15	Row14	Florentino de Mateo	27	M	XS	Cúcuta	Transversal 7 #21-46	gabinovicens@lerma-be	Artritis	Católica	CT-2025-000015	573128102940	3293408	14
16	Row15	Merche Toro Saldaña	39	F	XXL	Bogotá	Avenida 16 #95-24	bertochamorro@gmail.co	Ninguna	Católica	CT-2025-000016	573759836835	10599747	15
17	Row16	Cruz Herranz Gil	27	M	L	Bucaramanga	Carrera 12 #78-45	jaanleopoldo@acunaes	Ninguna	Católica	CT-2025-000017	573989881185	9201680	16

Figura 6: *Resultado nodo counter generation*

Objetivo:

Generar un contador único por registro para usarlo luego como parte del seudónimo.

Técnica de anonimización:

Seudonimización (preparación) (Se crea un identificador numérico que permitirá reemplazar el nombre real.)

Configuración:

- Arrastrar el nodo.
- Activar generación de contador incremental.

Resultado:

Se añade la columna Counter con valores 1, 2, 3, ..., n.

SERVICIO GEOLÓGICO COLOMBIANO

4. String Manipulation

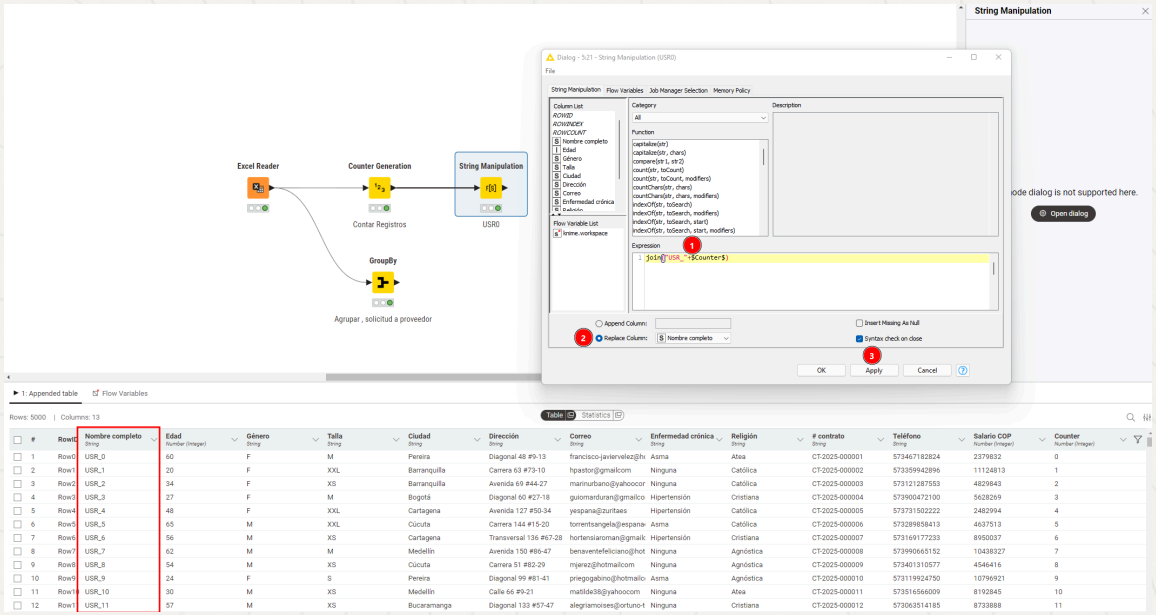


Figura 7: *configuración nodo string manipulation*

Objetivo:

Crear un identificador seudonimizado para reemplazar el nombre real del empleado.

Técnica de anonimización aplicada: Seudonimización (Se reemplaza el nombre por un código irreversible, como “USR_1”.)

Configuración: Expresión usada: join(“USR”, “Counter”) · Activar Replace column para reemplazar la columna de nombre real.

Resultado: Cada empleado queda identificado como USR_0, USR_1, USR_2..., sin exponer su nombre.

5. Rule Engine – Agrupación de Edad

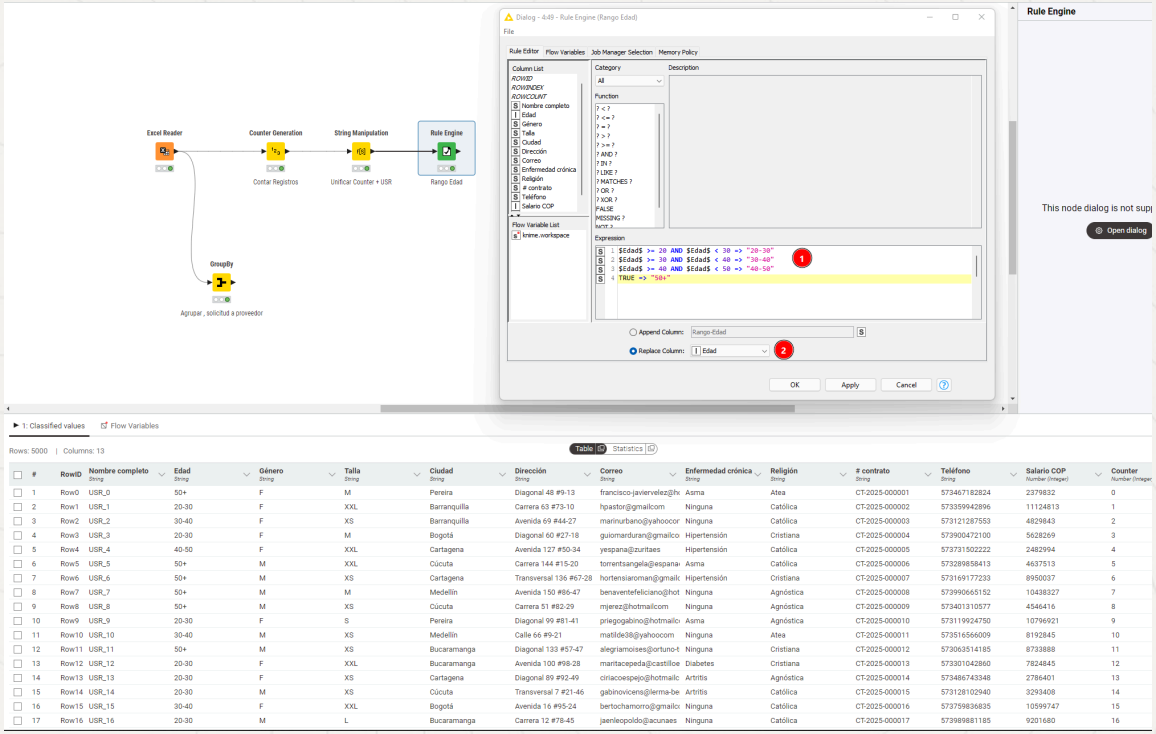


Figura 8: *configuración nodo y reglas de edad*

Objetivo: Generalizar la edad en rangos para evitar exposición de un dato personal preciso.

Técnica de anonimización aplicada: Generalización (Se reemplaza la edad exacta por un intervalo.) Configuración: Reglas usadas:

Edad >= 20 AND Edad < 30 => “20-30”

Edad >= 30 AND Edad < 40 => “30-40”

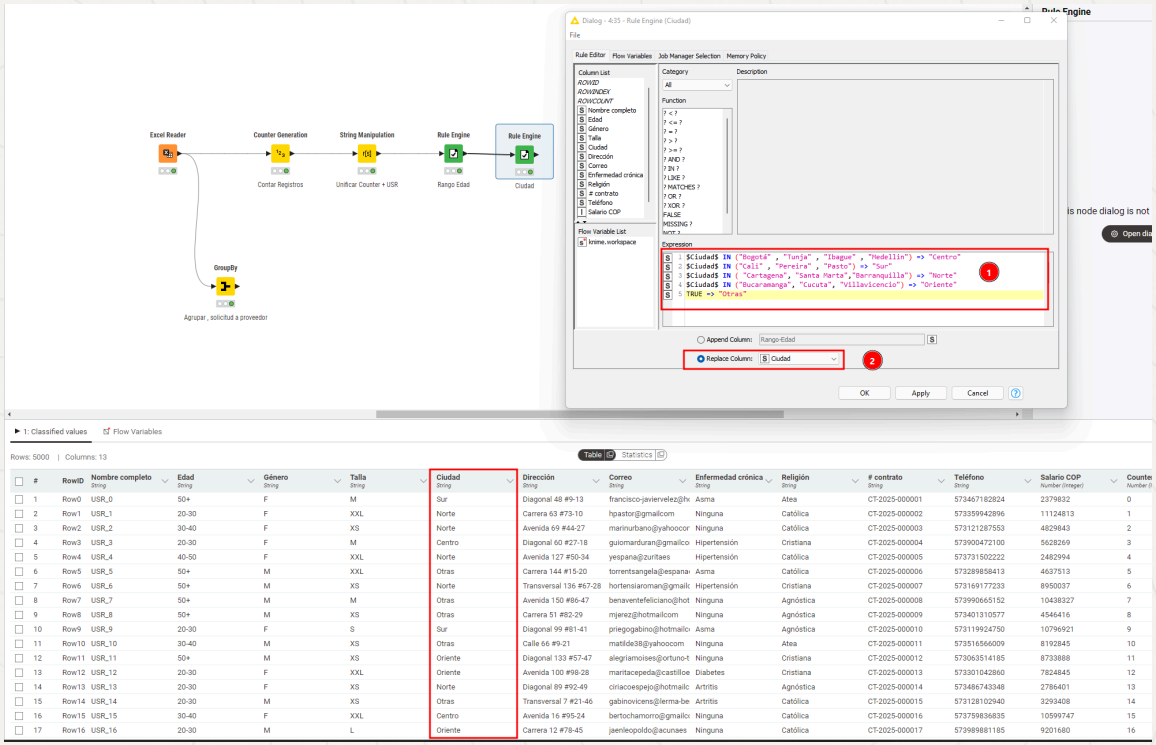
Edad >= 40 AND Edad < 50 => “40-50”

TRUE => “50+”

Reemplazar columna Edad.

Resultado: La edad deja de ser un dato exacto y se convierte en un rango menos identificable

6. Rule Engine – Agrupación por Ciudad



#	RowID	Nombre completo	Edad	Género	Talla	Ciudad	Dirección	Correo	Enfermedad crónica	Religión	# contrato	Teléfono	Salario COP
1	Row0	USR_0	50+	F	M	Sur	Diagonal 48 #9-13	francisco-javiervelez@hc	Asma	Atea	CT-2025-000001	573467182824	2379832
2	Row1	USR_1	20-30	F	XXL	Norte	Carrera 63 #73-10	hpastor@gmailcom	Ninguna	Católica	CT-2025-000002	573359942896	11124813
3	Row2	USR_2	30-40	F	XS	Norte	Avenida 69 #44-27	marinurbano@yahoocor	Ninguna	Católica	CT-2025-000003	573121287553	4829843
4	Row3	USR_3	20-30	F	M	Centro	Diagonal 60 #27-18	guiomarduran@gmailco	Hipertensión	Cristiana	CT-2025-000004	573900472100	5628269
5	Row4	USR_4	40-50	F	XXL	Norte	Avenida 127 #50-34	yespana@zuritaes	Hipertensión	Católica	CT-2025-000005	573731502222	2482994
6	Row5	USR_5	50+	M	XXL	Otras	Carrera 144 #15-20	torrentsangela@espana	Asma	Católica	CT-2025-000006	573289858413	4637513
7	Row6	USR_6	50+	M	XS	Norte	Transversal 136 #67-28	hortensiaroman@gmailc	Hipertensión	Cristiana	CT-2025-000007	573169177233	8950037
8	Row7	USR_7	50+	M	M	Otras	Avenida 150 #86-47	benaventefeliciano@hot	Ninguna	Agnóstica	CT-2025-000008	573990665152	10438327
9	Row8	USR_8	50+	M	XS	Otras	Carrera 51 #82-29	mjerez@hotmailcom	Ninguna	Agnóstica	CT-2025-000009	573401310577	4546416
10	Row9	USR_9	20-30	F	S	Sur	Diagonal 99 #81-41	priegogabino@hotmailc	Asma	Agnóstica	CT-2025-000010	573119924750	10796921
11	Row10	USR_10	30-40	M	XS	Otras	Calle 66 #9-21	matilde38@yahoocom	Ninguna	Atea	CT-2025-000011	573516566009	8192845
12	Row11	USR_11	50+	M	XS	Oriente	Diagonal 133 #57-47	alegriamoises@ortuno-t	Ninguna	Cristiana	CT-2025-000012	573063514185	8733888
13	Row12	USR_12	20-30	F	XXL	Oriente	Avenida 100 #98-28	maritacepeda@castillce	Diabetes	Cristiana	CT-2025-000013	573301042860	7824845
14	Row13	USR_13	20-30	F	XS	Norte	Diagonal 89 #92-49	ciriacoespejo@hotmailc	Artritis	Agnóstica	CT-2025-000014	573486743348	2786401
15	Row14	USR_14	20-30	M	XS	Otras	Transversal 7 #21-46	gabinovicens@lerma-be	Artritis	Católica	CT-2025-000015	573128102940	3293408
16	Row15	USR_15	30-40	F	XXL	Centro	Avenida 16 #95-24	bertochamorro@gmailc	Ninguna	Católica	CT-2025-000016	573759836835	10599747
17	Row16	USR_16	20-30	M	L	Oriente	Carrera 12 #78-45	jaenleopoldo@acunaes	Ninguna	Católica	CT-2025-000017	573989881185	9201680

Figura 9: *configuración nodo y reglas de ciudad*

Objetivo:

Generalizar la ciudad del empleado en grandes regiones para reducir riesgo de identificación.

Técnica de anonimización aplicada:

Generalización

Configuración:

Reglas usadas: Ciudad IN (“Bogotá”) => “Centro”

Ciudad IN (“Cali”, “Cartagena”, “Santa Marta”) => “Occidente”

Ciudad IN (“Bucaramanga”, “Medellín”, “Villavicencio”) => “Oriente”

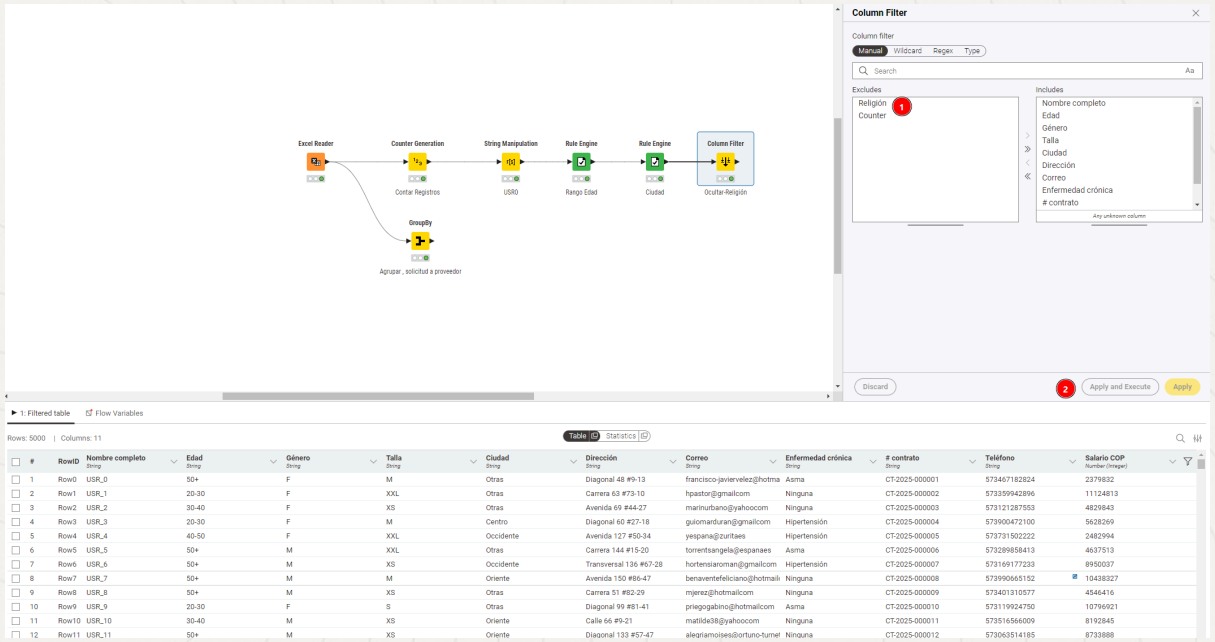
TRUE => “Otras”

Reemplazar columna Ciudad.

Resultado:

Las ciudades específicas se convierten en regiones, disminuyendo el riesgo de identificación directa o indirecta.

7. Column Filter-Religión



#	RowID	Nombre completo	Edad	Género	Talla	Ciudad	Dirección	Correo	Enfermedad crónica	# contrato	Teléfono	Salario COP
1	Row0	USR_0	50+	F	M	Otras	Diagonal 48 #9-13	francisco-javiervelez@hotma	Asma	CT-2025-000001	573467182824	2379832
2	Row1	USR_1	20-30	F	XXL	Otras	Carrera 63 #73-10	hpastor@gmailcom	Ninguna	CT-2025-000002	573359942896	11124813
3	Row2	USR_2	30-40	F	XS	Otras	Avenida 69 #44-27	marinurbano@yahoocom	Ninguna	CT-2025-000003	573121287553	4829843
4	Row3	USR_3	20-30	F	M	Centro	Diagonal 60 #27-18	guiomarduran@gmailcom	Hipertensión	CT-2025-000004	573900472100	5628269
5	Row4	USR_4	40-50	F	XXL	Occidente	Avenida 127 #50-34	yespana@zuritaes	Hipertensión	CT-2025-000005	573731502222	2482994
6	Row5	USR_5	50+	M	XXL	Otras	Carrera 144 #15-20	torrentsangela@espanaes	Asma	CT-2025-000006	573289858413	4637513
7	Row6	USR_6	50+	M	XS	Occidente	Transversal 126 #67-28	hortensiaroman@gmailcom	Hipertensión	CT-2025-000007	573169177233	8950037
8	Row7	USR_7	50+	M	M	Oriente	Avenida 150 #66-47	benaventefeliciano@hotmail	Ninguna	CT-2025-000008	573990665152	10438327
9	Row8	USR_8	50+	M	XS	Otras	Carrera 51 #82-29	mjerez@hotmailcom	Ninguna	CT-2025-000009	573401310577	4546416
10	Row9	USR_9	20-30	F	S	Otras	Diagonal 99 #81-41	priegogabino@hotmailcom	Asma	CT-2025-000010	573119924750	10796921
11	Row10	USR_10	30-40	M	XS	Oriente	Calle 66 #9-21	matilde38@yahoocom	Ninguna	CT-2025-000011	573516566009	8192845
12	Row11	USR_11	50+	M	XS	Oriente	Diagonal 133 #57-47	alegriamoises@ortuno-turner	Ninguna	CT-2025-000012	573063514185	8733888

Figura 10: *configuración nodo para ocultar columnas*

Objetivo Eliminar la columna Religión del dataset

Fundamento:

La religión es un dato sensible y debe ser tratado únicamente cuando exista una justificación clara y proporcional. Según el principio de minimización de datos, solo deben conservarse los atributos estrictamente necesarios. En este ejercicio, la religión no aporta valor a la asignación de tallas, envío de notificaciones, estadísticas por ciudad o validaciones básicas de salud ocupacional.

Mantenerla aumentaría el riesgo de reidentificación o exposición de información sensible sin beneficio operativo.

Técnica de anonimización

Supresión (eliminación del atributo) Se elimina por completo la columna sensible para reducir riesgo y cumplir buenas prácticas de privacidad.

Configuración

1. Insertar el nodo Column Filter.
2. En Exclude, seleccionar la columna: Religión.
3. Aplicar → Aceptar.

Resultado

La columna Religión se elimina del conjunto de datos, reduciendo riesgo y manteniendo solo la información estrictamente necesaria para el proceso.

8. String Manipulation – Enmascaramiento de Correo

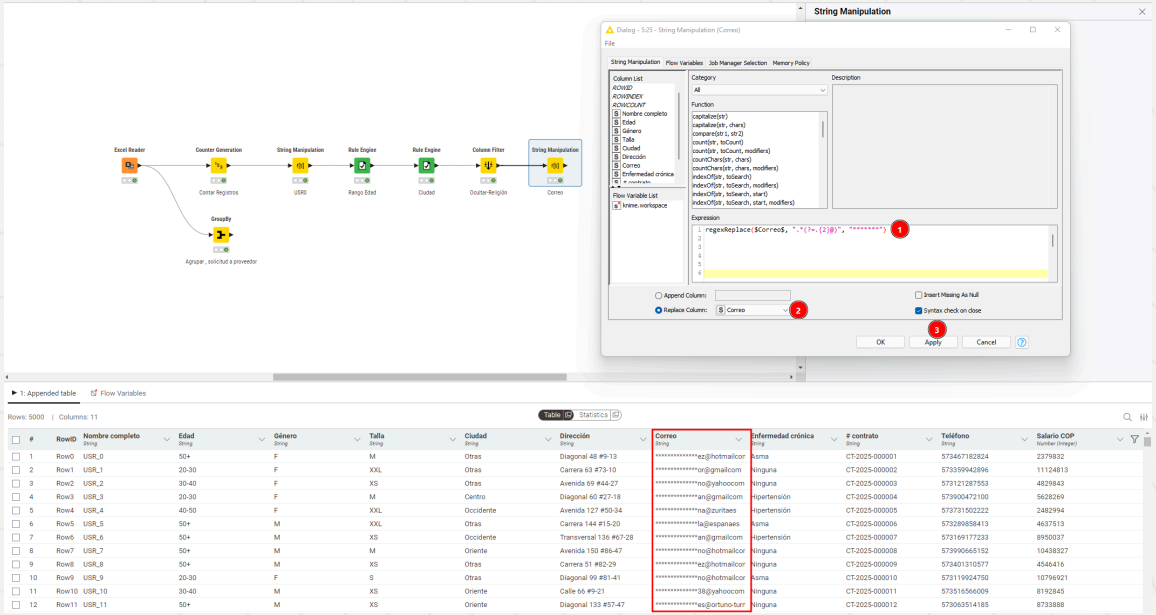


Figura 11: *configuración nodo enmascaramiento correo*

Objetivo:

Ocultar parcialmente la información del correo electrónico, manteniendo solo los dos primeros caracteres antes del símbolo @.

Técnica de anonimización

Seudoanonimización: Enmascaramiento (Elimina información sensible del identificador personal sin alterar el dominio del correo.)

Configuración:

Expresión usada: `regexReplace(Correo, ".*(?=.{2}@)", "*******")`

Descripción: · `regexReplace` busca todo lo que está antes de los últimos dos caracteres previos al símbolo @. · `"*******"` reemplaza ese contenido, dejando visibles únicamente los dos últimos caracteres antes del @ y el dominio.

Ejemplo:

- Entrada: `juan.perez@gmail.com`
- Salida: `******ez@gmail.com`

Reemplazar columna: Correo

Resultado:

El correo deja de mostrar el identificador completo del usuario, reduciendo el riesgo de reidentificación al conservar únicamente la estructura mínima necesaria para análisis posteriores.

9. String Manipulation – Enmascaramiento de Teléfono

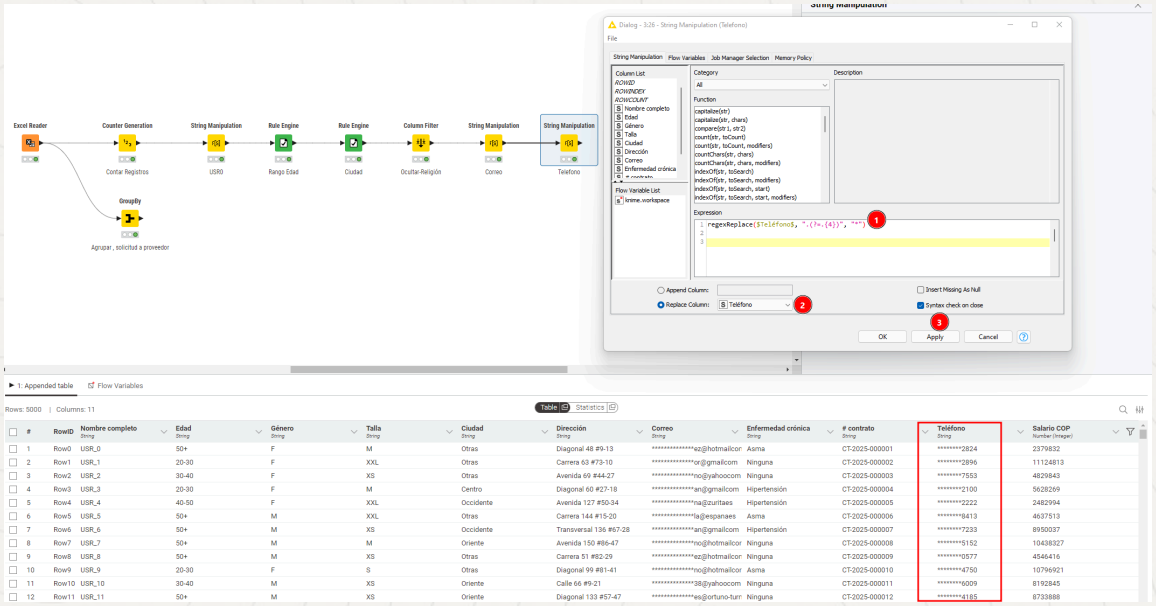


Figura 12: *configuración nodo enmascaramiento teléfono*

Objetivo:

Reducir la exposición del número telefónico ocultando todos los dígitos excepto los últimos cuatro.

Técnica de anonimización:

Seudonimización – Enmascaramiento parcial (Se reemplazan los dígitos iniciales por asteriscos, conservando únicamente la parte final del número.)

Configuración:

Nodo: String Manipulation

Reemplazar columna: Teléfono

Expresión usada:

```
regexReplace(Teléfono, ".(?=.{4})", "*")
```

Descripción de la lógica: El patrón `.(?=.{4})` selecciona cada carácter del número telefónico excepto los últimos cuatro y lo reemplaza por un asterisco (*).

Resultado:

El número queda en un formato protegido, por ejemplo:

`3204567890 → ****7890`

Esto disminuye la exposición del dato personal. No constituye anonimización completa, sino seudonimización, ya que conserva una parte reconocible del número.

10. GUID Generator – Tokenización de Contrato

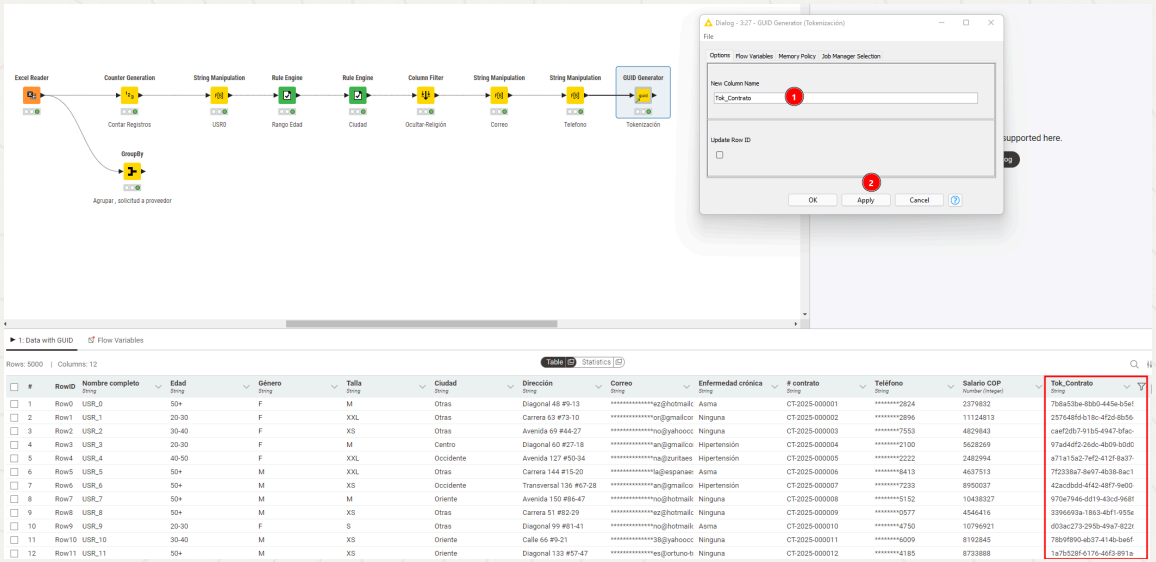


Figura 13: *configuración nodo tokenización*

Objetivo:

Reemplazar el número de contrato original por un identificador único que no permita inferir el valor real, reduciendo el riesgo de reidentificación.

Técnica de anonimización: Seudonimización – Tokenización

¿Qué es tokenización? Es el proceso de sustituir un valor real por un token aleatorio (en este caso un GUID). El token no contiene información del dato original, pero permite seguir usando la columna como identificador interno, sin revelar el valor sensible.

Configuración: Incluir nombre de columna Tok_Contrato

Resultado: La columna Contrato queda reemplazada por un token GUID totalmente aleatorio, como: **f47ac10b-58cc-4372--a567-0e02b2c3d479**. Lo cual permite trabajar con identificadores sin exponer el número real del contrato.

11. Hash Strings – Enfermedad crónica

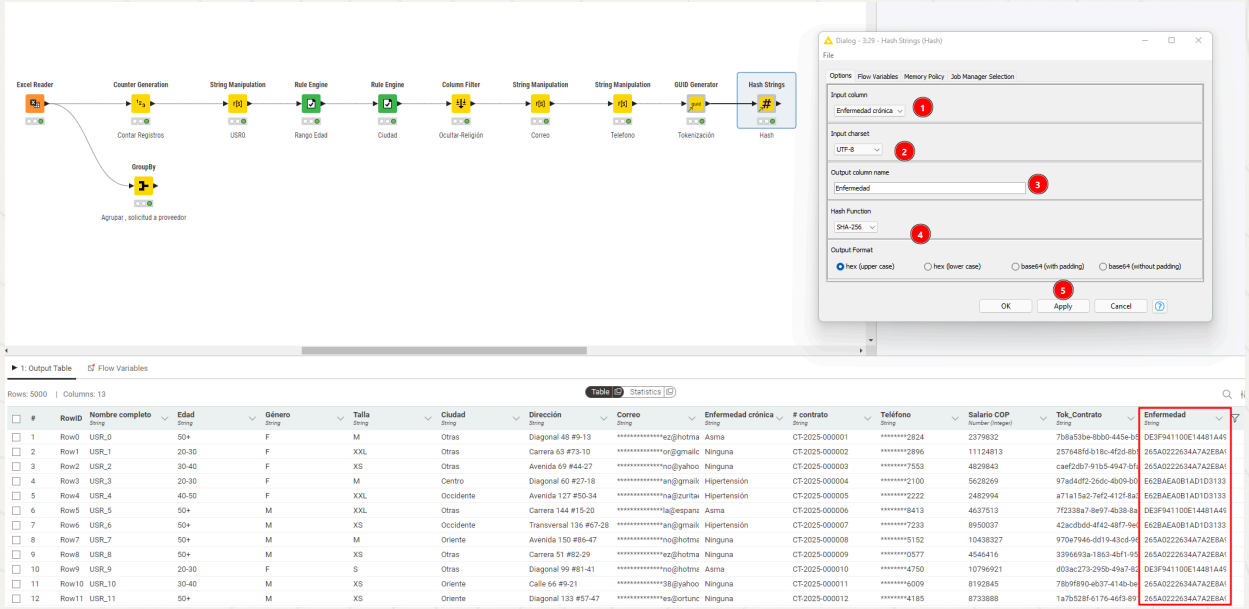


Figura 14: *configuración nodo hash*

Objetivo:

Proteger el dato sensible **Enfermedad crónica** mediante un proceso criptográfico irreversible que impida reconstruir el valor original, reduciendo el riesgo de exposición de información de salud.

Técnica de anonimización:

Seudonimización – Hashing (Función criptográfica irreversible)

¿Qué es un hash?

Un hash es una transformación criptográfica que convierte un valor original en una cadena alfanumérica de longitud fija.

Propiedades clave:

· Irreversible: no es posible recuperar el valor original a partir del hash. · Determinístico: el mismo valor de entrada siempre produce el mismo hash. · Unidireccional: diseñado únicamente para transformar, no para descifrar. · Alta entropía: dificulta ataques de adivinación o fuerza bruta.

Este mecanismo se utiliza para proteger datos altamente sensibles como información de salud, biometría o credenciales.

Configuración:

1. Insertar el nodo Hash Strings en el flujo.
2. Abrir la configuración del nodo.
3. En **Input column**, seleccionar: Enfermedad crónica.
4. En **Output column name**, escribir: Enfermedad (reemplaza la original).
5. En **Hash Function**, elegir: SHA-256 (estándar fuerte recomendado para protección de datos personales).
6. En **Output Format**, seleccionar: hex (upper case).
7. Apply → OK.

Resultado:

El dato sensible **Enfermedad crónica** se transforma en una cadena irreconocible, por ejemplo:

```
A3F79B94C6F5361F8E8760D78C9A8    32D0ABE1C22D13509C7E1A28BB0C8
E1E5F6
```

Esto protege la condición de salud del empleado, impidiendo que un tercero pueda deducir o reconstruir el diagnóstico original.

12. Column Filter – Eliminación de Atributos Sensibles

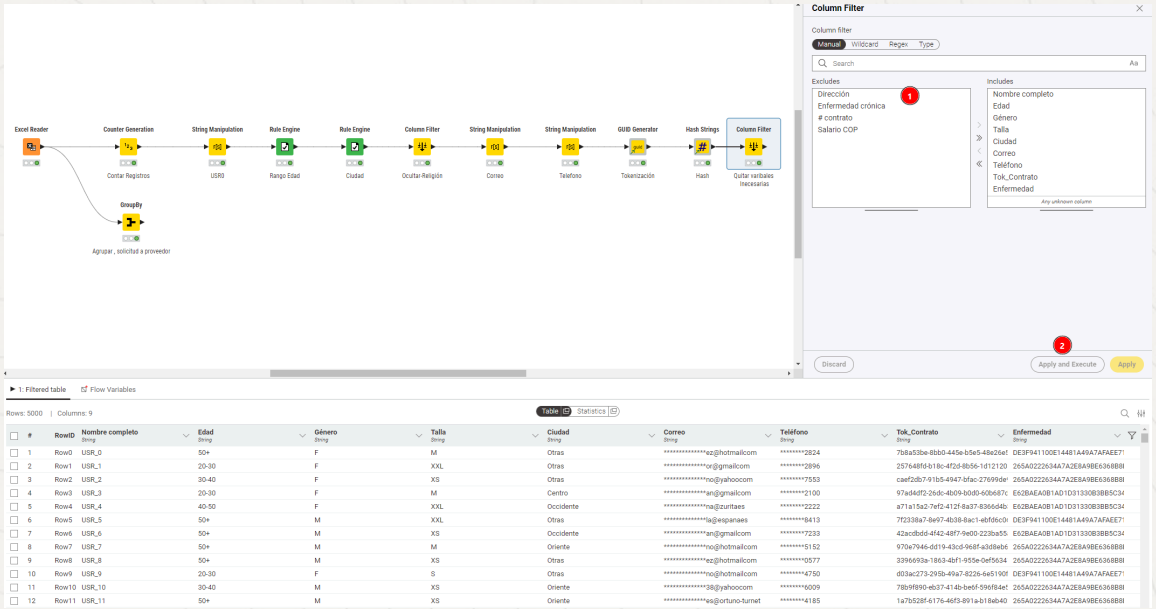


Figura 15: *eliminación atributos sensibles*

Objetivo:

Suprimir columnas que contienen datos personales o sensibles que no son necesarios para el proceso de bienestar, reduciendo el riesgo de identificación directa o indirecta.

Técnica de anonimización:

Supresión (eliminación del atributo) Consiste en retirar completamente columnas que no deben ser tratadas porque:

- No aportan a la finalidad del procesamiento.
- Incrementan el riesgo de reidentificación.
- Contienen información sensible según la normativa aplicable.

Fundamento:

Las variables Dirección, Enfermedad crónica, Contrato y Salario COP incluyen información que:

- No es requerida por el Comité de Bienestar ni por el proveedor.
- No aporta a la asignación de uniformes, estadísticas por ciudad o contacto operativo.
- Contiene datos sensibles (salud) o de alto riesgo (salario, dirección y número contractual).

La eliminación de estos atributos cumple el principio de minimización de datos y reduce la exposición innecesaria de información personal.

Verificación de columnas permitidas:

En **Includes** deben mantenerse únicamente las columnas necesarias:

- Nombre completo , edad , género , talla , ciudad, correo , teléfono , Tok_contrato , enfermedad

Resultado:

Las columnas sensibles quedan eliminadas completamente del conjunto de datos. El dataset resultante conserva solo los atributos estrictamente necesarios, ajustándose a los principios de minimización y proporcionalidad.

13. Rule Engine – Agrupación de Tallas

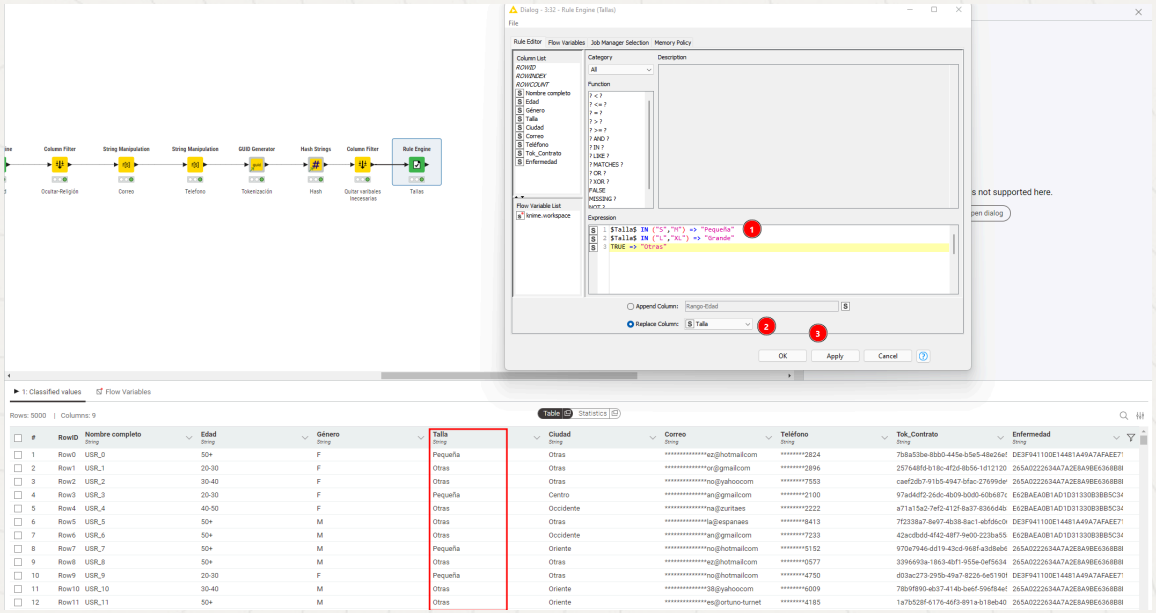


Figura 16: *agrupación tallas*

Objetivo:

Generalizar las tallas individuales en categorías amplias, reduciendo el nivel de detalle y evitando la exposición de características físicas precisas.

Técnica de anonimización:

Generalización Consiste en sustituir valores exactos de talla (S, M, L, XL, etc.) por rangos o categorías menos específicas. Esto disminuye la granularidad del dato y reduce el riesgo de identificación por atributos físicos.

Configuración:

Se definieron las siguientes reglas en el nodo Rule Engine:

```
Talla IN ("S", "M") => "Pequeña"
Talla IN ("L", "XL") => "Grande"
TRUE => "Otras"
```

Interpretación de las reglas:

- Las tallas S y M se agrupan como Pequeña.
- Las tallas L y XL se agrupan como Grande.
- Cualquier otro valor (incluidos casos atípicos) se clasifica como Otras.

Modo seleccionado: Replace Column → Talla (Se reemplaza directamente la columna original).

Resultado:

La talla deja de ser un atributo exacto y se convierte en una categoría amplia, reduciendo la posibilidad de inferir características físicas específicas de una persona.

12. Análisis de Singularidad y K-anonymity

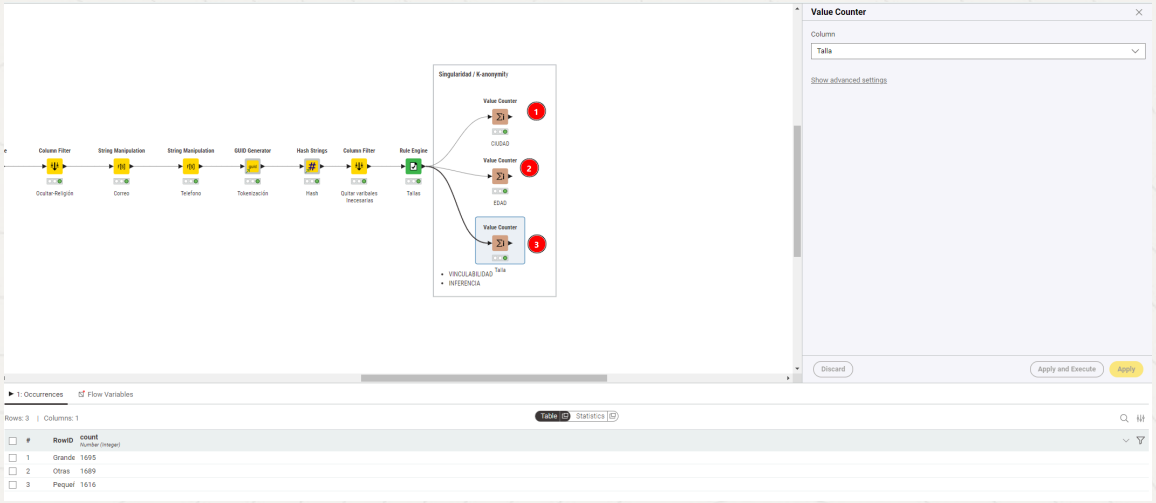


Figura 17: *Configuración Value Counter*

Objetivo:

Evaluar si existen valores o combinaciones de atributos que aparezcan muy pocas veces (singulares), lo que podría permitir reidentificación de personas aun después de aplicar técnicas de anonimización.

Técnica aplicada: K-anonymity

Se analiza cuántas veces aparece cada categoría en las variables consideradas cuasi-identificadores:

- Ciudad
- Edad (generalizada)
- Talla (generalizada)

Mientras mayor sea la frecuencia de cada categoría, mayor es el valor de “k”, y por lo tanto menor la probabilidad de reidentificar individuos.

Resultados de Value Counter
Varibale Ciudad

Categoría	Frecuencia
Centro	667
Occidente	1231
Oriente	123
Otras	1865

Análisis:

- La categoría Oriente (123) presenta el valor más bajo.
- Un k bajo (como 123) indica una posible vulnerabilidad de singularidad, pero aún es una frecuencia suficientemente amplia para evitar reidentificación directa.
- Las demás categorías tienen valores altos (≥ 667), lo cual reduce riesgo.

Variable: Edad

Categoría	Frecuencia
20-30	977
30-40	1004
40-50	947
50+	2072

Análisis:

- Todos los grupos de edad tienen valores superiores a 900, lo que representa un k alto.
- No hay singularidad.
- La generalización aplicada cumplió su propósito: hacer cada grupo suficientemente amplio.

Variable: Talla

Categoría	Frecuencia
Grandes	1695
Otras	1689
Pequeñas	1616

Análisis:

- Todas las categorías presentan frecuencias similares y superiores a 1600.
- Esto significa un nivel alto de k-anonymity.
- No existe riesgo de singularidad.

9.2. Interpretación – Vinculabilidad e Inferencia

9.3. Vinculabilidad:

La vinculabilidad ocurre cuando es posible relacionar registros entre sí aun si no contienen identificadores directos.

En este ejercicio:

- Las variables Ciudad, Edad y Talla generalizada tienen frecuencias relativamente altas.
- Debido a estos tamaños de grupo (k ≥ 123 en la categoría más pequeña), es difícil vincular un registro con otra fuente externa, porque no hay subgrupos extremadamente pequeños.
- La generalización en Edad y Talla reduce aún más la capacidad de vincular o diferenciar individuos.

Conclusión

El riesgo de vinculabilidad es bajo, especialmente en Edad y Talla. El único punto a vigilar es “Oriente”, con el menor k (123), aunque sigue siendo seguro.

9.4. Inferencia

La inferencia ocurre cuando un atacante puede deducir información sensible a partir de un grupo muy pequeño o un valor dominante. En este caso:

- Ninguna categoría presenta grupos pequeños (como 1, 2 o 3 casos), por lo que no se pueden inferir características específicas de una persona.
- Tampoco hay categorías donde más del 90% de una población tenga un mismo valor, que permitiría inferir algo de forma casi segura.
- La generalización (especialmente en Edad) ayudó a evitar conclusiones precisas sobre individuos.

Conclusión:

El riesgo de inferencia es muy bajo porque:

- Los grupos son amplios.
- No existen proporciones extremas que permitan deducciones sobre atributos individuales.

9.5. Riesgos Residuales

Aunque el proceso de anonimización aplicado reduce significativamente la posibilidad de reidentificación, es importante reconocer que ninguna técnica garantiza un riesgo cero. En este ejercicio persisten ciertos riesgos residuales que deben ser considerados:

1. Riesgo residual de reidentificación indirecta
Si bien no existen valores singulares ni combinaciones que generen subgrupos demasiado pequeños, es posible que bajo ciertos escenarios externos (p. ej., acceso a datos auxiliares o filtrado por condiciones adicionales no contempladas) se reconstruyan patrones que permitan acotar la identidad de un individuo. Este riesgo es bajo, pero inherente a cualquier método basado en generalización.

2. Riesgo de vinculación con fuentes externas
La K-anonymity protege contra ataques dentro del propio conjunto de datos, pero no elimina totalmente la vulnerabilidad frente a bases externas que un atacante pudiera poseer. Si otra base contiene categorías similares (por ejemplo, rangos de edad y ciudad), podría intentarse correlaciones probabilísticas. El uso de categorías suficientemente amplias reduce el impacto, pero el riesgo no desaparece por completo.

3. Riesgo de inferencia en grupos homogéneos
Aunque se aplicaron generalizaciones, aún existen categorías donde ciertos atributos tienen distribuciones ligeramente desbalanceadas. En grupos muy homogéneos, existe la posibilidad de inferir características sensibles de manera probabilística, incluso sin conocer la identidad del individuo (por ejemplo, si el 90 % de un grupo pertenece a una condición particular).

4. Riesgo derivado de variables cuasi-identificadoras no contempladas
Si en el futuro se incorporan nuevas variables (ocupación, fecha exacta, códigos geográficos específicos, etc.), podrían alterar el equilibrio logrado y generar nuevamente grupos pequeños o patrones únicos. La anonimización debe mantenerse como un proceso continuo, no como una acción única.

5. Riesgos asociados al reuso del dataset anonimizado
Si se realizan múltiples publicaciones o versiones del conjunto de datos, incluso con distintas capas de anonimización, podría darse un ataque de reconstrucción, donde un atacante combina varias versiones para aproximar valores originales. Esto suele ocurrir con conjuntos de datos compartidos en distintos contextos analíticos.

9.6. Conclusión general ejercicio Knime

El análisis aplicado mediante técnicas de anonimización sostenidas en K anonymity y evaluación de singularidad permitió verificar la solidez del tratamiento de datos implementado. Los resultados obtenidos demuestran que no existen valores singulares ni subgrupos peligrosamente pequeños, lo cual reduce significativamente la probabilidad de reidentificación directa. Asimismo, las transformaciones realizadas particularmente la generalización de variables como Edad y Talla contribuyeron a disminuir la precisión de atributos sensibles sin afectar el análisis estadístico global, mitigando de manera efectiva los riesgos asociados a la exposición de información personal.

Durante el proceso se empleó KNIME, una plataforma open source ampliamente utilizada en analítica, ciencia de datos y procesamiento seguro de información. Su naturaleza abierta ofrece varias ventajas:

Transparencia total en cada operación y nodo utilizado, lo cual facilita auditorías y verificaciones técnicas.

Reproducibilidad del flujo, permitiendo que los procesos de anonimización puedan ser revisados, versionados, automatizados o reaplicados en nuevos conjuntos de datos de manera consistente.

Flexibilidad y personalización, ya que admite integración con lenguajes como Python o R, así como extensiones adicionales de la comunidad.

Facilidad de uso, gracias a una interfaz visual que permite construir y documentar todo el flujo mediante nodos, sin requerir un conocimiento avanzado de programación.

Enfoque modular, que hace posible probar, comparar y ajustar diferentes técnicas de anonimización sin alterar el conjunto original.

El uso del nodo Value Counter permitió evaluar cuantitativamente los grupos resultantes de las transformaciones, confirmando que cada categoría mantiene tamaños suficientemente altos para garantizar un valor de k robusto. Como consecuencia, los riesgos de vinculabilidad la posibilidad de relacionar registros entre sí e inferencia la deducción de atributos sensibles a partir de patrones dominantes se encuentran en niveles bajos, incluso en los grupos con menor frecuencia relativa.

En conjunto, este ejercicio evidencia que es posible aplicar de manera eficiente prácticas de anonimización utilizando herramientas open source sin afectar la calidad analítica del dato. La solución implementada logra un equilibrio adecuado entre protección de datos personales, utilidad estadística y sostenibilidad técnica, convirtiéndose en un enfoque sólido para el tratamiento seguro de información en entornos organizacionales.

10. Referencias

- Servicio Geológico Colombiano. (s. f.). *Política de protección de datos personales del Servicio Geológico Colombiano*. https://www2.sgc.gov.co/AtencionAlCiudadano/Documents/Politicas-de-Tratamiento/Politica-de-Proteccion-de-Datos-Personales--del-SGC.pdf
- Asociación Española de Cumplimiento de Protección de Datos. (2024). *Guía completa de cumplimiento normativo en protección de datos*. https://aecpd.org/wp-content/uploads/2024/11/AECPD_Guia-completa-de-cumplimiento-normativo-en--proteccion-de-datos.pdf
- Departamento Administrativo Nacional de Estadística. (2024). *Guía de anonimización de datos*. https://www.dane.gov.co/files/sen/registros-administrativos/guia-anonimizacion-datos2024.pdf
- Archivo General de la Nación. (2020). *Guía de anonimización de datos estructurados*. https://www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Consulte/Recursos/Publicacionees/Guia_de_Anonimizacion-min.pdf
- Congreso de la República de Colombia. (2009). *Ley 1273 de 2009 por medio de la cual se modifica el Código Penal y se crea un nuevo bien jurídico tutelado denominado de la protección de la información y de los datos*. http://www.secretariasenado.gov.co/senado/basedoc/ley_1273_2009.html
- Congreso de la República de Colombia. (2012). *Ley 1581 de 2012 por la cual se dictan disposiciones generales para la protección de datos personales*. http://www.secretariasenado.gov.co/senado/basedoc/ley_1581_2012.html
- Data Protection Commission. (2019). *Anonymisation and pseudonymisation*. https://www.dataprotection.ie/sites/default/files/uploads/2019-06/190614%20Anonymisation%20and%20Pseudonymisation.pdf
- arbe. (s. f.).

  Hash strings [Workflow en KNIME Hub]. KNIME. https://hub.knime.com/arbe/spaces/Public/Hash%20Strings~uWE0_gg9wHIeXCYR/current-state