TFG EN INGENIERIA INFORMÁTICA, ESCUELA DE INGENIERÍA (EE), UNIVERSIDAD DE AUTONOMA DE BARCELONA (UAB)

Prueba de Concepto de Herramienta de Data Masking

Autor. Jakob Kotnik Lopez

Resumen — Enmascaramiento de datos es un proceso de transformación de los datos orientado a la protección, en el que es primordial intentar mantener el realismo de los mismos. Es esencial para proteger la privacidad, prevenir el robo de datos, cumplir con regulaciones, garantizar la seguridad en entornos de desarrollo y pruebas, y permitir análisis seguros en entornos de BI (Bussiness intelligence). Actualmente es un proceso de gran utilidad y valor para las empresas dentro del sector de la tecnología o cualquiera que trabaje con grandes tamaños de datos. Existen una gran cantidad de empresas que ofrecen herramientas de enmascaramiento y encriptación, aunque todas requieren de un proceso de personalización para cubrir las necesidades de cada cliente. En este proyecto se llevará a cabo una automatización de las preferencias de enmascaramiento y encriptación para así agilizar la configuración de una herramienta de enmascaramiento que se seleccionará haciendo un análisis de mercado de las empresas y sus herramientas más competentes en la actualidad.

Palabras clave —Enmascaramiento de datos (dinámico/estático), Automatización de procesos y preferencias, Anonimización, Seudonimización, Descubrimiento de datos.

Abstract — Data masking is a data transformation process aimed at protection, where maintaining the realism of the data is crucial. It is essential for safeguarding privacy, preventing data theft, complying with regulations, ensuring security in development and testing environments, and enabling secure analytics in BI environments. Currently, it is a highly valuable and useful process for companies within the technology sector or anyone dealing with large data volumes. There are numerous companies offering masking and encryption tools, although all require customization to meet each client's needs. In this project an automation of the masking and encryption preferences will be carried out to streamline the configuration of a masking tool, which will be selected through a market analysis of the most competent companies and tools available today.

Keywords — Data masking (dynamic/static), automation of processes and preferences, anonymization, pseudonymization, Data discovery.

1. INTRODUCCIÓN

En el contexto actual de la tecnología y la seguridad de la información, la protección de datos sensibles es una prioridad indiscutible para las organizaciones en todos los sectores. NTTDATA, como una empresa líder en soluciones tecnológicas, reconoce la importancia crítica de salvaguardar la confidencialidad y la integridad de la información que gestionan sus clientes como dictan las leyes de protección de datos, como la RGPD (Reglamento General de Protección de Datos) o la LOPD (Ley Orgánica de Protección de Datos). En respuesta a esta necesidad, surge el proyecto de desarrollar una prueba de concepto (PoC) de una herramienta de enmascaramiento de datos, que incluye la implementación de un proceso automatizado para mejorar la eficiencia y precisión en la protección de datos sensibles.

El Data Masking se trata de una técnica crucial en la protección de la información sensible. Su objetivo principal es proteger la privacidad de los datos al ocultar o modificar elementos específicos dentro de conjuntos de datos, manteniendo al mismo tiempo la utilidad de estos para ciertos fines operativos, como pruebas de desarrollo, análisis o demostraciones. Se emplea para asegurar que los datos confidenciales, como números de tarjetas de crédito, información personal identificable o datos financieros sensibles, no sean visibles o no estén accesibles para aquellos que no necesitan verlos o manipularlos directamente mediante diferentes técnicas.

A parte de ayudar a cumplir con las leyes de protección de datos otorga beneficios como:

E-mail: 1493499@uab.cat

Tutor Universidad: Carlos García Calvo (Departamento de Ciencias de la Computación)

Tutores Empresa: Pol Rojas Bartomeus

Mención y Curso: Computación. Curso 2023/24

E-mail de respaldo: kotnikjakob72@gmail.com

- Obtención de datos relacionales, estables y fidedignos.
- Eficacia con menos cantidad de registros.
- Protección de la información.
- Reducción de los costes seguridad.
- Rapidez en la generación de pruebas.
- Rapidez de procesamiento.
- Agilidad del proceso y calidad del desarrollo.
- Extracción flexible, a partir de un filtrado.
- Las distintas técnicas permiten aplicarlo tanto en entornos productivos como en no productivos o de pruebas.

2. TÉCNICAS DE ENMASCARAMIENTO DE DATOS

Hay distintos tipos de data masking y se pueden diferenciar por la reversibilidad de los datos después de aplicarlos:

- Anonimización: consiste en desvincular completamente los datos personales de los datos identificativos, es decir, cuando se anonimizan datos personales, se produce un nuevo conjunto de datos completamente disociado del individuo al que pertenecen, haciendo imposible que a través de esos datos anonimizados se pueda identificar o reidentificar a dicho individuo.
- Seudononimización: consiste en tratar los datos personales sin los datos identificativos del interesado, pero sin suprimir la vinculación entre los datos que consigan determinar la persona titular de los mismos.

O por el momento en el que se aplican:

- Persistent Data Masking: Se aplica enmascaramiento directamente en la base de datos antes de que se acceda a los datos (se suele aplicar en un repositorio clonado de la BD para no corromper la original).
- Dynamic Data Masking: Se aplica enmascaramiento en el momento que se acceden los datos, es decir, enmascaramiento a tiempo real personalizado que aplica las reglas correspondientes al usuario que solicita los datos según su rol.

Por otro lado, la Encriptación es una técnica de seguridad fundamental que convierte datos en un formato ilegible o cifrado mediante el uso de algoritmos y claves. El objetivo principal es proteger la confidencialidad de la información, asegurando que solo aquellos con las claves adecuadas puedan acceder a los datos en su formato legible original. También se podría considerar una técnica de enmascaramiento de datos, aunque las principales diferencias entre ellos son:

- El enmascaramiento de datos se utiliza principalmente para proteger campos de datos específicos dentro de un registro o conjunto de datos y se enfoca en preservar la utilidad y el formato mientras se ocultan detalles sensibles.
- La encriptación se utiliza para proteger un registro o conjunto de datos completo, cifra completamente los datos, proporcionando seguridad hermética, aunque inutilizando los datos si no se poseen las claves necesarias.

3. OBJETIVOS

Esta PoC no solo demostrará la eficacia y viabilidad de las técnicas de enmascaramiento y encriptación en un entorno empresarial real, sino que se pondrá un énfasis especial en el diseño e implementación de un proceso automatizado que garantice la eficiencia y flexibilidad necesarias para cumplir con los requisitos específicos de NTTDATA.

Partiendo de la base de que el ámbito de trabajo es en el sector de seguridad del dato el cual es uno de los pilares indispensables de las compañías para lograr una correcta gestión de los datos. Ofrece las medidas de defensa para impedir los accesos no autorizados, la pérdida o robo y/o la corrupción de los datos a lo largo de todo su ciclo de vida.

Por lo tanto, este proyecto tiene como objetivo ofrecer:

- Protección de la información: Ayuda a las empresas a tener su información protegida contra accesos no autorizados, evitando así filtraciones o pérdidas de datos o incluso ataques cibernéticos que pueden afectar a la integridad y reputación de la empresa.
- Cumplimiento normativo: Asegura el cumplimiento de regulaciones y estándares legales relacionados con la protección de datos, minimizando el riesgo de multas por incumplimiento con los gastos económicos que conlleva
- Confianza del cliente: Cuando los clientes saben que sus datos están seguros y protegidos, tienen más confianza en la empresa, lo que puede llevar a relaciones comerciales más sólidas y una reputación positiva en el mercado.

- Integridad: Es la garantía de que los datos no serán manipulados, modificados o alterados. Para ello, aparte de limitar el acceso a personas o sistemas autorizados, se protegerán mediante técnicas de cifrado.
- Confidencialidad: La información solo debe ser accesible a aquellas personas o sistemas autorizados, de manera que se protegerá mediante procedimientos de identificación y autenticación de usuarios, controles de acceso físico y lógico y contraseñas robustas.
- Disponibilidad: Es la capacidad de poder acceder a los datos en cualquier momento que sea necesario, siempre que se esté autorizado para ello. Se puede garantizar mediante la creación de copias de seguridad y respaldo de las bases de datos y la gestión de accesos.

El resultado final de este proyecto será la generación de un manual de uso detallado que proporcione instrucciones claras sobre cómo utilizar correctamente la herramienta de enmascaramiento y encriptación, junto con el proceso automatizado. Esto asegurará una correcta implementación y maximizará la protección de datos sensibles en NTTDATA, demostrando así el valor agregado que esta solución puede aportar a la organización.

4. METODOLOGÍA

Para conseguir los objetivos establecidos, se puede seguir una metodología basada en (cascade, agile) pasos secuenciales que abarque desde la identificación de necesidades y requisitos hasta la evaluación de la herramienta en entornos empresariales reales. Más detalladamente se seguirán estos puntos:

1. Análisis de necesidades y requisitos: Analizar la normativa aplicable (por ejemplo, GDPR) y los estándares de seguridad de datos relevantes para identificar los requisitos legales y regulatorios.
2. Análisis de mercado y selección de herramientas: Realizar un análisis exhaustivo del mercado de herramientas de enmascaramiento y encriptación de datos y evaluar las herramientas disponibles en función de los requisitos identificados y seleccionar la más adecuada para la implementación en NTTDATA.
3. Construcción del entorno de desarrollo: Configurar un entorno de desarrollo que reproduzca fielmente el entorno de producción de NTTDATA, incluyendo una base de datos adecuada para los casos de uso previstos.
4. Implementación y desarrollo de reglas de la herramienta de Data masking: Implementar la herramienta seleccionada y desarrollar reglas de enmascaramiento y encriptación personalizadas para cumplir con los requisitos específicos de NTTDATA.
5. Pruebas exhaustivas herramienta Data Masking: Realizar pruebas exhaustivas para garantizar la eficacia y precisión de la herramienta de enmascaramiento, así como del proceso de automatización. Si se detecta cualquier fallo o anomalía durante las pruebas corregir el problema.
6. Diseño e implementación de la herramienta de automatización: Diseñar e implementar un robot o proceso automatizado que pueda gestionar el enmascaramiento y la encriptación de datos de manera eficiente y personalizada. Garantizar que la sea flexible y adaptable a los cambios en los requisitos o procesos de NTTDATA.
7. Evaluación en el entorno de desarrollo: Implementar la herramienta y el proceso de automatización en el entorno de desarrollo empresarial creado y evaluar la viabilidad y la eficacia de la herramienta.
8. Generación de manual de uso: Elaborar un manual de uso detallado que proporcione instrucciones claras y concisas sobre cómo utilizar correctamente la herramienta de enmascaramiento y encriptación, junto con el proceso de automatización. Incluir ejemplos prácticos, consejos y mejores prácticas para maximizar la eficacia y la seguridad en el uso de la herramienta.

4. PLANIFICACIÓN

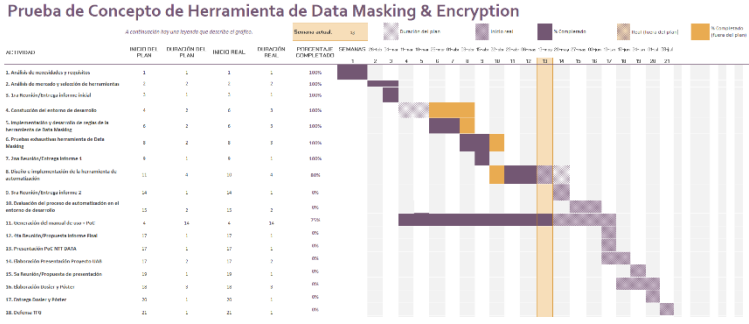


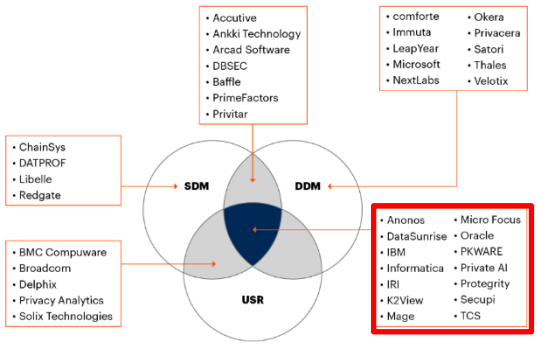
Figura 1: Diagrama de Gantt planificación
Ver anexo para mejor visualización

5. ANÁLISIS DE MERCADO

En este análisis de mercado se han estudiado las diferentes empresas que cubren todos los tipos de

enmascaramiento según el informe de gartner [3] (ver anexo):

Figura 2: Esquema de Gartner



6. ELECCIÓN

Después de analizar detenidamente todas las opciones presentadas, es evidente que todas ofrecen servicios muy similares y amplios en cuanto a enmascaramiento de datos y protección de la privacidad. Sin embargo, se ha concluido que SecuPi es la herramienta más adecuada para la empresa varias razones clave que se detallan a continuación:

- **Poca Experiencia Previa**: En primer lugar, la empresa ya ha tenido una experiencia positiva trabajando con SecuPi pero aún no ha explorado todas sus posibilidades y quiere contrastar su potencial con el de otras empresas con las que ya ha trabajado más profundamente como Informatica o Oracle. La prueba de concepto y sus mejoras permitirán explorar mejor su usos y estudiar las ventajas sobre las otras herramientas.

- **Reputación Sólida en el Mercado**: Además, SecuPi cuenta con una sólida y establecida reputación en el mercado de seguridad de datos. Su reconocimiento como proveedor confiable y líder en el sector brinda la confianza de que se está invirtiendo en un producto y servicio de alta calidad. Las numerosas referencias positivas y casos de éxito documentados fortalecen aún más la decisión, asegurando que la herramienta cumple con los estándares más exigentes de seguridad y privacidad.

- **Compatibilidad y Recomendación Interna:** Un factor adicional que ha influido en la selección de SecuPi es la recomendación interna basada en criterios operacionales y estratégicos. La empresa ha evaluado detenidamente las opciones y ha considerado que, dada la infraestructura tecnológica actual y los requisitos específicos de los proyectos, SecuPi es la solución que mejor se alinea con los objetivos a largo plazo. Este respaldo interno no solo subraya la confianza en la herramienta, sino que también se entiende como un mandato implícito para elegir esta solución sobre otras alternativas.

En resumen, la decisión de optar por SecuPi se fundamenta en la experiencia previa positiva, su reputación establecida en el mercado y la recomendación estratégica interna de la empresa. Estos factores combinados hacen que SecuPi sea la opción más lógica y eficiente para las necesidades de enmascaramiento de datos y protección de la privacidad.

7. SECUPI

SecuPi ofrece una herramienta avanzada de enmascaramiento de datos que utiliza técnicas sofisticadas para proteger la privacidad y la integridad de la información sensible. Su objetivo principal es garantizar que los datos confidenciales, como números de tarjetas de crédito, información personal identificable (PII) y datos financieros sensibles, no sean visibles o accesibles para aquellos que no tienen la autorización adecuada. Además, se integra fácilmente con los sistemas y aplicaciones existentes de la organización, lo que facilita su implementación y uso en entornos empresariales complejos. Esta herramienta nos permite cumplir con algunos de nuestros objetivos:

- Protección de la privacidad: proporciona una capa adicional de seguridad al ocultar datos sensibles, evitando así la exposición no autorizada de información confidencial.
- Cumplimiento normativo: La herramienta ayuda a la organización a cumplir con las regulaciones de protección de datos, como el Reglamento General de Protección de Datos (GDPR), al garantizar que los datos sensibles estén protegidos adecuadamente.
- Minimización de riesgos: Al enmascarar datos sensibles, ayuda a reducir los riesgos de violaciones de datos y fugas de información, protegiendo la reputación y la integridad de la organización.
- Facilitar el desarrollo y las pruebas: SecuPI permite a los equipos de desarrollo y pruebas trabajar con conjuntos de datos realistas y seguros, sin exponer información confidencial a riesgos durante el proceso de desarrollo de software.

Para ello usa la siguiente arquitectura con:

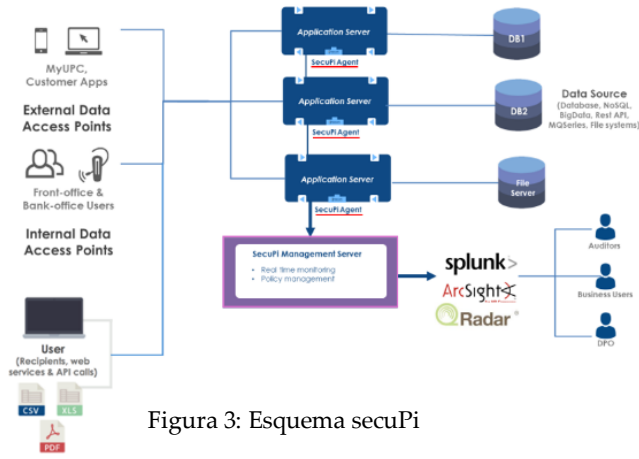


Figura 3: Esquema secuPi

- SecuPi Central Managment Server que incluye un repositorio central de datos y contiene todas las configuraciones requeridas por los agentes en tiempo de ejecución (configuraciones técnicas como claves de cifrado, así como políticas para aplicar a los datos).

- Agentes instalados en los servidores de aplicaciones (por ejemplo, WebSphere, Weblogic, Tomcat, JBoss) o en herramientas de acceso directo a la base de datos. Interceptan las solicitudes de datos relevantes y se encargan de extraer del Server SecuPi las políticas de respuesta que se aplicarán (en un intervalo de tiempo predefinido).

8. ENTORNO DE DESARROLLO

La herramienta, tal como está acoplada en el entorno empresarial, requiere el uso de SQL Server Management Studio para crear una base de datos en un entorno seguro usando un escritorio remoto que se encuentra en una máquina virtual usada por la empresa expresamente para este tipo de pruebas de desarrollo. Como se aprecia en la planificación ha habido cierto retraso en este punto ya que ha habido bastantes problemas para conseguir los permisos y certificados necesarios para obtener acceso al escritorio remoto debido a mi situación como student y la seguridad de este tipo de entornos empresariales en una empresa de la envergadura de NTTDATA.

Finalmente, al conseguir todo lo necesario he procedido a crear una base de datos con 4 tablas para simular un caso de uso real:

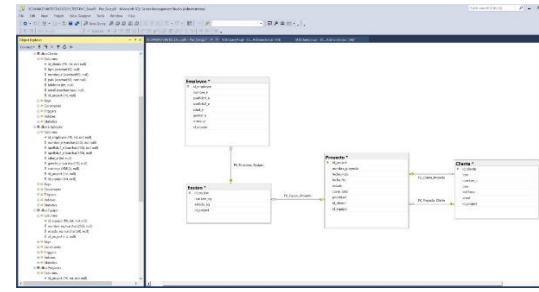


Figura 4: Diagrama Base de Datos de pruebas
Ver anexo para mejor visualización

Estas las he rellenado con conjuntos de datos no reales para empezar a hacer pruebas de la herramienta con ellos:

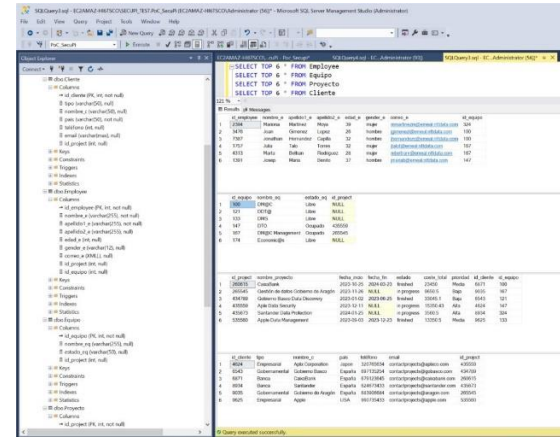


Figura 5: Contenido BD
Ver anexo para mejor visualización

9. PRUEBAS Y CONFIGURACIÓN DE LA HERRAMIENTA

En este apartado se estudia la herramienta y se hacen las pruebas correspondientes para comprobar que todo lo necesario funciona para seguir adelante con el proyecto. No se ha añadido en el informe por falta de espacio (Ver anexo).

10. PROCESO DE AUTOMATIZACIÓN

El proceso de automatización comprende tres etapas principales: la creación de una herramienta de descubrimiento de datos, el desarrollo de un recomendador de reglas de enmascaramiento basado en un árbol de decisiones y la creación de un fichero de configuración de SecuPi. Estas etapas están diseñadas para abordar la necesidad de identificar y clasificar diferentes tipos de datos presentes en conjuntos de datos, con el objetivo final de proporcionar recomendaciones para aplicar técnicas de enmascaramiento adecuadas. El proceso incluye la configuración automatizada de la

herramienta SecuPi mediante un archivo JSON para garantizar que las reglas de reescritura de datos, los mapeos de formatos y el acceso a los recursos se ajusten automáticamente a los datos del usuario.

1. **Etapa 1: Desarrollo de la Herramienta de Descubrimiento de Datos**

   La segunda etapa del proceso implica el desarrollo de una herramienta de descubrimiento de datos. Esta herramienta tiene como objetivo principal leer conjuntos de datos y clasificar los datos dentro de los tipos de datos previamente definidos. Inicialmente, se ha intentado implementar esta herramienta utilizando un algoritmo propio basado en Random Forest. Sin embargo, aunque el algoritmo funciona, los resultados obtenidos hasta el momento han sido insatisfactorios, lo que indica la necesidad de ajustar y mejorar el modelo o optar por otro camino. El proceso de desarrollo de la herramienta implica:

- Extracción y procesamiento de datos: La herramienta lee los conjuntos de datos proporcionados como entrada y realiza la extracción de datos para identificar patrones y estructuras.

- Clasificación de tipos de datos: Utilizando algoritmos de clasificación, la herramienta clasifica los datos identificados en los tipos de datos previamente definidos, como NOMBRE_PERSONA, IDENTIFICADOR, CODIGO POSTAL, CUENTA BANCARIA, TARJETA, DIRECCION, EDAD, EMAIL, FECHA, LUGAR y TELEFONO

  Integración con el recomendador: Una vez que los tipos de datos han sido identificados y clasificados, la herramienta pasa esta información al recomendador de reglas de enmascaramiento desarrollado en la etapa anterior. El recomendador proporciona recomendaciones para aplicar técnicas de enmascaramiento específicas a cada tipo de dato.

2. **Etapa 2: Desarrollo del Recomendador de Reglas de Enmascaramiento**

   El primer paso en el proceso de automatización implica el desarrollo de un recomendador de reglas de enmascaramiento. Este recomendador se basa en un árbol de decisiones, que es un modelo predictivo que utiliza un conjunto de reglas para inferir la clasificación de datos de entrada. El objetivo principal de este recomendador es leer los tipos de datos presentes en un conjunto de datos y recomendar la técnica de enmascaramiento más adecuada para cada tipo de dato.

   El proceso de desarrollo del recomendador implica:

- Definición de tipos de datos: Se identifican los tipos de datos relevantes que se pretenden enmascarar, como nombres de personas, identificadores, códigos postales, cuentas bancarias, tarjetas, direcciones, edades, correos electrónicos, fechas, lugares y números de teléfono.

- Entrenamiento del modelo: Se utiliza un conjunto de datos de entrenamiento que contiene ejemplos de cada tipo de dato junto con la técnica de enmascaramiento aplicada. Este conjunto de datos se utiliza para entrenar el árbol de decisiones, que aprende patrones y relaciones entre los tipos de datos y las técnicas de enmascaramiento.

- Evaluación del modelo: Se evalúa el rendimiento del modelo utilizando un conjunto de datos de prueba separado. Esto ayuda a garantizar que el recomendador pueda generalizar y proporcionar recomendaciones precisas para tipos de datos no vistos durante el entrenamiento.

3. **Etapa 3: Configuración Automatizada de SecuPi**

   Al finalizar todo el proceso, se procederá a modificar un archivo JSON de configuración de la herramienta SecuPi. Este archivo JSON permitirá que, al cargarse en la aplicación, se configure automáticamente de manera que los data rewrites (reescritura de datos), los format mappings (mapeos de formato) y el resource access (acceso a recursos) queden creados y ajustados automáticamente a los datos que tenga el usuario. Esto asegurará una integración fluida y eficiente de las reglas de enmascaramiento recomendadas en el entorno del usuario.

Flujo de Trabajo

El flujo de trabajo comienza con el proceso de descubrimiento de datos, donde la herramienta de descubrimiento analiza y clasifica los datos presentes en un conjunto de datos dado. Una vez que se han identificado los tipos de datos, esta información se pasa al recomendador de reglas de enmascaramiento, que proporciona recomendaciones sobre las técnicas de enmascaramiento más adecuadas para cada tipo de dato. Finalmente, se realiza la configuración automatizada de SecuPi mediante la modificación de un archivo JSON que asegura que las reglas y

mapeos se ajusten automáticamente a los datos del usuario, facilitando una implementación eficiente y efectiva de las técnicas de enmascaramiento.

11. RESULTADOS INICIALES

- Recomendador de Reglas de Enmascaramiento: Inicialmente, el recomendador de reglas de enmascaramiento ha mostrado un funcionamiento muy bueno. Dado que se basa en un árbol de decisiones, el proceso es bastante sencillo y escalable. El árbol de decisiones, al ser un modelo predictivo simple y eficiente, permite hacer recomendaciones precisas y rápidas. En esencia, el recomendador sugiere la técnica de enmascaramiento que aparece con mayor frecuencia como opción para el tipo de dato específico que se está analizando. Esta metodología ha demostrado ser efectiva, asegurando que las recomendaciones sean confiables y consistentes.

	DATA TYPE	MASKING RULE
2	CUENTA BANCARIA	
3	TARJETA	
4	DIRECCION	
5	EDAD	
6	GENERO	
7	EMAIL	
8	FECHA	
9	NOMBRE APELLIDO	
10	POBLACION	
11	PAIS	
12	PROVINCIA	
13	TELEFONO	

	DATA TYPE	MASKING RULE
2	CUENTA BANCARIA	encrypt_CUENTA TARJETA
3	TARJETA	encrypt_TARJETA
4	DIRECCION	encrypt_DIRECCION
5	EDAD	encrypt_EDAD
6	GENERO	encrypt_GENERO (crear)
7	EMAIL	encrypt_EMAIL
8	FECHA	encrypt_DATE
9	NOMBRE APELLIDO	encrypt_NOMBRE APELLIDO
10	POBLACION	encrypt_LUGAR
11	PAIS	encrypt_LUGAR
12	PROVINCIA	encrypt_LUGAR
13	TELEFONO	encrypt_TELEFONO

- Herramienta de Descubrimiento de Datos: En contraste, el desarrollo de la herramienta de descubrimiento de datos ha sido más desafiante. La implementación inicial utilizando un algoritmo propio basado en Random Forest no ha proporcionado los resultados esperados. El algoritmo Random Forest, aunque es conocido por su capacidad para manejar una variedad de tipos de datos y realizar clasificaciones precisas, ha mostrado dificultades significativas en este contexto. En particular, ha cometido numerosos errores al reconocer y clasificar correctamente los datos numéricos.

```
INPUT: 'Datos test 1': ['jakob
kotnik','Passeig de les roses, 40,
Castellar del valles, España,
55,'SAMANTiago@GMAIL.COM','687125366','33
3654783','08211','ZXCH789012345283401234'
,'4447893224451234','687125366',
'26/01/1999']
```

```
OUTPUT: 'Datos test 1':
array(['TELEFONO', 'DIRECCION', 'EDAD',
'EMAIL', 'TELEFONO',
'TELEFONO','TELEFONO', 'TELEFONO',
'TELEFONO', 'TELEFONO', 'FECHA']
```

Los problemas principales observados incluyen:

1. Errores de Clasificación: El algoritmo tiene dificultades para distinguir entre diferentes tipos de datos especialmente los datos numéricos, como cuentas bancarias, tarjetas de crédito y códigos postales. Estos errores de clasificación impactan negativamente en la precisión general de la herramienta.

2. Underfitting: Se ha observado el modelo no se ajusta bien a los datos de entrenamiento ni a los datos de prueba, lo que reduce su capacidad para generalizar y clasificar correctamente nuevos datos.

Para poder visualizar estos resultados de manera más precisa y clara he realizado pruebas de validación cruzada tanto para los datos de Train como para los datos de Test:

Traindata:

```
Mean accuracy score from cross-validation: 0.7545454545454545

Individual accuracy scores:
Fold 1: 0.7636363636363637
Fold 2: 0.740909090909091
Fold 3: 0.7909090909090909
Fold 4: 0.6909090909090909
Fold 5: 0.7863636363636364
```

Testdata:

```
Mean accuracy score from cross-validation: 0.5581818181818182

Individual accuracy scores:
Fold 1: 0.5181818181818182
Fold 2: 0.6090909090909091
Fold 3: 0.5545454545454546
Fold 4: 0.6
Fold 5: 0.509090909090909
```

Para entender mejor estos resultados también he querido mostrar las matrices de confusión de cada set de datos:

Figura 6: Normalized Average Confusion Matrix (Train Data)

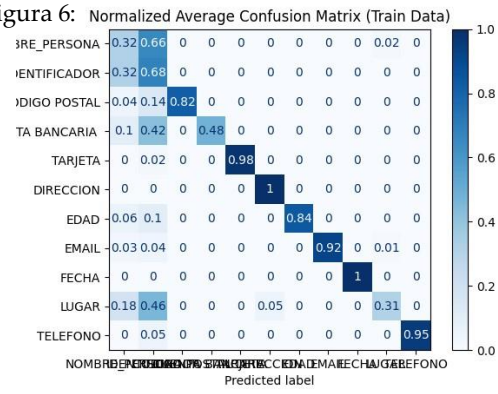
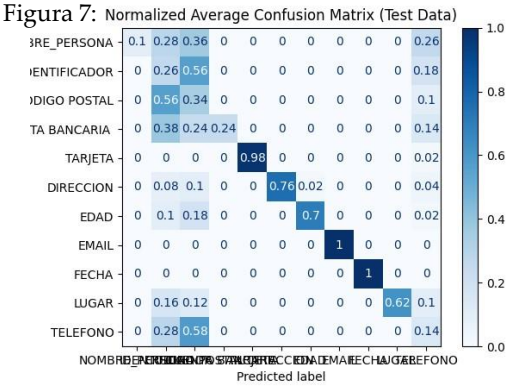


Figura 7: Normalized Average Confusion Matrix (Test Data)



Quiero destacar que, para llevar a cabo las pruebas de entrenamiento y validación del modelo presentado, he generado conjuntos de datos tanto de entrenamiento como de prueba. Estos conjuntos fueron creados internamente, utilizando datos que diseñé específicamente para el propósito de este estudio. La razón detrás de esta acción radica en la necesidad de contar con datos controlados y específicos que reflejen con precisión el problema que se está abordando. Al crear nuestros propios conjuntos de datos, se tiene la capacidad de adaptarlos a las características y requisitos del modelo que estamos evaluando, asegurando así que el rendimiento del modelo sea evaluado en condiciones realistas y relevantes para nuestra aplicación.

Por otra parte, la generación de conjuntos de datos propios nos permite poder hacer cualquier tipo de prueba sobre datos ficticios que no suponen un riesgo para la empresa NTTDATA, manteniendo así un entorno de desarrollo seguro.

Próximos Pasos

- Revisar y Optimizar el Modelo: Ajustar los parámetros del Random Forest y experimentar con otras técnicas de preprocesamiento de datos para mejorar la clasificación de datos numéricos

- Explorar Otros Algoritmos: Considerar el uso de otros algoritmos de clasificación, como redes neuronales o modelos preentrenados, que podrían ofrecer mejores resultados para la clasificación de los datos

- Validación y Evaluación Continua: Implementar un proceso de validación cruzada más riguroso para evaluar el rendimiento del modelo y asegurarse de obtener unos resultados decentes.

12. RESULTADOS FINALES: IMPLEMENTACIÓN CON API DE AZURE OPENAI

La implementación con la API de OpenAI en Azure ha resultado en mejoras significativas en los resultados y el funcionamiento de la integración. Esto se debe a varios factores clave:

1. Modelos Preentrenados Avanzados: La API de OpenAI proporciona acceso a modelos de lenguaje natural preentrenados avanzados, como GPT-3.5, que están diseñados para comprender y generar texto con gran precisión. Estos modelos han demostrado tener una capacidad superior para clasificar datos y comprender el contexto, lo que ha llevado a una mejora general en la precisión de la clasificación de datos.

2. Mayor Capacidad de Generalización: Los modelos preentrenados de OpenAI están entrenados en grandes conjuntos de datos que abarcan una amplia gama de dominios y temas. Esto les permite generalizar mejor y adaptarse a diferentes tipos de datos y contextos, lo que resulta en una mayor precisión y coherencia en la clasificación de datos.

Debido al uso de herramientas de IA avanzada, se ha tenido que solicitar y obtener permisos para acceder a la API de OpenAI en Azure. Esto ha implicado un proceso de trabajo para cumplir con los requisitos y procedimientos necesarios para garantizar un uso ético y responsable de estas herramientas. A pesar de cumplir con todos los requisitos es importante señalar que, finalmente, no se pudieron conseguir los permisos necesarios para usar la herramienta directamente. Debido a esto, las pruebas se realizaron de manera externa, utilizando el dispositivo de un amigo que contaba con los permisos y la suscripción activa para utilizar la API

de OpenAI en Azure. Esta limitación implicó que no pudiéramos evaluar la herramienta con grandes conjuntos de datos y se llevaron a cabo pruebas con las siguientes listas de datos reducidas:

```
Name = ['jaume', 'jakob', 'miguel',
'William', 'MOHAMED', 'GUILLERMO']
Phone = ['687125366', '+34 627 899 125',
'937 145 208', '756789012', '901234567',
'987654321']
Address = ['Calle del Prado, 55, Madrid,
España','Rue du Faubourg Saint-Honoré, 48,
París, Francia','Avenida Libertador, 1234,
Buenos Aires, Argentina','Avenida
Ámsterdam, 150, Ciudad de México,
México','Avenida Atlântica, 1702, Río de
Janeiro, Brasil','Avenue des Ternes, 80,
París, Francia']
Bank_Account=['ES9121000418450200051332','E
S2321000418450200051333','ES762100041845020
0051334','ES9121033418450235651332','ES7261
000418450204555332','ES91210042351672000513
32']
CIF = ['B12345678', 'C87654321',
'A11223344', 'B58023454', 'B25386658',
'A19203645']
C_Postal = ['08211', '06435', '09132',
'08313', '05672', '04022']
Credit_Card = ['5432 6789 0123 4567','4567
8901 2345 6789','9012 3456 7890 1234','4567
8901 2345 6789','5678 9012 3456 7890','8901
2345 6789 0123']
Age = ['28', '33', '22', '45', '34', '55']
ID = ['ID46198', 'ID13579', 'ID38671',
'ID75940', 'ID29463', 'ID24587']
Mail=['christopher.moore@fastmail.com','jam
es.clark@yopmail.com','stephanie.lewis@outl
ook.es','emily.smith@yahoo.com','john.doe@g
mail.com','david.jones@hotmail.com']
Date_Birth = ['28/09/2000', '05/10/1980',
'25/01/2008', '02/07/2006', '26/08/1970',
'14/09/1998']
City = ['París', 'Valencia', 'Caracas',
'Campinas', 'Girona', 'Hospitalet de
Llobregat']
Countries = ['Andorra', 'China', 'Japón',
'Estados Unidos', 'Alemania', 'Argentina']
```

Prompt usado:

```
Eres un modelo de clasificación que clasifica los string de texto en una de las siguientes categorías:
- NOMBRE Ejemplo: JAUME
- IDENTIFICADOR Ejemplo: ID90624
- CODIGO POSTAL Ejemplo: 08211
- CUENTA BANCARIA Ejemplo: ES9420805801101234567891
- TARJETA Ejemplo: 5432 6789 0123 4567
- DIRECCION Ejemplo: Carrer del Mar, 10, Barcelona, España
- EDAD Ejemplo: 52. La edad siempre tiene que estar entre 18-65 años
- EMAIL Ejemplo: jaumekotnikbetriu@gmail.com
- FECHA Ejemplo: Ejemplo: 20/11/1972 Ejemplo: 1999/03/12 Ejemplo: 11-10-2015
- TELEFONO Ejemplo 1: +34 687 125 366, Ejemplo 2: 687125366
- LUGAR Ejemplo: Barcelona
- PAIS Ejemplo: España
- CIF Ejemplo: B76365789
Nota: Responde solo con el nombre de la categoría. No incluyas explicaciones ni disculpas en tus respuestas.
    String a clasificar: {{$input}}
    Respuesta:
```

Los resultados obtenidos con la implementación de la API de OpenAI en Azure han sido muy prometedores, mostrando una notable precisión en la clasificación de datos en la mayoría de las categorías. Como el objetivo principal es el descubrimiento preciso de datos, para mejorar los resultados, se implementa un enfoque adicional de cálculo de la moda de las predicciones para cada columna en la base de datos (se muestra en las Keys de los diccionarios interiores siguientes). Este método nos permite trabajar más fácilmente con los resultados obtenidos en las otras fases del proyecto.

```
results = {
    "Name": {"NOMBRE": ["NOMBRE", "NOMBRE",
"NOMBRE", "NOMBRE", "NOMBRE", "NOMBRE"]},
    "Phone": {"TELEFONO": ["TELEFONO",
"TELEFONO", "TELEFONO", "IDENTIFICADOR",
"TELEFONO", "IDENTIFICADOR"]},
    "Address": {"DIRECCION": ["DIRECCION",
"DIRECCION", "DIRECCION", "DIRECCION",
"DIRECCION", "DIRECCION"]},
    "Bank_Account": {"CUENTA BANCARIA":
["CUENTA BANCARIA", "CUENTA BANCARIA",
```

```
"CUENTA BANCARIA", "CUENTA BANCARIA",
"CUENTA BANCARIA", "CUENTA BANCARIA"]},
    "CIF": {"CIF": ["IDENTIFICADOR",
"IDENTIFICADOR", "IDENTIFICADOR",
"IDENTIFICADOR", "CIF", "IDENTIFICADOR"]},
    "C_Postal": {"CODIGO POSTAL": ["CODIGO
POSTAL", "CODIGO POSTAL", "CODIGO POSTAL",
"CODIGO POSTAL", "CODIGO POSTAL", "CODIGO
POSTAL"]},
    "Credit_Card": {"TARJETAS": ["TARJETA",
"TARJETA", "TARJETA", "TARJETA", "TARJETA",
"TARJETA"]},
    "Age": {"EDAD": ["EDAD", "EDAD",
"EDAD", "EDAD", "EDAD", "EDAD"]},
    "ID": {"IDENTIFICADOR":
["IDENTIFICADOR", "IDENTIFICADOR",
"IDENTIFICADOR", "IDENTIFICADOR",
"IDENTIFICADOR", "IDENTIFICADOR"]},
    "Mail": {"EMAIL": ["EMAIL", "EMAIL",
"EMAIL", "EMAIL", "EMAIL", "EMAIL"]},
    "Date_Birth": {"FECHA": ["FECHA",
"FECHA", "FECHA", "FECHA", "FECHA",
"FECHA"]},
    "City": {"LUGAR": ["LUGAR", "LUGAR",
"LUGAR", "LUGAR", "LUGAR", "LUGAR"]},
    "Countries": {"PAIS": ["LUGAR", "PAIS",
"PAIS", "PAIS", "PAIS", "PAIS"]} }
```


Como se puede observar, el modelo presenta un alto grado de precisión, pero aún persisten ciertos errores de clasificación, especialmente en los campos CIF y Teléfono. Es importante destacar que estos errores ocurren utilizando un prompt relativamente simple. Con la redacción de prompts más específicos y detallados, es posible mejorar significativamente la precisión en la clasificación de estos tipos de datos ya que, proporcionando más ejemplos y especificaciones contextualmente relevantes, podemos guiar mejor el proceso de aprendizaje y el modelo sería capaz de ir mejorando continuamente. Si se quiere ver el resultado de la configuración final de Secupi con el archivo JSON generado con los resultados del descubrimiento de datos y el recomendador de reglas se puede ver en el anexo (apartado resultados)

13. CONCLUSIONES

El proyecto de implementación de una herramienta de enmascaramiento de datos en NTTDATA, junto con el desarrollo de un proceso de automatización, ha concluido con resultados notablemente positivos, destacando en varias áreas clave que reflejan el éxito de las etapas integradas de automatización.

Resultados Exitosos de la Automatización: La integración de las tres etapas principales del proceso de automatización — desarrollo del recomendador de reglas de enmascaramiento, creación de una herramienta de descubrimiento de datos, y configuración automatizada de la herramienta SecuPi — ha acabado siendo un éxito. El sistema ha demostrado su capacidad para clasificar y enmascarar adecuadamente los tipos de datos, proporcionando soluciones personalizadas y seguras que se ajustan a las necesidades específicas de la empresa. Gracias al proyecto se han creado 2 funcionalidades de gran utilidad que la herramienta Secupi no ofrece.

Implementación Efectiva y Eficiente: La configuración automatizada de la herramienta SecuPi ha permitido que los ajustes necesarios en la reescritura de datos, los mapeos de formato y el acceso a los recursos se realicen con gran precisión y poca complejidad, lo que resulta en una integración eficiente y efectiva del entorno de usuario de SecuPi.

Potencial para Mejoras Futuras: Con los resultados obtenidos y la infraestructura establecida, el proyecto abre la puerta a futuras innovaciones en la gestión y protección de datos. La capacidad de adaptarse y escalar las soluciones permitirá la implementación de tecnologías de protección de datos avanzadas y responder eficazmente a los cambiantes requisitos legales y técnicos de los clientes/usuarios que utilicen la herramienta. Por otro lado, por mucho que se hayan usado estas listas reducidas el tiempo de ejecución para resolverlas ha sido de 3min y 30s. Una de las posibles mejoras futuras es obtener los permisos para usar la API de manera completa y explorar la posibilidad de cargar listas enteras, de modo que no sea necesario realizar una llamada a la API por cada elemento de las listas, sino una sola vez por lista.

En conclusión, aunque hemos logrado avances significativos con el uso de la API de OpenAI en Azure, queda claro que existe un potencial considerable para optimizar aún más los resultados. Se espera que, con la implementación de prompts más específicos, la adaptación continua del modelo y la capacidad de procesar grandes conjuntos de datos se pueda agilizar y mejorar la precisión de la clasificación de datos.

REFERENCIAS

[1] Data masking: Qué es, tipos, funciones y objetivos | Grupo Atico34 (protecciondatos-lopd.com)
[2] Beneficios del data masking (powerdata.es)
[3] Gartner Reprint
[4] ¿Qué es la ocultación de datos y cuándo puede utilizarse? (powerdmarc.com)
[5] ¿Data Masking o Data Encryption? Como elegir la mejor técnica para cada caso de uso. - Dataspurs
[6] Data Masking, ¿qué es y qué ventajas tiene? - icaria Technology
[7] Data Masking: What It Is, Techniques and Examples | Informatica | Informatica
[8] What is "state of the art" in IT security? — ENISA (europa.eu)
[9] What "State of the Art" in IT Security Will Satisfy European Regulators? | WireWheel
[10] The state of data masking: It's essential—so get it right | TechBeacon
[11] La automatización inteligente, entre las tendencias en transformación digital para 2023 - Big Data Magazine
[12] Data Masking vs. Data Encryption: How Do They Differ? (techtarget.com)

Referencias análisis de mercado

[13] Dynamic Data Masking - SecuPi ,
[14] SecuPi - Get Your Data Protected and Privacy Ready
[15] Data Access Governance - SecuPi
[16] The SecuPi Platform - SecuPi
[17] Database Data Masking Solution | DataSunrise
[18] DataSunrise - SETI & SIDIF Dominicana (ss-d.com.do)
[19] Dynamic Data Masking Solution | DataSunrise
[20] Dynamic Data Masking | DataSunrise Data & DB Security
[21] Anonos: Full-spectrum data security platform to build performant data products
[22] Data pseudonymization solution | Protect sensitive data with a GDPR-compliant data privacy technology (anonos.com)
[23] Anonos: Protected Master Data Management for AI and Analytics
[24] Data Safeguarding: Understanding Masking & Protection | Anonos
[25] Data Masking transformation (informatica.com)
[26] https://www.informatica.com/content/dam/informatica-com/en/collateral/data-sheet/persistent-data-masking_data-sheet_6990.pdf
[27] Cloud Data Masking – Anonymize Data for Trust | Informatica
[28] Creating an Unstructured Data Transformation (informatica.com)
[29] Introduction to Oracle Data Masking and Subsetting
[30] Data Masking | Oracle España
[31] Data Masking (oracle.com)
[32] https://www.oracle.com/es/security/database-security/data-masking/#rc30p6
[33] InfoSphere Optim Data Privacy for Unstructured Data | IBM
[34] Masking data - IBM Documentation
[35] Enmascarar datos con reglas de protección de datos (IBM Knowledge Catalog) - Docs | IBM Cloud Pak for Data as a Service
[36] IBM InfoSphere Optim Data Privacy | IBM
[37] Unstructured Data Masking: More Effective via Business Entities (k2view.com)
[38] Data Masking Tools | K2View
[39] What is Data Masking? Techniques and Best Practices Guide | K2view
[40] Data Masking | K2View Support
[41] Unstructured Data Masking | IRI DarkShield
[42] Data Masking Software (iri.com)
[43] Dark Data Masking | IRI DarkShield
[44] Data Masking as a Service (iri.com)
[45] Data Masking Tools - IRI Data Protector Suite
[46] Data Masking (microfocus.com)
[47] Data Masking Guide (microfocus.com)
[48] Best Data Anonymization tool | Mage Dynamic Data Masking (magedata.ai)
[49] Mage Data : Best Data Security Platform For Enterprises
[50] Data Masking Solutions | PKWARE®
[51] PK Masking Datasheet - PKWARE®
[52] PK Dynamic Masking - PKWARE®
[53] DS_PKMasking_2021.pdf (hubspotusercontent-na1.net)
[54] Private AI | Identify, Redact & Replace PII (private-ai.com)
[55] Products - Text - Private AI (private-ai.com)
[56] The Explainer: Dynamic Data Masking and Monitoring (protegrity.com)
[57] Put on that Mask: Why Protegrity is Offering Dynamic Data Masking
[58] Protegrity Announces Latest Version of Platform with New Dynamic Data Masking Capabilities (datanami.com)
[59] Role of Data Masking & FHE in Safeguarding Customer Privacy (tcs.com)
[60] TCS MasterCraft™ DataPlus for Compliance to Data Privacy Regulations
[61] Enable Data Privacy Using Software-based Approach for Data Masking (tcs.com)