Memoria del proyecto de Machine Learning.

Gestión integral de los residuos documentales en una empresa de Gestión de Residuos Industriales.

Presentado por **Francisco Olivenza Millón** el 20 de mayo de 2025.

1. Introducción

La empresa *Gestión Integral de Residuos Industriales* ha acumulado un extenso fondo documental desde 1980, organizado conforme a estándares archivísticos como ISAD(G). Dada la necesidad de optimizar recursos y espacio digital, surge la necesidad de un sistema de apoyo a decisiones que automatice el tratamiento de sus documentos: conservar, expurgar o digitalizar.

Debido a la sensibilidad inherente a la información documental que manejan las empresas, especialmente en lo que respecta a datos personales, históricos o estratégicos, no fue posible obtener bases de datos reales para el desarrollo de este proyecto. Las organizaciones suelen ser muy celosas con el acceso a su Sistema de Gestión Documental, incluso para fines académicos o experimentales.

Por esta razón, se optó por la **generación de datos sintéticos** que replican la estructura jerárquica y lógica funcional de un archivo institucional real, siguiendo los principios de la norma ISAD(G) y el modelo de descripción multinivel: Fondo → Sección → Serie → Subserie → Expediente → Unidad documental. Esta estructura se complementó con atributos adicionales que no se encuentran en la norma ISAD(G), pero que son fundamentales para un análisis automatizado del tratamiento documental, tales como:

- Porcentaje de datos personales
- Grado de ilegibilidad (estado de conservación)
- Número de usos del documento en su vida útil
- Nivel de importancia (estimado a partir del uso y de la función)

Estos atributos se incorporaron para reflejar no solo la estructura archivística, sino también aspectos normativos, técnicos y de conservación que influyen en la toma de decisiones sobre la eliminación, conservación o digitalización de documentos. En conjunto, esta base de datos sintética busca ofrecer un entorno verosímil y robusto para entrenar modelos de clasificación documental automatizada con base en reglas reales.

Este proyecto implementa una solución basada en aprendizaje automático para predecir el tratamiento más adecuado para cada unidad documental, partiendo de una base de datos sintética generada a partir de reglas archivísticas, funcionales y legales.

2. Problema a solucionar

A medida que el volumen documental crece año tras año, tanto en soporte físico como digital, la empresa *Gestión Integral de Residuos Industriales* se enfrenta a un desafío cada vez mayor: decidir de forma eficiente, segura y justificada qué documentos deben **conservarse**, cuáles pueden **expurgarse** y cuáles requieren algún tipo de **tratamiento digital** (como restauración o digitalización).

La toma de decisiones sobre el tratamiento documental ha estado tradicionalmente en manos de equipos humanos, lo que implica:

- Elevado coste de tiempo y personal.
- Riesgo de errores subjetivos o inconsistencias.
- Dificultad para aplicar políticas archivísticas y legales de forma homogénea.

Enfoque del proyecto:

El proyecto plantea una solución basada en **aprendizaje automático (Machine Learning)**, capaz de:

- Aprender patrones a partir de variables como el número de usos, el estado de conservación, la importancia funcional o la sensibilidad de los datos personales.
- Predecir el tratamiento documental más adecuado para cada unidad documental, de acuerdo con políticas establecidas por la organización.
- Identificar posibles errores o decisiones contradictorias ya tomadas en registros previos (auditoría documental).

Formalización técnica:

Desde el punto de vista del aprendizaje automático, se trata de un problema de **clasificación multiclase supervisada**, donde la variable objetivo es *tratamiento_recomendado,* con tres clases:

- Conservar
- Expurgar
- Tratamiento digital requerido

Además, se contempla el uso de técnicas **no supervisadas** para detectar:

- Documentos que presentan patrones atípicos respecto a lo esperado.
- Posibles errores de etiquetado humano, mediante modelos como **Isolation Forest** .

El modelo debe ajustarse no solo a los datos, sino también a las **reglas archivísticas y criterios institucionales**, tales como:

- Documentos de subseries críticas (p. ej., Junta Directiva o Propiedad Intelectual) no pueden ser expurgados.

- Documentos ilegibles deben considerarse para digitalización o eliminación, según su importancia.
- La presencia de datos personales condiciona el tratamiento conforme a normativas de protección de datos (como el RGPD).

Objetivos específicos:

- Reducir costes de almacenamiento (físico y digital).
- Aumentar la eficiencia de los procesos archivísticos.
- Automatizar decisiones que cumplan con criterios técnicos, legales y archivísticos.
- Proveer una base para auditorías documentales y depuración de decisiones anteriores.

Lo bueno es que la Empresa invirtió en la gestión documental continua de una de sus plantas, de la cual extraeremos la estructura y criterio de división de la clasificación, para aplicarlo en la de otras plantas, hallando el fallo humano y rellenando nulos.

3. Elaboración de los datos

Dado que no se disponía de una base real por motivos de privacidad y confidencialidad, se optó por **generar sintéticamente una base de datos archivística estructurada** desde cero, siguiendo principios y modelos de descripción documental basados en ISAD(G), Records in Contexts y políticas de conservación. Esta generación parte de una lógica jerárquica y funcional, construyendo cada nivel documental desde el **Fondo** hasta las **Unidades Documentales** (UD), con atributos adicionales esenciales para su análisis mediante modelos de machine learning.

3.1. Estructura jerárquica

El sistema documental se construyó siguiendo esta jerarquía:

- **Fondo**: 1 solo fondo principal.
- **Secciones**: Representan grandes áreas funcionales (e.g., Jurídica, Producción, Logística, Administración).
- **Series**: Cada sección tiene sus propias series documentales (e.g., "Cumplimiento normativo", "Tratamiento térmico", "Recursos Humanos").
- **Subseries**: Mayor detalle funcional dentro de una serie (e.g., "Juicios laborales", "Certificaciones técnicas").
- **Expedientes**: Agrupaciones documentales homogéneas generadas según los criterios de cada subserie.
- **Unidades Documentales (UD)**: Documentos individuales con atributos específicos que son la base del análisis.

Cada nivel está identificado por un código (por ejemplo, SE-01, SS-03, etc.) y se relaciona jerárquicamente con el superior mediante claves foráneas. El objetivo es generar una **estructura relacional completa**, que simule fielmente un archivo institucional real.

3.2. Código de generación

A continuación se describen las principales secciones del código y su lógica:

Generación de niveles jerárquicos

Todo el proceso de generación se encuentra extensamente detallado en el JupiterNotebook asociado a esta memoria, por lo que no vamos a repetirnos.

El proceso comienza con la definición del fondo único y la generación de secciones. Para cada sección, se definen múltiples **series** asociadas, y para cada serie, varias **subseries**. Estas subseries son clave para determinar reglas específicas de conservación y tratamiento.

Generación de expedientes y unidades documentales

Para cada subserie se generan de forma aleatoria múltiples expedientes, y para cada expediente un número variable de unidades documentales:

num_expedientes = Se generan los que elijamos.

num_uds_por_expediente = np.random.randint(3, 8) (Apetura, al menos un doc y cierre)

Cada **unidad documental** se enriquece con atributos esenciales para el modelo:

- porcentaje_datos_personales: entre 0% y 100%, condicionado por la subserie.
- estado_conservacion: nivel de ilegibilidad (de 0 = perfecto a 100 = ilegible).
- numero_usos: veces que ha sido consultado (proxy de valor e importancia).
- subserie_codigo, expediente_codigo: referencia jerárquica.
- importancia: alta, media o baja, derivada del número de usos o función crítica

Asignación del tratamiento documental

Basado en las combinaciones de atributos anteriores, se decide el tratamiento_recomendado:

- **Tratamiento digital requerido**: para documentos ilegibles o de importancia alta dentro del periodo de retención.
- **Expurgar**: si tiene baja importancia,alto nivel de ilegibilidad o bajo uso y/o está fuera del periodo de retención.
- **Conservar**: si el documento está en buen estado y tiene un uso razonable.

4. Carga y revisión de datos

Los archivos CSV generados (fondo.csv, seccion.csv, serie.csv, etc.) se leen con pandas. Se realiza una verificación de:

- Estructura jerárquica consistente entre niveles.
- Formato y tipos de datos válidos.
- Coherencia en los valores asignados (porcentajes entre 0–100, correspondencia entre códigos padre-hijo, etc.).

5. Análisis exploratorio de datos (EDA)

Una vez generada la base sintética de unidades documentales y cargada desde el archivo .csv, se realiza un análisis exploratorio de datos (EDA) con el objetivo de:

- Comprender la distribución y características de las variables.
- Identificar posibles valores atípicos o inconsistencias.
- Visualizar relaciones entre variables que puedan ser relevantes para la predicción del tratamiento documental.

5.1. Variables principales:

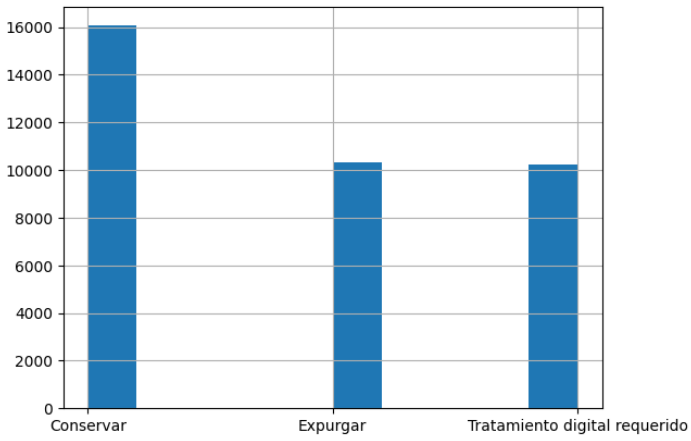
Entre las columnas más relevantes destacan:

- codigo_ud: identificador único de la unidad documental.
- expediente_codigo, subserie_codigo: permiten reconstruir la jerarquía documental.
- estado_conservacion: nivel de deterioro/ilegibilidad (0 = perfecto, 100 = ilegible).
- porcentaje_datos_personales: de 0% a 100%, importante para cumplimiento normativo.
- importancia: clasificada como **Baja**, **Media** o **Alta**.
- tratamiento_recomendado: variable objetivo (**Conservar**, **Expurgar**, **Tratamiento digital requerido**).

5.2. Distribución de clases:

Variable tratamiento (target):

Se evalúa el equilibrio de clases en la variable objetivo:

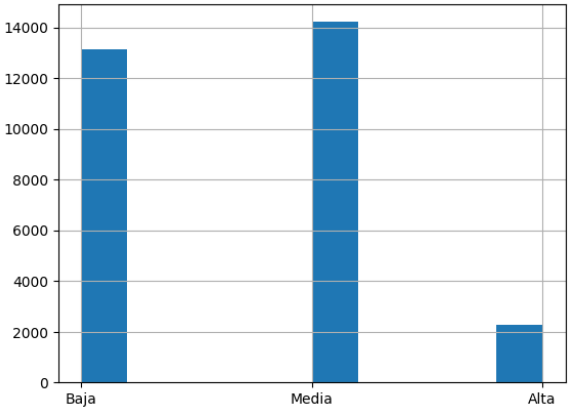


Observación:

- Las clases están moderadamente equilibradas, aunque se observa una mayor proporción de documentos clasificados como Conservar.
- Esto es consistente con la lógica aplicada: muchos documentos con son de apertura o cierre y tienen ese tratamiento por defecto, a pesar de ser reiterativos, a veces ser ilegibles y suponer una carga.

Variable importancia

La variable importancia es **categórica ordinal**, con valores:

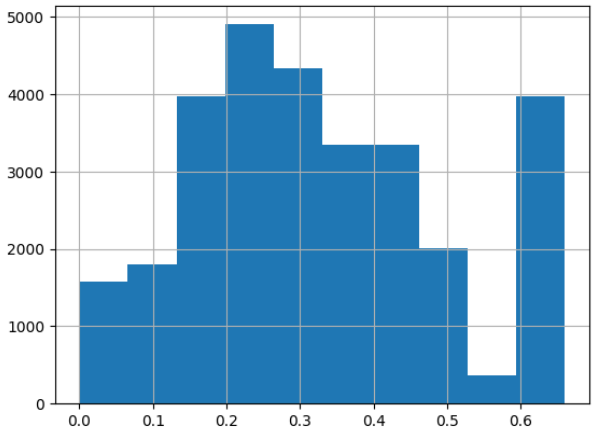


- **Alta**: documentos críticos (por función o uso).
- **Media**: valor operativo intermedio.
- **Baja**: escaso uso o relevancia funcional.

Observaciones clave:

- Más del 90% de los documentos de **importancia Alta** son tratados como Tratamiento digital requerido, en línea con la política definida para documentos de Junta Directiva, I+D+i, Litigios, etc.
- En los documentos de **importancia Media**, predomina la conservación o el tratamiento digital, dependiendo de su estado de conservación.
- En los de **importancia Baja**, predomina el Expurgar.

Variable datos personales:



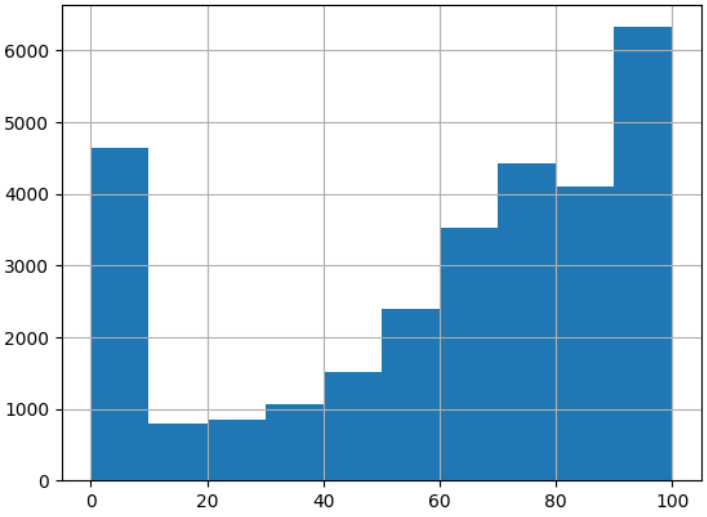
Observaciones:

- La subserie **Recursos Humanos** muestra sistemáticamente valores superiores al 60%, lo que confirma que se cumplió la regla de generación: mínimo del 60% de datos personales.
- Subseries como **Clientes**, **Asuntos jurídicos** o **Atención médica** también presentan niveles elevados.
- La presencia elevada de datos personales condiciona los tratamientos, ya que ciertos documentos no pueden conservarse ni digitalizarse sin procesos de anonimización.

Variable ilegibilidad:

Observaciones:

- Parece que hay una relación clara entre la antigüedad de un documento y su 'estado de conservación'.
- Muchos de los documentos de baja importancia tienen mal estado de conservación.
- Gran porcentaje de 0% porque gran cantidad de documentos no se ven afectados.

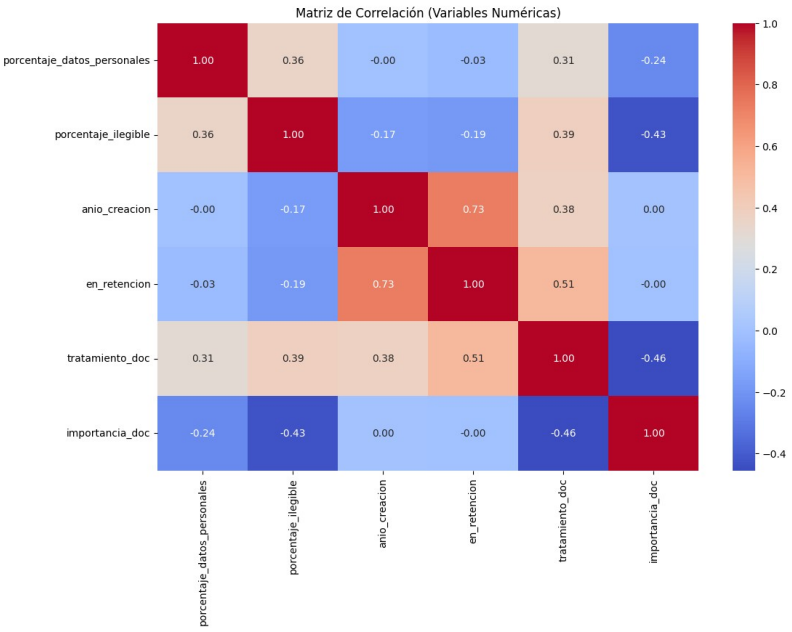


Otras variables creadas ad hoc:

Variables como 'doctip', que codifican el tipo documental, 'estado', que resumen los documentos entre abierto o cerrado, como un booleano, 'en_retencion', que dividen la base de datos entre los documentos que se encuentran o no en este periodo, y el año de creación del documento, solo el año.

Correlaciones y visualización conjunta:

Aunque las variables no son muchas ni altamente correlacionadas, se puede observar cierta consistencia y sinergia entre ellas, no siendo ninguna cercana a 0.



Conclusión del EDA

Este análisis exploratorio muestra que:

- La lógica de generación de los datos sintéticos se ha aplicado de forma consistente.
- Las variables muestran relaciones coherentes con la clase objetivo.
- importancia, estado_conservacion y porcentaje_datos_personales serán claves para entrenar modelos clasificadores.
- No se observan errores graves ni valores atípicos incompatibles, lo que permite pasar a la siguiente fase de preprocesamiento y modelado.

6. Limpieza de datos y selección de modelos

6.1. Limpieza de datos

En esta etapa se ha definido una función central, modelo_pruebas, que permite aplicar un flujo de limpieza general a cualquier conjunto de datos nuevo. Este enfoque modularizado y reutilizable facilita que el pipeline se aplique automáticamente sobre nuevas bases con problemas reales como valores nulos, códigos inconsistentes o errores humanos frecuentes.

La función realiza varias operaciones clave:

- **Gestión de valores nulos**: Se imputan automáticamente valores ausentes según el tipo de variable (por ejemplo, rellenando campos vacíos con modos, medias, o valores por defecto según la lógica del campo).
- **Normalización y transformación de campos**: Se aseguran formatos homogéneos para variables categóricas, fechas o descripciones que podrían variar entre documentos.
- **Cruce con otras tablas** (exp, ssub): La limpieza no se hace de forma aislada sobre la unidad documental. Se incorporan metadatos del expediente y de la subserie para enriquecer el dataset y contextualizar mejor cada unidad.

Tras la limpieza, se genera un DataFrame consolidado, sin nulos, listo para ser procesado.

6.2. Preparación para modelado

Una vez limpio el dataset, se realiza la división clásica entre variables independientes (X) y la variable objetivo (y), que en este caso es tratamiento_doc. Esta variable toma tres valores:

- Conservar
- Expurgar
- Tratamiento digital requerido

La base se divide en conjunto de entrenamiento y de prueba utilizando train_test_split, con el objetivo de validar los modelos de forma robusta.

6.3. Modelado con múltiples algoritmos

Se ha implementado un sistema de comparación entre modelos usando GridSearchCV, con 5-fold cross-validation y búsqueda en grilla de hiperparámetros. Los modelos comparados fueron:

- RandomForestClassifier
- XGBoostClassifier
- CatBoostClassifier

Cada uno con su propia grid de parámetros adaptada, por ejemplo:

- RandomForest: número de árboles (n_estimators) y profundidad (max_depth)
- XGBoost: tasa de aprendizaje (learning_rate) y regularización (max_depth)
- CatBoost: optimizado para categorías sin necesidad de encoding manual

La métrica principal utilizada fue **accuracy**, que mide la capacidad del modelo para discriminar correctamente entre las tres clases del tratamiento documental. Su porcentaje de acierto. Al ser una base de datos perfecta, sin errores típicos, y con tres categorías de target muy marcadas, el accuracy para train y test no bajó de 0.9996, en el peor caso.

6.4. Modelos aplicados sobre datos imperfectos

Tras entrenar los modelos sobre la base "perfecta", se aplicaron también sobre un segundo conjunto de datos, más representativo de un entorno real: documentos con errores humanos, tratamientos mal asignados y valores nulos.

Este segundo conjunto se limpia igualmente con modelo_pruebas y luego se vuelve a dividir para aplicar los modelos entrenados. Se ha comprobado que el rendimiento baja ligeramente en este contexto, lo cual es esperable, pero sigue siendo útil para detectar inconsistencias. Tiene un accuracy de cerca del 0,90.

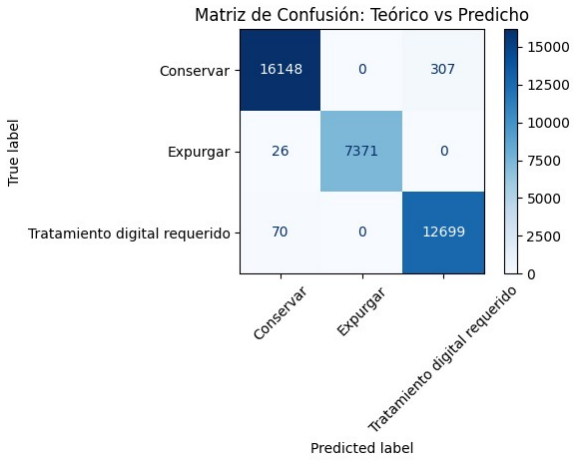
6.5. Modelado de detección de errores

Finalmente, se incorpora un modelo no supervisado con IsolationForest, entrenado para detectar outliers dentro de los datos ya tratados. Se ha observado que aquellos registros marcados como anómalos suelen coincidir con errores en el tratamiento documental (como unidades que deberían haber sido expurgadas, pero figuran como conservadas). Este procedimiento se detalla mejor en el JupiterNotebook asociado a esta memoria, en el que se explica por qué se a elegido IsolationForest.

Además, se implementa un sistema que sugiere un **tratamiento alternativo**, cuando se detecta que la predicción del modelo no coincide con el tratamiento registrado originalmente. Cambiándolo automáticamente.

6.6. Evaluación del rendimiento

Se ha aplicado una matriz de confusión entre los valores finalmente obtenidos y modificados con unos valores que se tenían guardados desde la generación documental de los expedientes y unidades documentales con fallos, esos mismos valores, pero sin los fallos, dando lugar a la siguiente gráfica:



En ella podemos observar que el 98,9% de los resultados son correctos, por lo que se ha reducido casi en diez puntos la predicción inicial.

Además, en lo que más se confunde el modelo es al clasificar la documentación que tiene los tratamientos 'Tratamiento digital recomendado' y 'Conservar', porque ambos son posibles en retención. No es un error crítico y su revisión se puede plantear dentro del flujo de trabajo normal de la empresa.

Lo importante es que se ha reducido al máximo el número de unidades críticas a revisar, unas 26 de 38000, lo que es viable para una revisión manual y va a suponer un ahorro importante de dinero y tiempo.

7. **Conclusiones, futuros proyectos y aplicaciones.**

La descripción multinivel, sobre todo desde el nivel subserie hasta unidades documentales, ha ido nutriendo la base de datos de unidades documentales, sobre la que finalmente se realizaba el tratamiento. Variables como estado, año de creación o tipo documental hacen que las predicciones sean más limpias.

Si bien partíamos de la base de que nos proporcionaban una base de datos trabajada y sin errores para llevar a cabo nuestro primer entrenamiento, el trabajo de modelado de datos y relleno de nulos para poder auditar los trabajos que han llevado a cabo las otras subsidiarias de la Empresa me parece relevante, pues ahorra una gran cantidad de trabajo a la hora de revisar documentación, pues solo un 1,1% tiene un tratamiento dudoso, siendo crítica su revisión en solo 26 unidades documentales (siendo seguramente alguno de apertura o cierre, teniendo claro su tratamiento desde un principio de la revisión).

De la forma en la que hemos estructurado los modelos, la empresa podría introducir todas las bases de datos de sus subsidiarias y estas serían escrutadas, auditadas y corregidas hasta el 98%, a pesar de que los errores sean notorios. Además, en el caso de que sepamos que una empresa sigue otros criterios para su clasificación documental (considera un mayor/menor número de usos como corte entre las importancia, tiene un porcentaje mayor/menor para considerar que un documento es ilegible o confidencial, o decide de una manera más tajante sobre el expurgo, porque no cuenta con los medios para un tratamiento digital) solo tendríamos que re-entrenar el RandomForest.

Mientras realizaba este trabajo no dejaba de pensar en las aplicaciones complementarias que se le podrían añadir.

- Un lector OCR que clasifique directamente, identificando los patrones como tipo documental, nombre, porcentaje de ilegibilidad o datos personales, o el año, ya que muchos de estos documentos son digitalizados. Dejando solo la labor de clasificación, que me parece complicada sin una ruta pre establecida por el Cuadro de Clasificación.
- Si se implementa en un Sistema de Gestión Documental (SGD) se puede hacer que llegado a cierto nivel de confianza sobre la predicción tome la iniciativa y realice la acción de conservar, copiándolo en una carpeta diferente, o expurgar, eliminándolo del sistema, dejando un testigo, que en el caso de esta Empresa son los documentos de apertura y cierre.
- Con respecto a ese tipo documental, los documentos de apertura/cierre ocupan un espacio masivo en el Archivo, tanto digital como físico. Si bien son documentos que sirven para dejar constancia de la actividad de una Empresa. Una de las tareas que quedaría pendiente es que si todos los documentos que se contienen en ese expediente, que no sean ellos mismos, son expurgados, ellos también lo sean, dejando un testigo en una nueva clase documental, que podría ser incluso un excel o un csv del expurgo de ese documento, anotando la fecha y el responsable del mismo.
- Las aplicaciones en el ámbito archivístico me resultan infinitas, fuera ya del propio ámbito empresarial y siendo generalista, la potencia que tienen estos modelos para entender y clasificar contenidos los hacen muy relevantes para encontrar patrones en la documentación y poder realizar descripciones más ricas y detalladas. Hay una sección en la ISAD(G) llamada 'Alcance y contenido' que se podría automatizar con esta identificación y que resulta del todo tediosa su elaboración.