Máster

Máster en Big Data: Tecnología y Analítica Avanzada

Procesamiento y Transformación de Datos para la Predicción de Riesgos de Crédito Empresarial.

Autor
Lawrence Javier Minguillan Van Kapel

Supervisor
Eduardo Lobo Fenouil

Madrid
Mayo 2025

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Procesamiento y Transformación de Datos para la Predicción de Riesgos de Crédito Empresarial.

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2024/25 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha

sido tomada de otros documentos está debidamente referenciada.

Fdo.: **Lawrence Javier Minguillán Van Kaple**

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: **Eduardo Lobo Fenouil**

Resumen

Este proyecto se centra en el desarrollo de una herramienta de extracción de información financiera que servirá de preámbulo para la creación de modelos predicativos avanzados para evaluar el riesgo crediticio de empresas corporativas. La metodología empleada combina técnicas de extracción de información de datos basada en "fuzzy logicçombinada con una interfaz final para la extracción y evaluación de probabilidad de devolución de los créditos.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Resumen ejecutivo del proyecto

1.1. Contexto y objetivo

El contexto y objetivo de desarrollar una herramienta de extracción de datos financieros de los çredit packs.[en] una institución como Ebury están relacionados con el proceso de evaluación del crédito y la toma de decisiones sobre la financiación de las operaciones comerciales de los clientes. Los çredit packs"son conjuntos de información financiera y documentación que se utilizan para evaluar la viabilidad financiera de una empresa cliente y para determinar si cumple los requisitos necesarios para obtener un crédito para una operación comercial especificada.

Básicamente, cuando Ebury considera conceder líneas de crédito sin seguro o con seguro reducido, la decisión recae en el comité de crédito. Una vez que este comité aprueba una línea de crédito y el cliente está de acuerdo con la tarificación, se desencadena un proceso específico. Ese proceso implica generar un enlace Avoka (que incluye términos y condiciones, tamaño de la línea de crédito, precio y convenios) y cualquier documento adicional requerido por el comité de crédito. Estos documentos pueden incluir información sobre la naturaleza del negocio del cliente, el sector industrial, la estructura de propiedad, así como la due diligence mejorada y una evaluación de riesgo significativa.

En cuanto a la funcionalidad de la herramienta de extracción de datos financieros, su objetivo principal sería automatizar y eficientizar la recopilación, el análisis y la presentación de información relevante contenida dentro de los çredit packs". Esto sería esencial para tomar decisiones informadas sobre la concesión de créditos para financiar operaciones comerciales. Esta herramienta permitiría a Ebury evaluar rápidamente la solvencia de un cliente y la probabilidad de recuperar el crédito otorgado, basándose en el cumplimiento de los pagos y el ciclo de conver-

sión de efectivo de las operaciones comerciales financiadas.

Por ejemplo, para financiar la compra de bienes y servicios, se necesitan facturas elegibles que muestren un valor monetario futuro (para poder devolver el financiamiento). Los bienes financiados deben ser vendidos o utilizados en la fabricación de otros productos que se venderán dentro de los 150 días siguientes a la financiación, y los ingresos de estas ventas se utilizarían para repagar el préstamo.

1.2. Estado del arte

El estado del arte del proyecto presentado aborda diversas áreas clave que son fundamentales para la implementación de un sistema de evaluación predictiva de riesgos crediticios para empresas. Este análisis se centra en la integración y uso de tecnologías modernas y avanzadas. Está dividido en varias secciones:

Extracción, Transformación y Carga (ETL) La metodología ETL (Extraction, Transformation, and Load) es crucial para el procesamiento de grandes volúmenes de datos financieros. La etapa de extracción implica recolectar datos de diversas fuentes, incluyendo sistemas internos y externos. La transformación consiste en limpiar, normalizar y preparar los datos para su análisis. Finalmente, la carga implica insertar los datos procesados en un sistema de almacenamiento o base de datos. En el proyecto, se hace hincapié en la automatización de este proceso para mejorar la eficiencia y precisión. La alternativa de usar puramente python con normas de extracción seria utilizar un modelo de OCR para analizar directamente los caracteres.

Desarrollo de la Aplicación Web La implementación de una aplicación web interactiva utilizando Streamlit es uno de los puntos destacados del proyecto. Streamlit permite crear aplicaciones web rápidas y eficientes para la visualización de datos y la interacción del usuario. Esta herramienta facilita a los analistas la manipulación y análisis de datos sin necesidad de conocimientos avanzados de programación. La alternativa a streamlit podría haber sido usar HEX, pero implicaría tener el código almacenado en su interfaz y no en github, o flask, pero hubiera implicado un mayor desarrollo de front-end.

Uso de Tecnologías en la Nube El proyecto aprovecha los servicios de Google Cloud Platform (GCP) para el almacenamiento y procesamiento de datos. GCP proporciona una infraestructura escalable y segura que permite manejar grandes volúmenes de datos y realizar análisis complejos. La integración con herramientas

como Google BigQuery y Looker facilita la creación de dashboards y reportes personalizados que son esenciales para la toma de decisiones basadas en datos. Se podria haber usado cualquier proveedor cloud, pero actualmente se usa GCP.

Evaluación Predictiva y Modelado La evaluación predictiva de riesgos crediticios se basa en modelos avanzados de machine learning, como XGBoost, que permiten predecir con alta precisión el comportamiento crediticio de las empresas. Estos modelos analizan datos históricos y patrones para identificar riesgos potenciales y ayudar a tomar decisiones informadas. La implementación de estos modelos en el sistema permite una evaluación continua y automatizada del riesgo. La alternativa a este modelo podría haber sido una regresión logística o un árbol de decisión que fuese mucho más entendible por el equipo de riesgos.

Automatización y Eficiencia Operativa Uno de los principales beneficios del proyecto es la automatización de tareas repetitivas y manuales, lo que libera tiempo para que los analistas se concentren en actividades de mayor valor añadido. La creación de un proceso automatizado semanal para la recolección y análisis de datos mejora significativamente la eficiencia operativa y reduce el riesgo de errores humanos. La alternativa a este proceso sería una ejecución manual semanalmente.

1.3. Descripción de la empresa

Ebury es una institución financiera que proporciona servicios de gestión de riesgo de divisas, pagos internacionales y soluciones de préstamos flexibles a empresas. Utiliza una combinación de experiencia financiera y tecnología avanzada para ayudar a las empresas a manejar su riesgo cambiario (FX), realizar pagos internacionales de manera eficiente y financiar sus compras para crecer más rápido.

Con cobertura global, Ebury permite a sus clientes operar en más de 130 monedas, incluyendo divisas de mercados emergentes que otros proveedores pueden tener dificultades para ofrecer. Además, proporciona líneas de crédito no garantizadas de hasta £3 millones durante un máximo de 150 días.

La compañía apoya a sus clientes, independientemente de su tamaño, para competir a nivel mundial, ofreciendo soluciones técnicas innovadoras que brindan flexibilidad financiera y funcionalidad previamente limitada a grandes multinacionales. Ebury se enorgullece de contar con el apoyo y la confianza de inversores líderes mundiales como Banco Santander, 83North, Angel CoFund y Vitruvian.

Además, Ebury está autorizada y regulada por la Financial Conduct Authority (FCA) del Reino Unido como una Institución de Dinero Electrónico y como Firma de Inversión para proporcionar asesoramiento y ejecutar operaciones en productos derivados de MiFID, como los contratos de divisas a término (FX Forwards).

La empresa ofrece apoyo dedicado y soluciones personalizadas a una variedad de clientes, incluyendo empresas e-commerce, ONGs y organizaciones benéficas, donde servicios como la captación de fondos en múltiples monedas o la gestión del riesgo de cambio de divisas son esenciales.

1.4. Descripción del proyecto

El proyecto en cuestión tendría como objetivo principal mejorar la eficiencia y precisión en el proceso de evaluación del crédito. Esto se lograría mediante la automatización de la recopilación de datos de los çredit packs"que son analizados por los analistas de crédito de la entidad financiera cada semana. Los çredit packsçontienen información financiera vital que permite evaluar la solvencia y el riesgo de crédito de las empresas.

La extracción automatizada semana a semana permitiría tener una visión más actualizada y rápida del estado de crédito de los clientes que están siendo evaluados sin tener que depender de procesos manuales, lo que puede ser propenso a errores y consumir mucho tiempo.

Además, la creación de una aplicación utilizando Streamlit facilitaría la ejecución de tareas .ªd hoc"por parte del equipo de crédito. Streamlit es un marco de trabajo que permite a los desarrolladores crear aplicaciones de datos en Python de manera ágil y visual. La aplicación permitiría a los usuarios seleccionar específicamente çredit packsz extraer la información relevante de ellos cuando sea necesario. Esto significa que se podría tener flexibilidad además de la extracción semanal automatizada para realizar análisis puntuales según la demanda.

Con la información extraída, la aplicación podría calcular automáticamente el rating o la calificación crediticia de las empresas, lo cual es esencial para presentar la propuesta de crédito frente a un comité de crédito. Tener el rating facilitaría el proceso de toma de decisiones y apoyaría argumentos con datos precisos, aumentando así las posibilidades de aprobar operaciones de crédito justas y bien fundamentadas.

La eficiencia mejorada no solo beneficiaría a los analistas de crédito al redu-

cir su carga de trabajo manual, sino que también permitiría presentar los çredit packs.ªnte el comité de crédito de una manera más estructurada y con mayor confianza en la precisión de los datos.

1.5. Propuesta de valor y aspectos financieros

Para analizar la propuesta de valor y los aspectos financieros que puede conllevar el proyecto, vamos a establecer algunos supuestos:

- El salario promedio de un analista de crédito: Supongamos que es de 30000€ anuales.
- Horas de trabajo al año: Un año de trabajo estándar sin contar vacaciones, ni fines de semana es aproximadamente de 1520 horas (38 semanas x 40 horas/semana).
- Cantidad de veces que buscan la empresa por semana: Supongamos que buscan la misma empresa 10 veces a la semana.
- Numero de empresas en las que los analistas trabajan semanalmente: 3 empresas.
- Numero de analistas en la empresa: 19 analistas.

1.5.1. Análisis de la situación actual:

- Tiempo dedicado a la búsqueda manual por empresa: 3 minutos.
- Tiempo dedicado a la búsqueda manual por semana: 3 empresas/semana * 3 minutos/empresa * 10 veces/semana = 90 minutos/semana.
- Tiempo anual invertido: 90 minutos/semana * 38 semanas/año = 3420 minutos/año.
- Horas anuales invertidas: 3420 minutos/año / 60 minutos/hora = 57 horas/año.

Calculamos el coste de este tiempo basado en el salario de los analistas:

- Coste por hora de trabajo de un analista: 30000€ / 1520 horas/año = aproximadamente 19,73€/hora.
- Coste anual de la búsqueda manual: 57 horas/año * 19.73€/hora = 1124€

1.5.2. Propuesta de valor con la automatización y creación del dashboard:

Supongamos que la automatización y el dashboard pueden reducir el tiempo de búsqueda por empresa en 2 minutos (quedando en 1 minuto por empresa).

- Tiempo dedicado a la búsqueda con el dashboard por empresa: 1 minuto.
- Tiempo dedicado a la búsqueda con el dashboard por semana: 3 empresas/-semana * 1 minuto/empresa * 10 veces/semana = 30 minutos/semana.
- Tiempo anual invertido con el dashboard: 30 minutos/semana * 38 semanas/año = 1140 minutos/año.
- Horas anuales invertidas con el dashboard: 1140 minutos/año / 60 minutos/hora = 19 horas/año.

Calculamos el nuevo coste con la reducción de tiempo:

Coste anual de la búsqueda con el dashboard: 19 horas/año * 19,73€/hora =374.87€ .

1.5.3. Ahorro con la implementación del proyecto:

- Ahorro anual en tiempo: 57 horas/año - 19 horas/año = 38 horas/año.
- Ahorro anual en costes: 1124 - 374 = 750€/analista.
- Ahorro anual en tiempo total: 38 horas/año * 19 analistas = 30,08 dias.
- Ahorro anual en costes totales: 750€/analista * 19 analistas = 14250€

1.5.4. Tiempo ahorrado con la presentación a mitad de semana:

Si normalmente tienen que esperar una semana para presentar un çredit pack", y con el nuevo sistema pueden presentarlo a mitad de semana, supongamos que esto significa que pueden presentar el pack 3 días antes.

- Si trabajan en 3 credit packs a la semana, eso sería un ahorro de 3 días/semana para la totalidad de los credit packs.
- El ahorro anual sería 3 días/semana * 38 semanas/año = 114 días/año.
- Esto no solo ahorra tiempo sino que también puede acelerar el ciclo de decisión y potencialmente permitir una rotación más rápida del capital, mejora de la utilización de recursos y posiblemente un mayor volumen de negocio.

1.5.5. Costes del Proyecto:

Para determinar la viabilidad desde el punto de vista financiero, se deben considerar los costes de desarrollo, implementación y mantenimiento del sistema de automatización y el dashboard. El coste de desarrollo es unicamente el de un becario trabando 20 horas a la semana con un salario aproximado de 600€ al mes, durante 4 meses. Es decir un coste total de 2400€. Lo que sigue dejando un beneficio anual de 11850€.

1.6. Conclusión

El proyecto propuesto presenta una valiosa oportunidad para Ebury al ofrecer una automatización avanzada y eficiente de las tareas relacionadas con la extracción y el análisis de datos financieros. La implementación de esta herramienta se traduce en una reducción significativa de los tiempos dedicados a las búsquedas manuales, lo que no solo disminuye los costos asociados con el trabajo analítico sino que también promete un notable aumento en la precisión y la velocidad de la toma de decisiones crediticias.

Con una inversión inicial moderada, el ahorro en el tiempo de trabajo de los analistas es sustancial, lo que evidencia la sólida viabilidad financiera del proyecto. La utilización de herramientas tecnológicas punteras es un factor diferencial clave que permitirá a Ebury mejorar sus servicios y capacidades de financiamiento, así como aumentar potencialmente el volumen de negocio gracias a una rotación más rápida del capital.

En conclusión, la introducción de esta innovación en el proceso de evaluación del crédito se perfila como un paso estratégico para optimizar las operaciones internas de Ebury, teniendo impactos positivos a corto y largo plazo en la eficiencia operativa, la satisfacción laboral y el rendimiento financiero.

Capítulo 2

Desarrollo del proyecto

2.1. Situación inicial

2.1.1. Operativa empresarial

El proceso de aprobación de préstamos por parte de los analistas de crédito de Ebury, que incluye su paso por el comité de crédito, es meticuloso y sigue una serie de pasos estratégicos. En cuanto a las limitaciones, se refieren a varios factores que pueden afectar el flujo del proceso y la toma de decisiones, tales como la calidad de la documentación presentada y el tiempo dedicado al análisis.

Inicialmente, se realiza una calificación preliminar del prospecto para garantizar que cumpla con los requisitos mínimos establecidos por Ebury. Este es un filtro esencial para identificar potenciales clientes que sean adecuados para el análisis de riesgo.

Una vez que el prospecto se califica como potencial cliente, se procede a la recopilación de documentos. El equipo de ventas tiene la responsabilidad de asegurarse de que toda la documentación relevante para la evaluación de riesgo sea recopilada y enviada a los analistas. La integridad y la calidad de esta documentación son vitales, ya que cualquier fallo o carencia en ella puede resultar en demoras en el proceso o incluso en la negación del préstamo.

Los analistas de riesgo de Ebury realizan un análisis detallado de las finanzas del cliente. Este paso involucra un examen exhaustivo de balances financieros, estados de cuentas de gestión, análisis de deudores y acreedores y otro tipo de declaraciones relevantes como las de IVA. Además, se incluye una evaluación del

historial del negocio que puede comprender chequeos en registros públicos y una discusión con los directores financieros de la empresa solicitante. La complejidad de esta etapa dependerá en gran medida de la calidad de los datos proporcionados, ya que si los documentos son ambiguos o incompletos, el análisis puede resultar más prolongado y complicado.

Seguido al análisis de riesgo, se tiene el proceso de aprobación del crédito, el cual se realiza a diferentes niveles dentro de la organización en función del tamaño de la línea de crédito solicitada. Este proceso incluye la aprobación del comité de crédito que se basa en la evaluación de riesgo presentada por los analistas.

Los seguros de crédito suelen ser un requisito común para aprobar la mayoría de las líneas de crédito. Sin embargo, la decisión de otorgar líneas sin seguro o con seguro reducido recae en el comité de crédito, y no está garantizada para todos los clientes. Esta es una limitación importante, ya que la ausencia de un seguro puede poner en riesgo la viabilidad del préstamo.

Una vez aprobado el crédito, existe una ventana de operación durante la cual los clientes pueden realizar transacciones sin necesidad de entregar nueva documentación. El tiempo de esta ventana es una restricción que limita el período durante el cual el cliente puede operar libremente antes de tener que presentar información financiera actualizada.

En cuanto al proceso de incorporación de los clientes al servicio, una vez que el comité de crédito ha dado su aprobación y se han acordado los términos y condiciones, se llevan a cabo una serie de comprobaciones para completar el proceso de onboarding. Este procedimiento es fundamental para garantizar el cumplimiento de los clientes y la prevención del fraude, aunque puede prolongar el tiempo hasta que el cliente esté finalmente listo para realizar sus operaciones financieras.

A través del portal en línea de Ebury (EBO), los clientes pueden enviar detalles de sus operaciones financieras y deben proporcionar información precisa sobre sus intercambios comerciales. La velocidad con la que los clientes pueden presentar esta información y la precisión de la misma puede impactar en la rapidez y eficiencia con la que se pueden procesar sus solicitudes de financiamiento.

Las principales limitaciones en este proceso son la calidad y precisión de la documentación entregada por los clientes y el tiempo necesario para realizar todas las verificaciones y análisis. La documentación deficiente puede resultar en demoras significativas o incluso en el rechazo de una solicitud de crédito. Además, otro

factor limitante es la presión temporal, ya que un análisis riguroso es indispensable para una buena gestión del riesgo, pero también se debe tener en cuenta la necesidad de los clientes de obtener un financiamiento oportuno. Estas restricciones exigen que tanto Ebury como sus clientes trabajen conjuntamente y de manera eficiente para garantizar un proceso fluido y exitoso.

La implementación de una herramienta automatizada para la extracción de datos financieros puede transformar el proceso de evaluación y aprobación de créditos de Ebury al ofrecer una solución a las limitaciones asociadas con la precisión de la documentación y los tiempos de análisis. Esta herramienta asegura la calidad y consistencia de los datos recogidos, lo que resulta en una reducción significativa de errores humanos. Al agilizar la recopilación de datos, los analistas de crédito pueden concentrarse en tareas de mayor valor, como el análisis interpretativo y la toma de decisiones estratégicas. Además, la capacidad de realizar extracciones ad hoc posibilita una respuesta rápida en situaciones que requieren una evaluación inmediata de la empresa, lo cual es crucial para satisfacer las necesidades urgentes de financiamiento de los clientes.

Con datos financieros siempre actualizados gracias a la extracción automática semanal, los analistas pueden operar con la información más reciente, permitiendo un mejor manejo del riesgo y toma de decisiones más informadas. La estandarización de los procedimientos de evaluación mejora la equidad y transparencia en el otorgamiento de crédito, al tiempo que reduce la carga administrativa tanto para Ebury como para sus clientes. En resumen, la automatización no solo mejora la eficiencia operativa sino que también fortalece la posición competitiva de Ebury en el mercado financiero.

2.1.2. Operativa tecnológica

El sistema original de extracción de datos estas enfocado únicamente a la extracción de ciertos datos de la hoja de balance, a partir de ahora referenciado como BS, por sus siglas en ingles Balance Sheet. Estos datos se pueden ver reflejados en la siguiente tabla: 2.1.

La búsqueda de estos datos esta fijada a ciertos campos en posiciones concretas lo que impide que el sistema funcione en caso de haber algún campo mal posicionado. La búsqueda se puede dividir en dos partes:

NOMBRE	DESCRIPCIÓN
bvd id number	Identificador único de Bureau van Dijk
Name	Nombre de la empresa
Currency	Moneda usada
Closing date	Ultima fecha con informacion financiera registrada
Execution date	Fecha de extraccion de los datos
Total current assets	Total de activos actuales
Total inventory	Inventario total
Trade debtors	Deudores
Total cash equivalent recibables	Total de equivalentes de efectivo recibidos
Total assets	Total de activos
Share holders funds	Fondos de los accionistas
Total current liabilities	Total de pasivos corrientes
Loans	Préstamos
Trade creditoes	Creditores comerciales
Turnover	Volumen de negocio
Profit and loss bedore tax	Beneficios y pérdidas antes de impuestos
Net profit loss	Beneficio neto

Cuadro 2.1: Campos originalmente extridos de los Credit Packs

1. Campo deseado

El formato original de búsqueda con el que trabajamos se basa en un enfoque riguroso que requiere precisión en la localización de datos. Se demanda que el programa use con exactitud la fila (índice) y la columna donde se espera encontrar la información solicitada.

Sin embargo, a simple vista, es evidente que este sistema de búsqueda tiene sus limitaciones y puede ser propenso a fallos. En situaciones donde los analistas cometen errores manuales durante la entrada de datos, o si deciden realizar cambios en la plantilla original para adaptarla a sus necesidades específicas, el sistema podría no reconocer estos ajustes. Esto se debe a que el sistema de búsqueda se diseñó para trabajar con un formato invariable y no está programado para adaptarse a modificaciones de posiciones o alteraciones de texto. En consecuencia, cualquier cambio no anticipado en la disposición de los datos o en el formato de la plantilla tiene el potencial de generar errores y afectar la eficacia de la búsqueda.

Para ilustrar mejor este concepto, haremos una representación de la búsqueda

de los datos financieros del campo "Financial assets". En esta ilustración, que se mostrará gráficamente, se podrá apreciar cómo el formato de búsqueda estático está diseñado para interactuar con la información dispuesta en la plantilla. La figura representará visualmente el campo "Financial assets"de manera que sea posible localizarlo a partir de la intersección precisa de la fila y columna correspondientes, destacando la importancia de mantener un orden estricto y sin modificaciones en la estructura de los datos para que el proceso de búsqueda funcione correctamente. Esto lo podemos apreciar en la figura 2.1

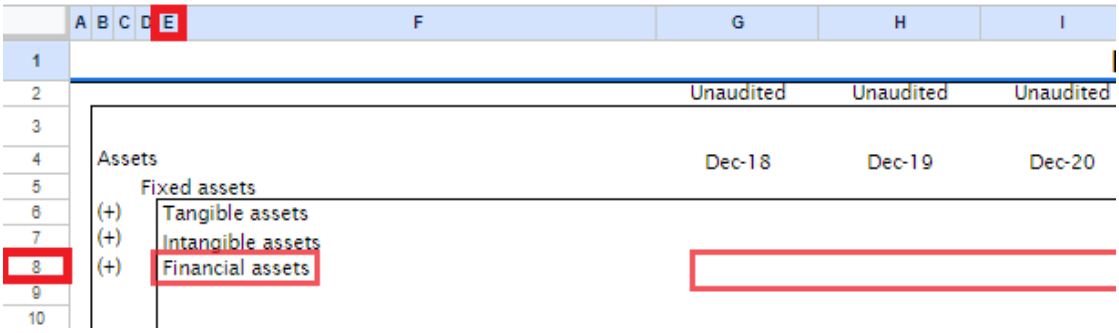

Figura 2.1: Busqueda de un campo especifico en el CP

2. Ultima fecha de actualización.

Después de haber conseguido identificar el campo, intentamos identificar las fechas mas actualizadas para las cuales ese campo especifico buscado tenga valores asociados.

Una vez que determinamos las fechas clave, procedimos a localizar la intersección exacta entre la fila correspondiente a "Financial assetsz las columnas que representan las fechas identificadas. Es en este punto de encuentro donde emergen los datos de interés que estamos buscando extraer. Los valores encontrados en esta intersección son los que, de manera meticulosa, procederemos a recopilar y almacenar en nuestra base de datos. Estos datos serán esenciales para la posterior utilización en una amplia gama de aplicaciones analíticas, lo que incluye la generación de herramientas especializadas y la construcción de dashboards intuitivos y detallados.

Es a través de estos dashboards y herramientas que podremos visualizar tendencias, crear proyecciones y realizar un sinfín de análisis cuantitativos que contribuyen directamente al proceso de toma de decisiones dentro de la organización.

La visualización de la intersección, subrayada en amarillo y la distribución de los datos relevantes queda claramente reflejada en la figura 2.2

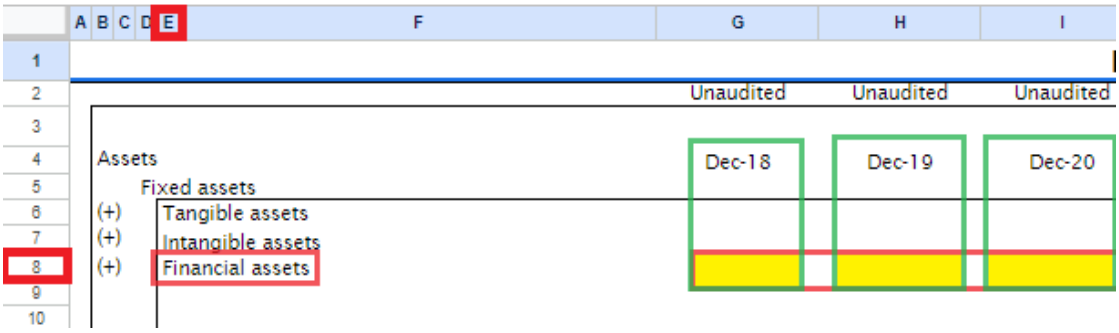

Figura 2.2: Búsqueda de las fechas mas actualizadas para el campo encontrado.

2.2. Desarrollo tecnológico

El desarrollo tecnológico se divide en tres etapas (ETL)

- Extraction: se centra en la extracción de los datos de los google sheets.
- Transform: se centra en la transformacion de los datos para poder almacenarlos en GCP.
- Load: se centra en el proceso de carga de los datos a GCP.

2.2.1. Extraction

El problema radica en el proceso de búsqueda y localización del campo deseado en la plantilla financiera. Mientras que el método actual para identificar la fecha es efectivamente un procedimiento aceptable, el verdadero desafío se encuentra en la flexibilidad requerida para encontrar los campos específicos que pueden estar distribuidos diversamente y con nomenclaturas variantes dentro de una plantilla del Balance Sheet (BS) más extensa de lo usual.

El desarrollo tecnológico para superar esta barrera consta de 6 pasos claramente definidos:

1. Implementar una lógica fuzzy que facilitará la identificación de todos los campos dentro de la plantilla expandida del BS. Esta técnica permite una búsqueda más tolerante a variaciones nominales y posicionales, pudiendo así extraer datos de campos que tienen nombres ligeramente distintos o que se encuentran en ubicaciones diferentes a las esperadas.

2. Desarrollar una plantilla específica para el Profit and Loss Statement (P&L) e implementar una búsqueda fuzzy. Esta aproximación incluirá el método de intersección por fechas vigente, adaptándolo a la lógica fuzzy para mejorar la precisión en la identificación de campos relevantes.

3. Crear una plantilla para el mom (month-over-month analysis) y llevar a cabo una búsqueda fuzzy de campos, esta vez introduciendo un novedoso método de intersección por fechas. Este método estará diseñado para rastrear y recuperar datos que muestren la evolución mensual del P&L a lo largo de un período de tres años.

4. Establecer un nuevo proceso de segmentación de la información, con el objetivo de delimitar un çuadrado.ªlrededor de los datos de interés encontrados en los Credit and Debt (CnD) schedules. Luego, llevar a cabo el mapeo del timeline de la deuda a sus columnas correspondientes y almacenar la información de forma sistemática.

5. Ajustar el proceso de extracción recién creado de tal manera que sea posible generar una nueva tabla. Esta tabla mantendrá el formato adecuado para ejecutar el análisis de rating de las compañías, preservando la integridad de los datos y los formatos necesarios para una evaluación precisa.

6. Desplegar el proceso y establecer una rutina semanal que automáticamente ejecute la extracción de la información de aquellas empresas sobre las cuales los analistas han realizado su trabajo durante la semana anterior. Esto asegura que se mantengan al día las bases de datos para análisis futuros y proporciona un flujo de información constante y actualizado a todos los interesados.

Estos pasos reflejan un enfoque estructurado y metódico para la mejora de la búsqueda y el registro de datos financieros, orientado a maximizar la eficiencia, la precisión y la accesibilidad de la información para los analistas.

A continuación entraremos en mas detalle en cada uno de los procesos.

Implementar una lógica fuzzy y ampliación de campos del BS

Para mejorar la capacidad de nuestro sistema de gestionar y analizar extensivamente los datos de un Balance Sheet (BS), proponemos implementar una lógica fuzzy para la identificación y categorización de los campos de datos. A continuación, detallamos los pasos para desarrollar esta metodología:

1. **Creación de una lista de campos a identificar**: Iniciamos por compilar un inventario exhaustivo de todos los posibles campos que podríamos encontrar en diversos formatos de Balance Sheet. Esto incluye no solo los campos

comunes, sino también aquellos menos frecuentes que podrían aparecer en situaciones específicas o en BS de diferentes países y sectores. Este diccionario de campos esta compuesto por un total de 53 campos.

2. **Desarrollo de una función de búsqueda fuzzy:** Programamos una función especializada que emplea técnicas de lógica fuzzy para identificar el campo más similar al que deseamos encontrar en un conjunto de datos. Esta función debe ser capaz de tolerar pequeñas desviaciones o diferencias en la denominación de los campos, reconocer abreviaturas, y adaptarse a variaciones lingüísticas, devolviendo la ubicación precisa del campo deseado (columna y fila) dentro del BS. Para hacer esto posible nos basamos en una lógica especifica apoyándonos en la librería difflib para obtener la información deseada.

3. **Almacenamiento de las columnas relevantes:** Una vez identificados los campos, procedemos a guardar la información de la columna correspondiente a cada campo en una nueva lista que posteriormente nos permitirá buscar en estas columnas con información posibles campos que no teníamos mapeados en la plantilla original. Esta acción es crucial para encontrar potenciales fallos a la hora de extraer campos de los CP y encontrar nuevos elementos.

4. **Extracción de datos de las filas mapeadas:** Utilizamos las columnas identificadas para extraer de manera sistemática la información contenida en las filas asociadas. Esto nos permite recopilar los datos específicos relacionados con cada campo de interés. Almacenamos esta información en un DataFrame en formato de fila. Este paso es crucial para poder mantener un formato estándar entre los diferentes CP que se estén analizando paralelamente.

5. **Búsqueda ampliada de nuevos campos:** Finalmente, con las columnas ya mapeadas, implementamos otro conjunto de búsquedas para detectar nuevos campos que podrían haberse añadido al BS y que no estén contemplados en la plantilla original. Este enfoque nos ayuda a mantener nuestro sistema actualizado con las últimas modificaciones y prácticas contables que pueden surgir con el tiempo. El formato de almacenamiento de esta información es similar al anterior, pero la información se transpone para poder tener un formato estandarizado de nombre de columnas.

Nuevo proceso de PnL

Para obtener la información del Profit and Loss Statement (PnL), que es un proceso muy similar al empleado para el Balance Sheet (BS) pero con campos distintos en otro Excel, seguimos estos pasos:

1. Iniciamos creando una lista de los campos deseados para la extracción de datos del PnL. Esta lista incluye elementos tradicionales como ingresos, coste de bienes vendidos (COGS), gastos operativos, EBITDA, ingresos y gastos financieros, impuestos, y el resultado neto, además de otros campos específicos relevantes para diferentes industrias o modelos de negocio.

2. A continuación, llevamos a cabo la extracción de los datos de los campos mapeados; estos son campos estandarizados que esperamos encontrar y que ya han sido identificados y estructurados previamente en nuestro sistema, facilitando así su recolección automática.

3. Para los campos que no han sido mapeados previamente o que son novedosos, utilizamos la técnica de búsqueda fuzzy para identificar y extraer la información relevante.

4. Por último, una vez recopilados todos los datos, los consolidamos en un formato unificado junto con el BS, que permite un análisis cómodo y la comparación entre diferentes PnL y periodos de tiempo paralelamente a la información de balance.

En este momento tenemos dos DataFrames: uno contiene la información estructurada del BS y el PnL y en otro la información desestructurada de estos mismos.

Nuevo proceso de MoM

El proceso para el análisis Month-over-Month (MoM) sigue la misma estructura detallada previamente en el proceso de extracción de información para el Profit and Loss Statement (PnL). La principal adaptación que se realiza es en la función de selección de fechas, la cual se ajusta para incluir y comparar los datos correspondientes a los últimos tres años, garantizando así una visión más amplia de las tendencias y cambios mensuales en los indicadores financieros relevantes.

Al igual que en el caso del PnL, continuamos con la recopilación y consolidación de datos, pero con un foco especial en el seguimiento y la comparación de la evolución a corto plazo que revelan los análisis MoM.

Esto es posible ya que la búsqueda no se limita a tres campos, sino que a 36 representando los 36 meses dentro de esos 3 años de búsqueda que queremos obtener.

Esta información la almacenamos en dos nuevos DataFrames. El primero es la información estructurada de MoM y el segundo, como es de esperar, la información desestructurada del MoM.

Por lo tanto, en este momento tenemos un total de 4 DataFrames, cada uno con información especifica y diferente.

Nuevo proceso de CnD

1. **Proceso de Preparación de los Datos**: Empezamos por un proceso de recorte de los archivos Excel originales. Esta etapa es fundamental para conseguir la estructura de datos necesaria que permita un procesamiento eficiente en DataFrames. Este procedimiento requiere una atención detallada para no perder información relevante de las empresas, que en este caso son nombradas directamente y no mediante campos estándares.

2. **Creación de DataFrames para Acreedores y Deudores**: Una vez ajustados los datos en Excel, procedes a la construcción de dos DataFrames distintos, uno para los acreedores y otro para los deudores. La identificación de los límites de cada uno es crucial y se fundamenta en criterios preestablecidos que permiten separar de manera precisa la información pertinente a cada grupo.

3. **Identificación de Empresas y Asociación de Deudas**: En esta etapa, focalizas en identificar las compañías listadas en cada uno de los DataFrames. Este paso involucra un análisis meticuloso para enlazar cada empresa con las deudas asociadas en función del eje temporal respectivo. La información detallada de las deudas, como montos y fechas de vencimiento, se almacena cuidadosamente para su evaluación y uso posterior.

Este procedimiento manifiesta un enfoque manual y personalizado para analizar la información financiera de las empresas relacionadas con los paquetes de crédito. Sin campos estandarizados, el proceso requiere una comprensión profunda y un manejo cuidadoso de los datos para asegurarse de que la información de acreedores y deudores se mantenga actualizada y precisa. Este enfoque minucioso es vital para garantizar la calidad y la integridad de los datos que serán fundamentales para las decisiones de crédito subsecuentes.

Al final de esta etapa podemos obtener un DataFrame con los acreedores de una empresa indicada y los periodos de deudas en la que se encuentra la información. Esto nos permitirá un pequeño vistazo a la relación que tienen nuestros clientes con sus deuda.

En este momento aparece un nuevo DataFrame que contiene unificada la información de los deudores y los acreedores. En total tenemos un total, ahora mismo de 5 DataFrames diferentes.

Adaptar BS para crear SPV

El proceso se centra en la transformación de la primera tabla descrita, para que estos puedan ser incorporados en una tabla específica que se usa para calcular la probabilidad de impago o default.

1. **Filtrado de Columnas**: Una vez que tienes los datos de la tabla original, el siguiente paso es filtrar las columnas que nos interesan. Esto significa eliminar toda la información que no es necesaria para el análisis que vas a realizar. La filtración de columnas ayuda a simplificar el conjunto de datos y facilita los cálculos posteriores.

2. **Renombrar Columnas**: Después de filtrar las columnas, necesitarás cambiar el nombre de las mismas para que coincidan con los nombres de las columnas de la tabla existente donde se va a hacer el cálculo de probabilidad de default. Es crucial que los nombres de las columnas coincidan exactamente, de lo contrario, el sistema o la herramienta que utilices podría no reconocer los datos.

   Por ejemplo, si en el Balance Sheet la columna de 'Total Assets' se etiquetó como 'TA', pero en la tabla existente la columna correspondiente se llama 'Assets_Total', entonces necesitarás cambiar el nombre de 'TA' a 'Assets_Total'.

2.2.2. Transformation

El proceso de transformación de datos para su adecuación a las normas y estructuras de Google Cloud Platform (GCP) se centra en tres partes específicas, que permitirán una integración limpia y eficaz de los datos:

1. **Estandarización de Nombres de Columnas**: La primera parte del proceso es crucial ya que se trata de la estandarización de los nombres de las columnas en los DataFrames. Este paso es importante para garantizar que no haya caracteres no estándar o espacios adicionales que puedan causar errores al procesar o consultar los datos en GCP. Los caracteres especiales como los acentos, guiones, espacios, símbolos y mayúsculas en algunos casos, deben eliminarse o reemplazarse siguiendo un patrón constante para todas las columnas. Esto puede implicar convertir todos los nombres de columnas a minúsculas y utilizar guiones bajos para separar palabras. Por ejemplo, una columna titulada .Añño-Producción"podría estandarizarse a .anyo_production".

2. **Transformación de Tipos de Columnas**: La segunda parte se centra en la conversión de los tipos de datos de las columnas para alinearlos con los tipos admitidos por GCP. Aquí es donde los datos numéricos, fechas, cadenas de texto y otros formatos se ajustan para coincidir con los correspondientes tipos de GCP como `INTEGER`, `FLOAT`, `STRING`, `DATE`, `TIMESTAMP`, etc. Esta transformación garantiza que, al cargar los datos en servicios como BigQuery, todas las columnas tengan los tipos de datos correctos para

consultas eficientes y sin errores de interpretación. Por ejemplo, si tienes una columna que contiene fechas en formato de cadena, necesitarás convertirla al tipo de datos `TIMESTAMP` o `DATE` de BigQuery.

3. **Registro de Fecha de Ejecución**: La tercera parte del proceso apunta a añadir un campo adicional en todos los DataFrames que indique la fecha de ejecución. Este procedimiento es vital para mantener un control sobre cuándo se procesaron los datos y para mantener una línea de tiempo de los registros procesados. Al añadir una columna con la fecha de ejecución, se pueden realizar fácilmente consultas históricas y análisis de tendencias. Esta columna se puede añadir con el valor actual de la fecha y hora al momento de la ejecución del proceso de transformación, asegurando consistencia en los registros.

2.2.3. Load

La etapa final del proceso de ETL, conocida como "Load.º Çarga", se enfoca exclusivamente en la transferencia de los DataFrames ya procesados hacia sus destinos finales dentro de la infraestructura de Google Cloud Platform (GCP). Esta fase es crucial, ya que es el momento en que los datos se hacen accesibles para el análisis y la toma de decisiones.

Durante la etapa de carga, se emplean mecanismos y servicios específicos de GCP para asegurar que la inserción de los datos sea eficiente y segura. Se pueden utilizar diversos productos de almacenamiento de GCP, como BigQuery, Google Cloud Storage o Cloud SQL, dependiendo de la naturaleza de los datos, pero para el caso de esta aplicación, usaremos tablas de BigQuery.

Es aquí donde se llevan a cabo las últimas comprobaciones de calidad, asegurando que los datos cargados sean consistentes con los esquemas predefinidos y que estén listos para ser utilizados por usuarios finales y aplicaciones de negocio. Además, se pueden implementar automatizaciones en esta etapa para que la carga de datos sea recurrente, permitiendo así una actualización constante de los datos en los sistemas de GCP.

El éxito de la etapa de carga es determinante para la eficiencia de los procesos de ETL en su conjunto. Por tanto, se pone especial atención en optimizar los tiempos de carga, en manejar adecuadamente las actualizaciones incrementales y en garantizar la integridad de los datos durante todo el proceso de transferencia.

2.2.4. Desplegar y crear proceso semanal

Como parte del proyecto, la implementación de nuestra aplicación Streamlit en Google Cloud Platform (GCP) es un paso clave. Este despliegue nos permitirá proveer a los usuarios un acceso continuo y consistente a la aplicación, así como garantizar un rendimiento optimizado. A continuación, se detallan los pasos para llevar a cabo este despliegue estratégicamente y cómo convertirlo en un proceso automatizado que ocurre de manera semanal.

Paso 1: Preparación de la Aplicación Streamlit

Antes del despliegue en GCP, es esencial confirmar que nuestra aplicación Streamlit opere de forma impecable en un entorno de desarrollo local. Este proceso no solo incluye una revisión exhaustiva del código fuente en busca de errores sintácticos o lógicos, sino también una exhaustiva verificación funcional que garantice que todas las características y flujos de trabajo de la aplicación estén actuando según se espera.

Durante esta etapa preliminar, también llevaremos a cabo tests automatizados que cubran distintos escenarios de uso y condiciones fronterizas para asegurarnos de que todas las partes de la aplicación son robustas y a prueba de fallos. Las pruebas automatizadas son una parte fundamental de las buenas prácticas de desarrollo de software, ya que ayudan a identificar problemas que de otra manera podrían quedarse sin detectar hasta etapas más avanzadas del desarrollo o incluso después del despliegue.

Además, profundizaremos en la revisión del código para garantizar que sigue las mejores prácticas de programación y estándares de calidad. Esto incluye la implementación de principios de diseño de software, tales como el mantenimiento de un código limpio y bien documentado, el uso de patrones de diseño donde sea apropiado y la adhesión a principios SOLID para una mejor estructuración y mantenimiento del código.

Por otro lado, nos ocuparemos de las dependencias del proyecto. Toda aplicación moderna se apoya en bibliotecas externas para funcionar correctamente, y la nuestra no será la excepción. Compilaremos una lista detallada de estas dependencias en un archivo `requirements.txt`. Este archivo es vital porque especifica exactamente qué librerías y versiones necesita nuestra aplicación para ejecutarse. La creación de este archivo se puede realizar mediante herramientas como `pip freeze`, lo cual asegura que las mismas versiones de las bibliotecas que funcionan localmente serán utilizadas en el entorno de producción, minimizando así

los problemas de incompatibilidad entre entornos.

Este archivo `requirements.txt` nos facilitará el proceso de construcción de nuestra imagen de Docker más adelante, ya que será usado para instalar automáticamente todas las dependencias cuando se construya la imagen del contenedor. Este procedimiento asegura que nuestra aplicación en el entorno de producción refleje fielmente el entorno de desarrollo local, creando coherencia y previsibilidad en el comportamiento de la aplicación una vez desplegada.

La meticulosa preparación y la validación de la aplicación Streamlit antes del despliegue son pasos que no podemos obviar, ya que nos colocan en la mejor posición posible para un lanzamiento exitoso en la plataforma de Google Cloud. Estos esfuerzos preliminares son una inversión que redundará en la estabilidad, seguridad y facilidad de mantenimiento de la aplicación a largo plazo.

Paso 2: Contenedorización con Docker

Para garantizar que nuestra aplicación Streamlit sea fácilmente desplegable en Google Cloud Platform (GCP), la encapsularemos en un contenedor utilizando Docker. Este paso es fundamental para crear un entorno de ejecución unificado, independiente del sistema subyacente, permitiendo que nuestra aplicación funcione igual, independientemente de donde se ejecute. La construcción del contenedor comienza con la creación de un `Dockerfile`. Este archivo actúa como un plano instructivo que detalla cómo se ensamblará la imagen del contenedor, incluyendo la base del sistema operativo (como una versión ligera de Linux y Python), las librerías y paquetes necesarios que serán instalados a partir de las dependencias especificadas en nuestro `requirements.txt`, así como la configuración de las variables de entorno y la definición del punto de entrada o el comando para ejecutar la aplicación.

Una vez definido el `Dockerfile`, procedemos a compilar la imagen del contenedor ejecutando un comando `docker build`. Este proceso crea un paquete que encapsula el entorno completo necesario para que la aplicación funcione. Como medida de control de calidad, antes de desplegar esta imagen en GCP, la ponemos a prueba en el entorno local usando `docker run`, asegurándonos de que todas las funcionalidades de Streamlit operan según lo planeado. A través de esta validación, podemos identificar y corregir cualquier problema antes del despliegue, lo que minimiza el riesgo de incompatibilidad y errores en el entorno de producción. Esta práctica nos provee la confianza necesaria para avanzar al siguiente paso crítico del proyecto, el despliegue en GCP, sabiendo que contamos con una aplicación robusta

y lista para ser utilizada por los usuarios finales.

Paso 3: Uso de Container Registry

Tras asegurar que la imagen del contenedor de nuestra aplicación Streamlit funcione adecuadamente, procederemos a subirla al Container Registry de Google Cloud Platform. Este paso lo realizaremos utilizando herramientas de GCP, en particular el comando `gcloud`, que nos permitirá empujar la imagen al registro tras autenticarnos y etiquetar la imagen con el nombre adecuado. Container Registry proporciona un servicio seguro y privado para alojar nuestras imágenes, lo que facilita el manejo de nuestras versiones y el despliegue en el cloud.

Subiendo la imagen a Container Registry, estableceremos la base para un proceso de despliegue ágil y seguro en GCP. Además, esta acción nos abre el camino para implementar una estrategia de integración y despliegue continuos, automatizando las actualizaciones de nuestra aplicación y habilitando un despliegue escalable y eficiente a través de servicios como Google Kubernetes Engine o Google Cloud Run. Esto es esencial para mantener nuestra aplicación actualizada y para adaptar su escala según las demandas de uso.

Paso 4: Despliegue Continuo de la Aplicación

Realizaremos el despliegue de la aplicación utilizando la imagen alojada en Container Registry. Dependiendo del servicio de hosting de GCP que hayamos seleccionado, el proceso de despliegue puede variar, pero siempre aseguraremos que el entorno de producción refleje fielmente lo que hemos probado localmente.

Paso 5: Automatización de la Actualización Semanal

Para mantener la aplicación actualizada sin intervención manual, implementaremos un flujo de trabajo automatizado con Cloud Scheduler. Este programará y desencadenará un trabajo de Cloud Run que se encargará de actualizar la aplicación. Configuraremos este trabajo para que se ejecute una vez a la semana, lo que permitirá que nuestra aplicación esté siempre al día con las últimas modificaciones y datos más recientes.

De esta forma ya satisfacemos la necesidad de un proceso semanal y on demand que nos permite tener dos servicios desplegados con sus correspondientes dockerfiles para que los servicios funcionen sin problemas.

Paso 6: Monitoreo y Mantenimiento

Por último, es fundamental que mantengamos un registro preciso del desempeño y las métricas de la aplicación mediante Google Cloud Monitoring y Logging. También revisaremos periódicamente las políticas de seguridad para asegurar que sólo los usuarios autorizados puedan acceder a la aplicación y realizar cambios en ella.

En resumen, este enfoque nos permitirá desplegar de forma eficiente nuestra aplicación Streamlit en GCP y mantenerla actualizada a demas de tener un proceso semanal, garantizando así el suministro constante de mejoras y correcciones necesarias para el éxito del proyecto.

2.3. Desarrollo de Web App

El desarrollo de la web app es un elemento crucial en la automatización de procesos para los analistas de crédito que buscan eficiencia y precisión en la extracción de datos financieros de empresas. La aplicación está diseñada con un enfoque específico en el usuario final, es decir, los expertos financieros que requieren una herramienta capaz de ejecutar manualmente el complejo proceso de extracción, transformación y carga de datos (ETL).

Para facilitar la interacción del usuario, se ha optado por utilizar el servicio de Streamlit, una solución eminentemente práctica en el desarrollo de aplicaciones web Python. Streamlit se caracteriza por su capacidad para crear interfaces de usuario atractivas, modernas y, sobre todo, funcionales con relativa facilidad. Esto permite a los desarrolladores centrar sus esfuerzos en el funcionamiento interno del sistema, conocido como el backend, sin descuidar la experiencia del usuario final. La estructura de la aplicación es intencionadamente sencilla y sin adornos innecesarios, lo que simplifica la curva de aprendizaje y facilita una adopción más rápida por parte de los analistas.

La aplicación tiene como una de sus principales funciones la capacidad de recibir un link directo de una hoja de cálculo de Google Sheets. Esto es fundamental ya que muchas empresas almacenan su información financiera en esta plataforma. El analista simplemente provee el ID del archivo Google Sheets deseado, y la aplicación inicia de manera eficiente el proceso de ETL. Este consiste en extraer los datos brutos, transformarlos según sea necesario (limpieza, normalización, asignación de tipos de datos, etc.) y cargarlos a un sistema donde se puedan almacenar

y manipular con mayor facilidad. En este caso, el proceso culmina con la carga de los datos en un DataFrame especializado.

El siguiente paso crítico en la funcionalidad de la app es la evaluación financiera de los datos. La aplicación no solo extrae datos, sino que también ejecuta análisis sofisticados para calcular la "probability of default" (probabilidad de impago) de la empresa evaluada. Este indicador es fundamental en el mundo de las finanzas, ya que proporciona una medida cuantitativa del riesgo de que la empresa en cuestión no cumpla con sus obligaciones financieras. La metodología para calcular este indicativo se basa en modelos estadísticos y financieros que procesan el DataFrame obtenido durante el proceso de ETL.

Una vez que la información sobre la probabilidad de impago ha sido calculada, la aplicación presenta los resultados al analista de una manera clara y concisa. Esto permite a los profesionales crear evidencia documental de que el proceso de análisis se ha completado satisfactoriamente. Además, la app garantiza que todos los datos procesados se almacenan de manera segura y correcta en BigQuery, la plataforma de almacenamiento en la nube de Google. Dicho almacenamiento no solo asegura la integridad de la información sino que también la hace accesible para análisis posteriores o para ser utilizada como parte de un sistema de inteligencia de negocios más amplio.

El diseño y la funcionalidad de la aplicación web son reflexivos y metódicos, proporcionando así un balance óptimo entre simplicidad y potencia. Al final, los analistas de crédito disponen de una herramienta que no solo ahorra tiempo y recursos sino que también aumenta la precisión de sus análisis y decisiones. Con una interfaz intuitiva, un proceso de ETL transparente y un análisis de datos robusto, esta aplicación promete ser un cambio de juego para los profesionales del sector financiero.

La visualización de la aplicación es la siguiente figura 2.3

Como se puede apreciar hay un recuadro dónde se puede indicar a la herramienta el link del creditpack. Una vez se ha introducido el link, se pulsa el botón y este inicia todo el proceso de ejecución del servicio de ETL. Dando lugar a la siguiente figura: 2.4

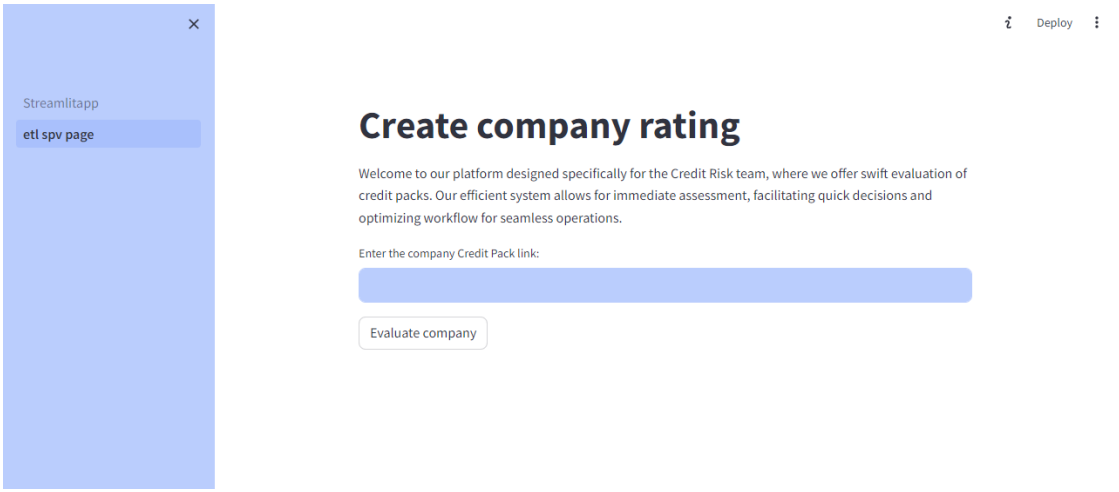

Figura 2.3: Rating de empresas y proceso ETL

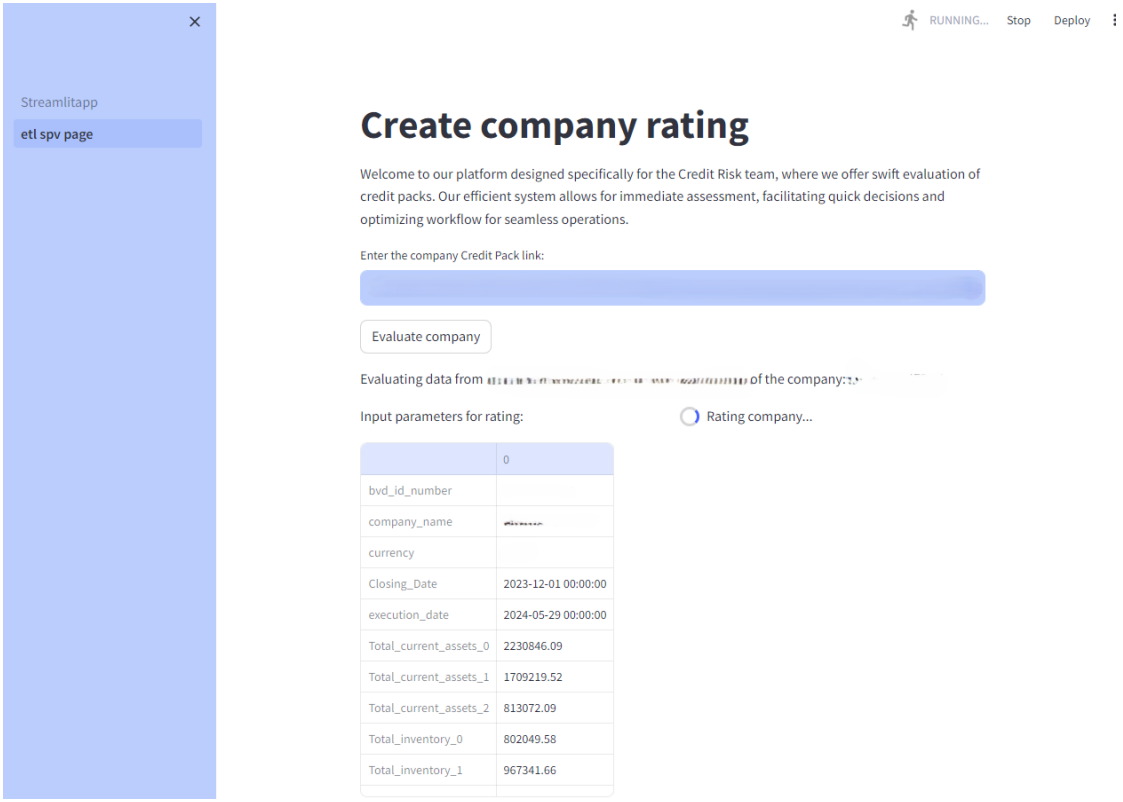

Figura 2.4: Streamlit ETL proceso.

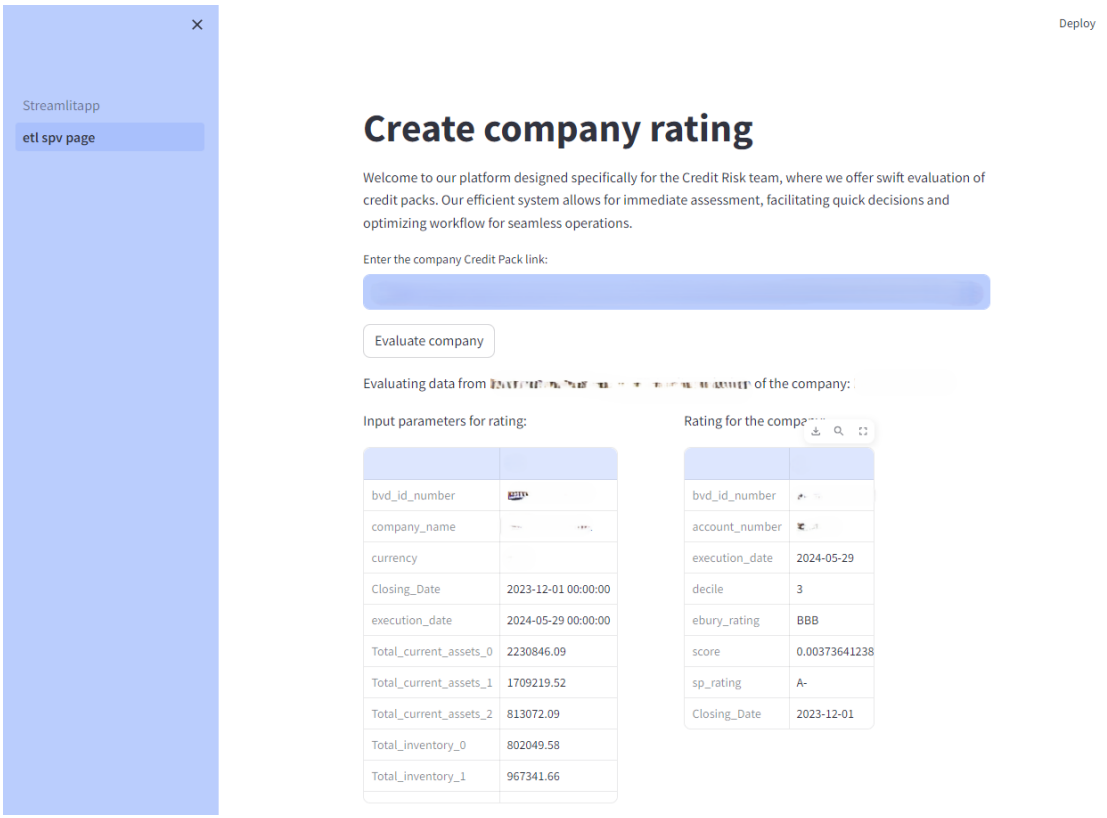

Figura 2.5: Rating process in Streamlit.

2.4. Próximos pasos

2.4.1. Desarrollo del Dashboard con Looker y Google BigQuery

3.1 Diseño Preliminar de Modelos de Datos en BigQuery El desarrollo del dashboard proyectado se fundamentará en un diseño lógico y estructurado de un modelo de datos hospedado en Google BigQuery. Este modelo proporcionará la relación y jerarquía necesaria para que Looker pueda mapear efectivamente los campos relevantes a nuestros KPIs financieros. El modelo constituirá el armazón del repositorio de datos y deberá ser construido con miras a maximizar la eficiencia y la optimización de las consultas futuras.

3.2 Definición Prospectiva de Métricas y KPIs En el umbral de la construcción del dashboard, una articulación precisa de métricas y KPIs financieros será formulada y establecida. Tales indicadores se pronostican para medir el rendimiento y viabilidad de los credit packs y para contribuir en un entendimiento más profundo de las tendencias y relaciones financieras. La configuración de estos se hará siguiendo los estándares analíticos y serán adaptados para su visualización y análisis a través de Looker.

3.3 Desarrollo Planeado de la Aplicación de Dashboard con Looker Looker, como plataforma primaria para la creación de nuestro dashboard, nos ofrecerá una interfaz flexible y poderosa para el diseño y la gestión de visualizaciones. Se utilizará LookML, el lenguaje de modelado de Looker, para definir las relaciones de los datos, y se crearán dashboards y reportes que sean escalables, accesibles y de alto rendimiento. La programación del dashboard prevé la utilización de componentes personalizados y Looker Blocks para agilizar el desarrollo.

3.4 Diseño Gráfico y Visualización de Datos La fase visual implica el diseño de diagramas, gráficos y tableros intuitivos con una estrecha colaboración entre analistas de datos y diseñadores UI/UX para garantizar que las visualizaciones transmitan la información de forma efectiva y coherente. Se elegirán cuidadosamente los gráficos y las paletas de colores, asegurándose que estarán en consonancia con los principios de gestión de la información y las mejores prácticas de diseño.

3.5 Facilitación de Interactividad al Usuario Uno de los aspectos cruciales del dashboard será la capacidad para que el usuario interactúe con los datos

presentados. Se desarrollarán funciones interactivas como filtros variables, búsquedas y selección de detalles, que permiten a los usuarios personalizar su análisis y profundizar en dimensiones específicas de los datos financieros.

3.6 Salvaguardas de Seguridad y Acceso El diseño y la implementación del dashboard tendrán un énfasis significativo en la seguridad. En Looker, el nivel de acceso y control de usuario será cuidadosamente configurado para garantizar que los datos financieros críticos estén protegidos y que solo personal autorizado podrá acceder a la información detallada.

3.7 Estrategia de Pruebas y Optimización de Rendimiento Un procedimiento comprensivo de pruebas se delineará para asegurar que tanto el modelado de datos como las visualizaciones del dashboard sean precisas. Esto incluirá la optimización de las consultas SQL en BigQuery y la funcionalidad del dashboard en Looker, con el objetivo de minimizar los tiempos de carga y maximizar la eficiencia de la recuperación de datos.

3.8 Implementación Estratégica y Mantenimiento Previsto El despliegue del dashboard será una operación coordinada que contemple una transición sin fricción al nuevo sistema. Acompañando al lanzamiento, se prevé un programa de mantenimiento que atienda actualizaciones necesarias, solicitudes de características adicionales, y ajustes en base a los comentarios recibido por los usuarios. Este ciclo de mejora continua asegurará que el dashboard permanezca relevante y valioso en el tiempo.

2.4.2. Evolución del modelo de evaluación de las empresas

La elección de modelos de regresión logística o de árboles de decisión se orienta a dotar a los analistas de crédito de un entorno comprensible y transparente para evaluar el riesgo de impago de las empresas. Ambas metodologías prometen una interpretabilidad superior en comparación con modelos más complejos de machine learning, lo cual se traduce en un mayor grado de confianza en las decisiones que se basan en sus predicciones.

La **regresión logística** es particularmente benévola en términos de su relación directa entre las variables de entrada y la probabilidad de un evento como el default. Posee una forma funcional sencilla que permite entender cómo cada factor impacta en el riesgo. Por ejemplo, un coeficiente más alto indica un mayor efecto sobre la probabilidad de impago. Esta linealidad y simplicidad estadística hacen que sea fácilmente explicable a partes interesadas con o sin formación técnica en

estadísticas.

Por otro lado, los **árboles de decisión** ofrecen una representación gráfica de las decisiones basada en reglas simples de decisión que se parecen al razonamiento humano. Este método divide progresivamente el dataset en base a criterios precisos que pueden ser rastreados a lo largo del árbol. Así, es posible explicar los motivos detrás de una calificación específica de riesgo, inspeccionando las divisiones y condiciones a lo largo del árbol.

Ambas técnicas permiten una implementación rápida, donde la preparación de datos y la selección de características pueden realizarse con herramientas convencionales de pre-procesamiento.

En la **implementación** de estos modelos, se emplearían las siguientes etapas:

- **Selección de Datos**: selección de variables financieras y no financieras relevantes de los datasets existentes.
- **Pre-procesamiento**: limpieza, normalización y transformación de datos para su adecuación en el modelo.
- **Construcción del Modelo**: utilizando herramientas, como `scikit-learn` en Python, se entrena el modelo con un conjunto de datos históricos.
- **Evaluación**: validación del rendimiento del modelo en base a métricas adecuadas de clasificación.
- **Visualización e Interpretación**: desarrollo de reportes o dashboard donde los resultados y pesos de las variables (para regresión logística) o la estructura del árbol y las reglas de división (árboles de decisión) pueden ser examinados por los analistas de crédito.

Optar por cualquiera de estas dos alternativas favorece un análisis accesible y una mayor adopción y confianza en el modelo, ya que cada factor de riesgo puede ser estudiado y justificado con detalle. Además, simplifica la colaboración y comunicación entre analistas técnicos y partes interesadas del lado de los negocios. Con todo esto, se detona un enfoque más democrático y transparente en la evaluación de solvencia de las empresas.

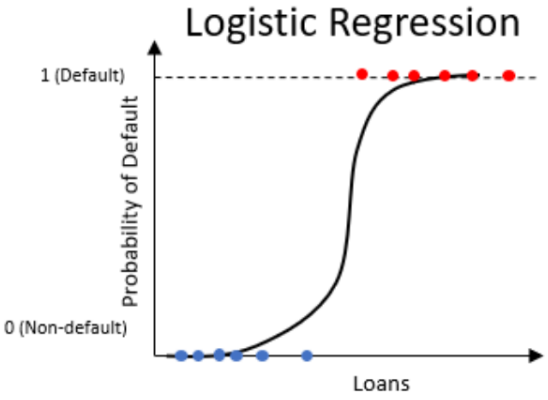

Figura 2.6: Ejemplo de un modelo de regresión logística.

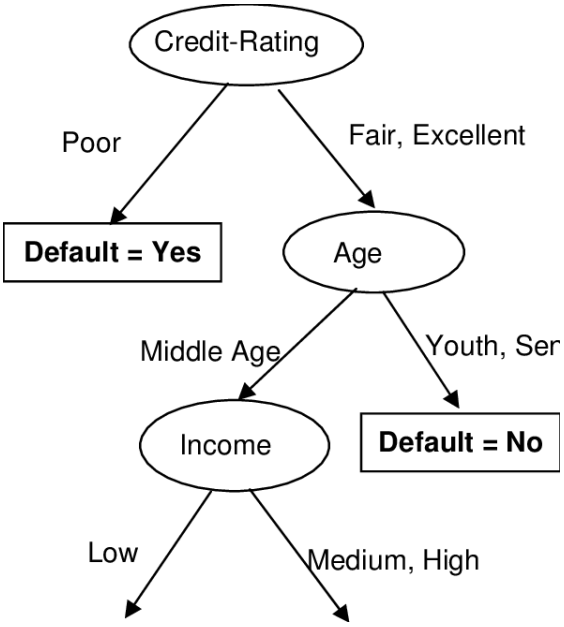

Figura 2.7: Ejemplo de un modelo de árbol de decisión.

2.4.3. Alternativa al Proceso ETL Tradicional

La alternativa propuesta integra herramientas de OCR basadas en aprendizaje profundo para la extracción inicial del texto desde PDFs escaneados, seguido de técnicas de NLP avanzadas que no solo transforman el texto en datos estructurados sino que también permiten la clasificación y comprensión de la semántica financiera involucrada. En cuanto a la última etapa , los modelos desarrollados podrán integrarse con sistemas de base de datos existentes, asegurando la carga de datos precisos y contextualmente enriquecidos.

1. **Extracción**: Construcción de un pipeline de OCR que utilice modelos de Machine Learning capacitados para interpretar con alta exactitud una amplia variedad de fuentes de texto financiero contenido en PDFs.

2. **Transformación**: Entrenamiento de modelos de NLP para procesar y estructurar el texto obtenido. Esto incluye la identificación de entidades financieras, normalización de formatos de datos y deducción de relaciones entre datos. Se valorará el uso de algoritmos preentrenados y redes neuronales especializadas.

3. **Carga**: Desarrollo de interfaces y adaptadores para integrar los datos procesados con sistemas de almacenamiento de datos, manteniendo la integridad y la conformidad con los esquemas de datos establecidos.

Ventajas sobre ETL Tradicional:

- **Automatización**: Minimiza la manipulación manual de datos, disminuyendo los errores de programación y el tiempo de procesamiento.
- **Flexibilidad**: Mejora la capacidad de gestionar diferentes formatos y disposiciones de documentos.
- **Escalabilidad**: Capacidad de procesar rápidamente grandes volúmenes de documentos.
- **Precisión**: Disminuye las tasas de error en la captura y clasificación de datos financieros.

El enfoque de ML en este componente del proyecto aspira a demostrar viabilidad técnica y mejoras funcionales sobre los métodos ETL tradicionales usados. Anticipamos que el sistema resultante, previa validación con conjuntos de datos de prueba y pruebas de estrés, ofrecerá ventajas notables en cuanto a eficacia y eficiencia en la gestión de documentos financieros en PDF, alineándose con los

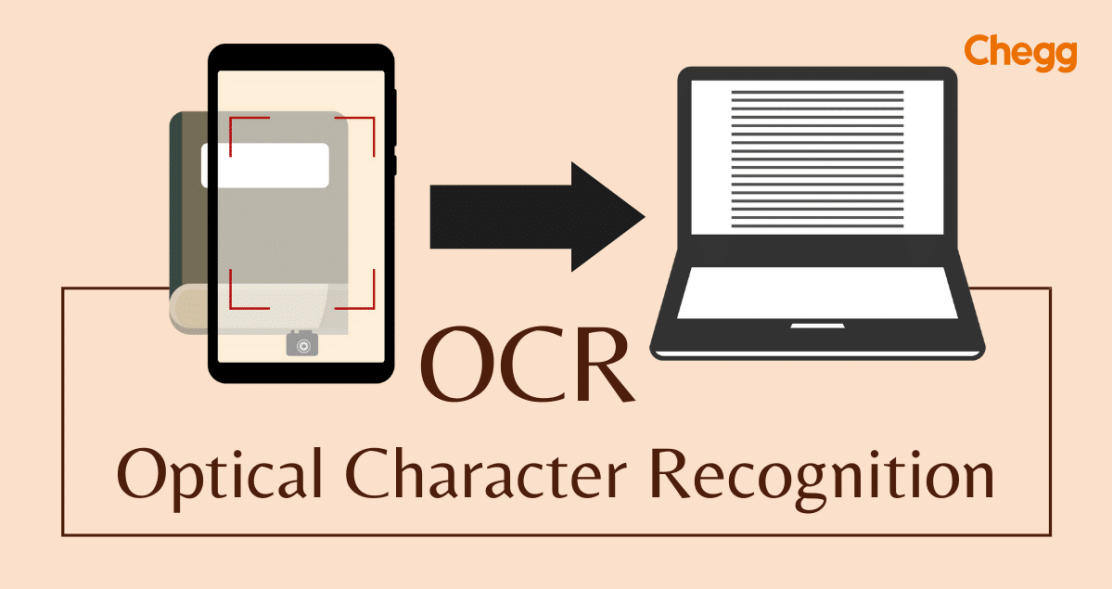

Figura 2.8: Place holder de un OCR.

objetivos estratégicos de modernización digital y optimización de procesos en el ámbito financiero.

La implementación de sistemas OCR para la extracción de datos financieros de documentos PDF puede enfrentarse a desafíos críticos, entre los cuales destacan:

1. **Calidad del Documento**: La precisión del OCR es altamente dependiente de la nitidez y claridad del texto en los documentos PDF, siendo vulnerables a errores con textos borrosos o distorsionados.

2. **Complejidad y Diseño de Documentos**: Los documentos con diseños complejos, como múltiples columnas y tablas, pueden entorpecer la correcta interpretación y extracción de datos.

3. **Costos de Implementación**: La adquisición, personalización y mantenimiento de sistemas OCR de alta calidad representan una inversión significativa.

4. **Seguridad y Privacidad**: El manejo de información confidencial mediante OCR obliga a adoptar medidas rigurosas de seguridad para proteger los datos contra accesos no autorizados y brechas de información.

Capítulo 3

Conclusiones y resultados finales

En conclusión, el proyecto desarrollado ha supuesto una auténtica transformación para la empresa, redefiniendo no solo la manera en la que operamos sino también el impacto en la moral y la eficiencia del personal. La implementación de Streamlit como herramienta forma parte de esta metamorfosis digital, proporcionando una solución práctica y efectiva a los desafíos de manejar el 'Credit Pack'.

Este avance resulta en una significativa aceleración en el flujo de trabajo diario, liberando a los colaboradores de las tediosas y repetitivas tareas asociadas con el análisis manual. Ahora, se pueden abordar con mayor facilidad análisis de datos críticos de manera dinámica, lo que permite a los equipos enfocarse en actividades de valor añadido que aprovechan su experiencia y conocimientos especializados.

El uso de Streamlit nos ha habilitado para realizar un crecimiento efectivo y preciso de empresas de relevancia estratégica, mejorando significativamente la capacidad de tomar decisiones basadas en datos. Ahora, podemos recolectar y procesar conjuntos de datos financieros de manera masiva y sistemática, lo cual es vital para anticipar tendencias del mercado y adaptar nuestras estrategias de negocio de manera proactiva.

No menos importante es el impulso que esta herramienta ha dado a la cultura de la innovación dentro de la empresa. Al adoptar tecnologías de vanguardia, reafirmamos nuestro compromiso con la mejora continua y establecemos un entorno que fomenta la curiosidad, el aprendizaje y la experimentación entre los miembros del equipo.

Mirando hacia adelante, el éxito de este proyecto proporciona un sólido fundamento sobre el cual construir. Nos brinda confianza en nuestra capacidad para enfrentar futuros desafíos mediante la aplicación de soluciones tecnológicas inno-

vadoras. Con una continua iteración y mejora de nuestras herramientas y procesos, el impacto duradero de este cambio será un aumento observable tanto en la satisfacción del personal como en la rentabilidad y sustentabilidad del negocio.

3.1. Estructura final del sistema

La estructura actual con la que esta montado el sistema se puede ver representado enlos siguientes diagramas:

- Proceso actual del ETL: Figura 3.1
- Proceso actual del ETL y el rating: Figura 3.2
- proceso actual de la extracción semanal: Figura 3.3

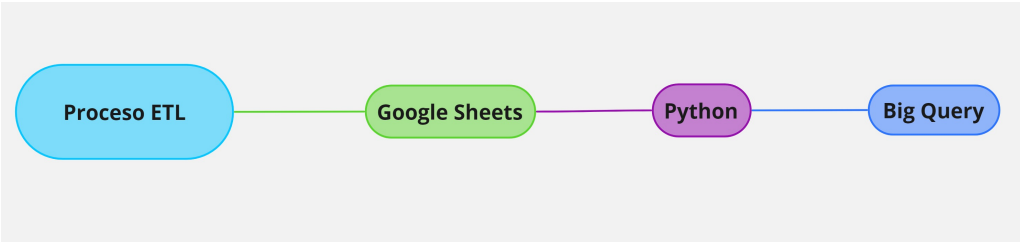

Figura 3.1: Proceso ETL actual.

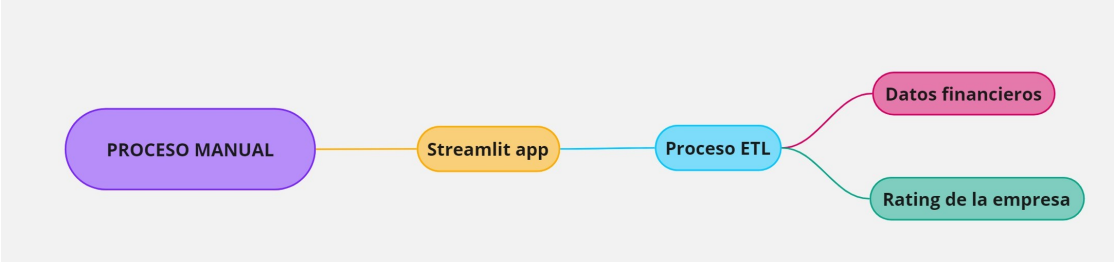

Figura 3.2: Proceso manual para el ETL y Rating actual

3.2. Estructura futura del sistema

En la estructura futura el único cambio relevante al sistema es el proceso del ETL, que se basara en OCR's. Figura 3.4

Figura 3.3: Proceso semanal para el ETL actual

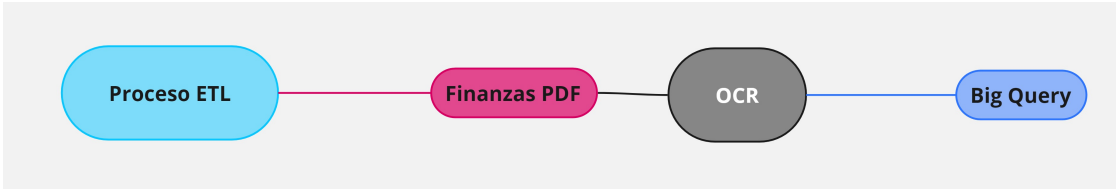

Figura 3.4: Futuro proceso con OCR

Capítulo 4

Bibliografía

1. Streamlit

- **Fuente:** Sitio web oficial de Streamlit.
- **Enlace:** `https://streamlit.io/`
- **Descripción:** Streamlit es una biblioteca de Python que permite la creación rápida y sencilla de aplicaciones web interactivas, especialmente útil para visualización de datos y desarrollo de aplicaciones de aprendizaje automático.

2. Documentación de Python

- **Fuente:** Python Software Foundation.
- **Enlace:** `https://www.python.org/doc/`
- **Descripción:** La documentación oficial de Python proporciona recursos exhaustivos, incluyendo guías, tutoriales y referencia completa de la biblioteca estándar de Python, facilitando el aprendizaje y desarrollo en este lenguaje de programación.

3. Google Cloud Platform (GCP)

- **Fuente:** Google Cloud.
- **Enlace:** `https://cloud.google.com/?hl=es_419`

- **Descripción:** Google Cloud Platform ofrece una amplia gama de servicios en la nube, desde almacenamiento hasta análisis de datos y aprendizaje automático. La documentación proporcionada por Google Cloud facilita la comprensión y el uso eficaz de estos servicios.

4. Documentación de Docker

- **Fuente:** Docker Documentation.
- **Enlace:** `https://docs.docker.com/`
- **Descripción:** Docker es una plataforma de software que simplifica la creación, implementación y ejecución de aplicaciones en contenedores. La documentación oficial de Docker ofrece una guía detallada para el uso efectivo de esta tecnología.

5. Google Sheets

- **Fuente:** Google.
- **Enlace:** `https://www.google.com/sheets/about/`
- **Descripción:** Google Sheets es una herramienta de hojas de cálculo en línea que permite la colaboración en tiempo real. La página de inicio proporciona información detallada sobre las características y funcionalidades de Google Sheets, así como recursos para su uso efectivo.