Anonimización

Contexto general

Este reporte documenta el proceso de anonimización y manejo ético de los datasets del CAETEC utilizados en este proyecto de Inteligencia Artificial Avanzada para la Ciencia de Datos. Los datos provienen de fuentes internas del socio formador (CAETEC) e incluye registros individuales de vacas, producción diaria y eventos registrados por l@s veterinari@s.

El conjunto de datos incluye archivos como los siguientes, considerando a XXX como identificadores únicos de distintas vacas del CAETEC.:

- `registro_ordeño.csv`
- `Todas la visitas a EO-PO XXX.csv (dataset de ordeño individual por vaca)`
- `Eventos de animales XXX.csv (dataset de ordeño individual por vaca)`

Dentro de los datasets no existen datos personales ni información confidencial de personas.

Políticas de acceso a los datos

Almacenamiento

- Los datos se encuentran resguardados en un Google Drive institucional: Data Exploration & Preparation Archivos csv
- El acceso a este repositorio requiere autorización previa de un miembro del equipo Vacas Saturno Saturnitas.
- No se permite el almacenamiento de los datos en dispositivos personales que no cuenten con cifrado.

Procesamiento

- El procesamiento de los datos está permitido únicamente en entornos de desarrollo locales y privados, o en notebooks seguros a los que solo puedan acceder las personas autorizadas por el equipo.

Acceso

- El acceso a los datos será concedido exclusivamente a personas verificadas dentro del espacio de trabajo en Notion o Google Drive, administrado por los miembros del equipo.

Verificación de datos

Debido al tipo de solución que el proyecto propone, se decidió ir en busca de datos adicionales para la evaluación de la factibilidad de la propuesta generada. Para esto los datos fueron recopilados de la máquina de ordeño DELAVAL, donde la doctora nos proporcionó los datos tal cual como salen de la maquila. Debido a esto los datos proporcionados cuentan con alguno de los siguientes atributos :

- Número de vaca
- Edad en dias
- Padre de la Madre - Semen
- Fecha de nacimiento
- Edad (a:mm)
- Padre de toro progenitor, ID/NRO de toro

Anonimizar los datos

Para este enfoque, se propone una anonimización básica únicamente como opción técnica, ya que los datos no contienen información personal de seres humanos ni requieren medidas avanzadas de protección.

La anonimización del ID de las vacas puede aplicarse de manera opcional cuando los datos se utilicen con fines de análisis, investigación o desarrollo de modelos, especialmente si el objetivo es compartir resultados fuera del entorno operativo del CAETEC o distribuir datasets a terceros sin exponer la estructura interna del rancho.

Si bien la identificación individual del ganado es un requisito legal para garantizar el control sanitario, la trazabilidad y la prevención de riesgos como zoonosis, fraude o pérdida, en contextos académicos o de análisis interno no existe obligación de anonimizar ni de conservar identificadores reales, quedando la decisión sujeta a las necesidades del proyecto.

En caso de elegir aplicar anonimización, esta permite evitar la exposición de información operativa o estructural del sistema productivo sin afectar la calidad del estudio, manteniendo la posibilidad de trabajar con datos representativos pero desasociados de su identificador original.

Técnica empleada para anonimizar los ID

Para anonimizar los datos, utilizaremos la técnica de sustitución, esta técnica permite sustituir el contenido de una columna de una base de datos por datos procedentes de una lista predefinida de valores ficticios, de modo que la información no pueda rastrearse hasta un individuo identificable. Esta técnica tiene la ventaja de mantener intacta la integridad de la información original.Data Anonymization Techniques - iNext

Al anonimizar los ID, se protege la información sensible del sistema productivo y de los registros oficiales, manteniendo la confidencialidad de la fuente sin afectar la calidad del estudio. Esta práctica permite cumplir con los objetivos de investigación mientras se resguarda la privacidad de los datos operacionales del rancho.

Para anonimizar los datos se utiliza la técnica de sustitución, que consiste en reemplazar el contenido de una columna de la base de datos con valores ficticios predefinidos. Esta técnica garantiza que:

- La información no puede rastrearse hasta un individuo identificable
- Se mantiene intacta la integridad referencial de los datos
- Las relaciones entre registros permanecen consistentes
- Los análisis estadísticos mantienen su validez

Tabla de sustitución de Identificadores de Datos

Carpeta de los archivos de las vacas:

https://drive.google.com/drive/u/1/folders/1JdYCPRt_OQrd2I1m_7gM1w3yCHedPl86

Identificador Original	Identificador Ficticio
1213	0001
1216	0002
1225	0003
1226	0004
1236	0005
1239	0006
1493	0007
1497	0008
1510	0009
1511	0010
1514	0011

1517	0012
1552	0013
1575	0014
1590	0015
2078	0016
2107	0017
2109	0018
2119	0019
2128	0020
2150	0021
5767	0022
6062	0023
6070	0024
6131	0025
6134	0026
6164	0027
6184	0028
6248	0029
8703	0030
8710	0031
8712	0032
8715	0033
8729	0034
8738	0035
8742	0036
8744	0037
8749	0038

8756	0039
8764	0040
8768	0041
8780	0042
8782	0043
8783	0044

Mecanismos y herramientas

La aplicación web que se desarrollará incorporará mecanismos automatizados de anonimización que transforman los datos sensibles en el momento de su carga, asegurando que ninguna información identificable quede expuesta en la base de datos operacional ni en las interfaces de usuario. Este documento explica cómo funcionan estos mecanismos de protección integrados en el flujo de ingesta de datos.

Proceso Automatizado Durante la Carga de Datos

Cuando un usuario carga datos crudos a la aplicación web, el sistema ejecuta automáticamente el siguiente flujo:

1 - Recepción de Datos Originales

El usuario importa archivos (CSV, Excel) o ingresa datos mediante formularios. Los datos llegan en su forma original conteniendo los identificadores reales de las vacas.

2 - Detección de Identificadores

El sistema identifica automáticamente las columnas que contienen IDs de vacas mediante el análisis de nombres de columnas y formatos de datos.

3 - Aplicación de Sustitución

Los identificadores originales se reemplazan automáticamente por sus correspondientes identificadores ficticios según la tabla de mapeo predefinida.

4 - Almacenamiento Anonimizado

Los datos transformados se almacenan en la base de datos PostgreSQL. Únicamente los identificadores ficticios quedan accesibles en el sistema operacional y en las interfaces de usuario para los roles que no sean administradores ni colaboradores.

5 - Tabla de Correspondencia Segura

La tabla de mapeo (ID original al ID ficticio) se almacena de forma separada y con acceso restringido, permitiendo la trazabilidad sólo cuando sea absolutamente necesario y autorizado.

Triggers y Middlewares

El sistema de auditoría de doble capa (triggers de PostgreSQL + middleware de aplicación) asegura que todos los accesos a datos queden registrados, proporcionando trazabilidad completa mientras se protege la confidencialidad de la información del rancho.

Consideraciones finales sobre la necesidad real de anonimización

Aunque este documento describe un proceso de anonimización mediante sustitución de identificadores, es importante aclarar que, debido a la naturaleza del dataset del CAETEC, la anonimización estricta no es un requisito obligatorio ni una medida de mitigación indispensable. Esto se debe a varios factores técnicos, éticos y legales:

1. No existe riesgo de exposición de datos personales

Los datasets proporcionados por CAETEC no contienen información de personas físicas, ni datos sensibles regulados por la Ley Federal de Protección de Datos Personales en Posesión de los Particulares (LFPDPPP).

Los registros incluyen únicamente atributos relacionados con **vacas**, como:

- Número de animal
- Edad en días
- Eventos reproductivos o sanitarios
- Producción de leche

Estos datos no permiten identificar ni inferir información sobre una persona humana, por lo que no existe riesgo de violación de privacidad o de tratamiento indebido de datos personales.

En consecuencia, la anonimización no responde a una necesidad legal, sino únicamente a un enfoque precautorio interno.

2. La anonimización total puede resultar contraproducente para el funcionamiento del sistema

El ecosistema digital propuesto para CAETEC incluyendo dashboards, trazabilidad y monitoreo del estado de cada animal depende del uso de identificadores reales para que el personal veterinario y operativo pueda reconocer inmediatamente a cada vaca.

Implementar anonimización rompería funcionalidades esenciales:

- La identificación en tiempo real de animales
- La visualización de su historial clínico y productivo
- La trazabilidad requerida por normativas como **NOM-001-SAG/GAN-2015**
- La correlación entre eventos (partos, tratamientos, ordeños)
- La correcta asignación de decisiones en la aplicación

Si cada vaca estuviera representada con un ID ficticio en la interfaz, el personal no podría reconocerla en campo, afectando la operación diaria del CAETEC.

3. La anonimización no aporta beneficios prácticos para CAETEC en el contexto actual

A diferencia de sistemas humanos donde la anonimización protege la privacidad individual, en este caso:

- Los animales no poseen derechos ARCO (acceso, rectificación, cancelación, oposición)
- Los identificadores no exponen procesos financieros, fiscales o personales
- No hay riesgos reputacionales, legales ni éticos asociados al uso de IDs reales
- El acceso ya está restringido mediante controles internos (Drive institucional, Notion, permisos de equipo)

4. Los identificadores reales son requeridos explícitamente por CAETEC para garantizar trazabilidad sanitaria

El propio CAETEC, como unidad productiva, está sujeto a regulaciones que exigen la trazabilidad individual del ganado. Para cumplir con estas normativas:

- Cada animal debe tener un identificador único y permanente
- Su historial debe ser accesible por personal autorizado
- Se requiere coherencia entre los registros digitales y los registros físicos
- La trazabilidad es fundamental en caso de auditorías o brotes sanitarios

Por tanto, la aplicación diseñada debe conservar los identificadores reales, no reemplazarlos, ya que su propósito operativo es precisamente facilitar la trazabilidad y el seguimiento sanitario en tiempo real.

Tabla de accesos

Como se indicó al inicio del documento, el acceso a los datos únicamente puede ser autorizado por los miembros del equipo y está bajo la administración de Vacas Saturno Saturnitas. Solo se concede acceso a personas verificadas dentro del espacio de trabajo institucional en Notion o Google Drive.

El repositorio de Google Drive permanece bajo la responsabilidad de a01710680@tec.mx

Cualquier solicitud de acceso deberá dirigirse a dicho miembro del equipo. Una vez validada la identidad y el propósito del uso de los datos, la persona solicitante será registrada en el log de accesos, cumpliendo con los mecanismos de control y trazabilidad establecidos.

https://drive.google.com/drive/folders/1JdYCPRt_OQrd2I1m_7gM1w3yCHedPl86?usp=drive_link

Nombre completo	Correo institucional	Rol / Motivo de acceso	Datos accedidos	Método de acceso (Drive/Notion/ App)	Autorizado por	Observaciones
@Ángel Mauricio Ramírez Herrera	A01710158 @tec.mx	Análisis y Desarrollo	Archivos consultados	Google Drive / Notion	a01710680@tec.mx	—
@Fidel Alexander Bonilla Montalvo	A01798199 @tec.mx	Análisis y Desarrollo	Archivos consultado	Google Drive / Notion	a01710680@tec.mx	—
@Kevin Ramirez Luna	A01711063 @tec.mx	Análisis y Desarrollo	Archivos consultado	Google Drive / Notion	a01710680@tec.mx	—
@Cristian Chávez Guía	A01710680 @tec.mx	Análisis y Desarrollo	Archivos consultado	Google Drive / Notion	a01710680@tec.mx	—
@Diego Antonio García Padilla	A01710777 @tec.mx	Análisis y Desarrollo	Archivos consultado	Google Drive / Notion	a01710680@tec.mx	—
@Guadalupe Paulina López Cuevas	A01701095 @tec.mx	Análisis y Desarrollo	Archivos consultado	Google Drive / Notion	a01710680@tec.mx	—
@José Eduardo Viveros Escamilla	A01710605 @tec.mx	Análisis y Desarrollo	Archivos consultado	Google Drive / Notion	a01710680@tec.mx	—