Diseño e Implementación de un Sistema de Clasificación Contable Automatizado para Facturas: Una Aproximación Mediante Recurrencias e Inteligencia Artificial

Juan Carlos Baján Castro

UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Diseño e Implementación de un Sistema de Clasificación Contable Automatizado para Facturas: Una Aproximación Mediante Recurrencias e Inteligencia Artificial

Trabajo de graduación en modalidad de Trabajo Profesional presentado por
Juan Carlos Baján Castro
Para optar al grado académico de Licenciado en Ingeniería en Ciencias de la Computación y Tecnologías de la Información

Guatemala, Octubre del 2024

Vo.Bo.:

(f) _____

Alan Gerardo Reyes Figueroa

Tribunal Examinador:

(f) _____

Alan Gerardo Reyes Figueroa

(f) _____

Alfonso Rodriguez Ongay

(f) _____

Eddy Omar Castro Jauregui

Fecha de aprobación: Guatemala, Noviembre.

Prefacio

En los últimos años, hemos sido testigos de una transformación sin precedentes en la forma en que las empresas operan y compiten. La irrupción de tecnologías digitales ha redefinido los mercados, alterado las expectativas de los clientes y creado nuevos modelos de negocio. Cada sector ha sido impactado, para bien o para mal, por la velocidad con la que el mundo ha cambiado. Esta tesis nace del deseo de contribuir, aunque sea modestamente, a la innovación y facilitar la vida cotidiana de las personas.

La concepción de este proyecto surgió del anhelo de identificar un nicho donde aplicar las estrategias de innovación y automatización aprendidas durante mi formación universitaria. A lo largo de diversas conversaciones con profesionales de distintas áreas, una oportunidad en particular captó mi atención debido a su proximidad y a la posibilidad de probar las diferentes fases del proyecto con mayor facilidad.

Deseo expresar mi más sincera gratitud al equipo de Pale por abrirme las puertas de sus oficinas, permitirme implementar los avances logrados, brindarme retroalimentación y confiar en mi capacidad de desarrollo para mejorar la calidad de su trabajo.

Confío en que este proyecto sea un ejemplo de cómo la combinación de una buena idea con un desarrollo adecuado puede renovar un sector del mercado, mejorando así la eficiencia y la calidad del trabajo en Guatemala.

Juan Carlos Baján Castro

Agradecimientos

La culminación de esta tesis no habría sido posible sin el apoyo y la colaboración de muchas personas y organizaciones, a quienes deseo expresar mi más sincera gratitud.

En primer lugar, quiero agradecer profundamente a Dios por permitirme estudiar en la mejor universidad de Guatemala, por bendecirme con una familia que me apoya y anima, por proveerme en cada momento y darme las herramientas necesarias para cumplir con determinación las metas que me he propuesto.

Agradezco también a mis padres, quienes siempre han confiado en mí, me han animado a seguir adelante y han dado todo lo que tienen para permitirme llegar hasta este momento. Me han impulsado a dar el 110 % en todo y me han enseñado que con esfuerzo y dedicación todas las metas son alcanzables.

Quiero expresar mi agradecimiento a la Universidad por orientarme adecuadamente y proporcionarme todo el conocimiento necesario para proyectar ideas hacia la vida real. Gracias por darme las herramientas para construir un futuro mejor y contribuir al desarrollo de una Guatemala más próspera.

A mi asesor, quien sin dudarlo decidió apoyarme y guiarme técnicamente para cumplir con el objetivo. Agradezco su dedicación y tiempo para estudiar mi proyecto y pensar en formas de llevarlo aún más lejos.

A Pale, la organización que acogió mis ideas, me proporcionó las herramientas para implementarlas y confió en mi capacidad de desarrollo.

A mis compañeros de estudio y amigos, gracias por su apoyo constante, por las largas horas de estudio compartidas y por ser una fuente continua de motivación.

Finalmente, agradezco a todas las personas y entidades que, de una manera u otra, contribuyeron al desarrollo y culminación de esta tesis. Su ayuda y apoyo han sido invaluables, y esta tesis es tanto mía como suya.

Índice

Lista de Figuras

Lista de Cuadros

Resumen

La presente tesis aborda el diseño e implementación de un sistema automatizado de clasificación contable para facturas, utilizando técnicas de inteligencia artificial (IA) y aprendizaje automático. Este proyecto surge ante la creciente necesidad de optimizar procesos contables en un entorno empresarial cada vez más competitivo y digitalizado. Tradicionalmente, la clasificación de facturas es una tarea manual que consume tiempo, es propensa a errores y limita la capacidad de los profesionales contables para enfocarse en actividades de mayor valor estratégico.

El sistema desarrollado emplea algoritmos avanzados de procesamiento de lenguaje natural (PLN) y aprendizaje profundo para analizar, categorizar y asignar de manera automática cada producto o servicio a su cuenta contable correspondiente. La solución propuesta promete mejorar significativamente la eficiencia operativa, reduciendo hasta en un 80 % el tiempo dedicado a la clasificación manual, al tiempo que asegura un mayor grado de precisión en el cumplimiento de las normativas fiscales vigentes.

El proyecto ha sido implementado utilizando una infraestructura serverless en Google Cloud, lo que garantiza su escalabilidad y flexibilidad para adaptarse a las necesidades de empresas de diferentes tamaños. Además, se ha integrado un sistema de retroalimentación continua que permite realizar ajustes y actualizaciones regulares, asegurando la adaptabilidad del sistema frente a cambios en las normativas fiscales o nuevas necesidades de los usuarios.

Los resultados obtenidos demuestran el éxito del sistema automatizado al mejorar tanto la precisión como la velocidad en la clasificación de facturas. Con esta solución, las empresas pueden optimizar sus procesos contables, reduciendo costos operativos y mejorando la calidad del trabajo. Este proyecto representa un paso importante hacia la modernización del sector contable, mostrando cómo la integración de tecnologías emergentes puede transformar prácticas empresariales tradicionales.

CAPÍTULO 1

Introducción

En un mundo cada vez más digitalizado, la necesidad de innovar en sectores tradicionales es imperiosa. La contabilidad, como pilar fundamental en la gestión financiera y operativa de las empresas, no ha quedado exenta de esta transformación. A pesar de los avances tecnológicos de las últimas décadas, muchos procesos contables siguen siendo manuales, repetitivos y propensos a errores humanos. La automatización de estos procesos, especialmente en la clasificación de facturas, se presenta como una solución clave para optimizar tiempos, mejorar la precisión y liberar a los profesionales para que se concentren en tareas de mayor valor estratégico.

Este proyecto nace con el objetivo de diseñar e implementar un sistema automatizado de clasificación contable que utilice inteligencia artificial y aprendizaje automático. Este sistema está diseñado para procesar facturas de manera rápida y precisa, asignando cada producto o servicio a la cuenta contable correspondiente. Al reducir significativamente la intervención manual, se busca transformar un proceso tradicionalmente laborioso en uno eficiente y libre de errores, mejorando así la capacidad operativa de las empresas y contribuyendo a la evolución del sector contable.

La implementación de tecnologías avanzadas en la clasificación de facturas no solo tiene el potencial de mejorar la eficiencia de los procesos contables, sino que también representa un avance hacia la digitalización integral de los sistemas empresariales. Con un enfoque en la integración de procesamiento de lenguaje natural (PLN) y aprendizaje profundo, esta investigación explora nuevas fronteras en la automatización de tareas contables, ofreciendo una solución robusta y escalable para un entorno fiscal y empresarial cada vez más dinámico.

En este contexto, el presente trabajo de tesis busca no solo automatizar una parte crítica del ciclo contable, sino también contribuir al debate sobre la modernización de la contabilidad, demostrando cómo las herramientas de inteligencia artificial pueden desempeñar un papel transformador en la creación de soluciones más inteligentes y eficientes. Este proyecto, además, aspira a establecer un precedente en la implementación de tecnologías emergentes en áreas que tradicionalmente han sido conservadoras en su adopción de la digitalización.

CAPÍTULO 2

Objetivos

2.1. Objetivo General

Crear un sistema basado en inteligencia artificial que automatice la clasificación de facturas, asignando cada producto o servicio a la cuenta contable correspondiente con alta precisión, para transformar el proceso contable tradicional y aumentar la eficiencia operativa.

2.2. Objetivos Específicos

- Implementar tecnologías avanzadas en el clasificador para utilizar técnicas de aprendizaje automático y procesamiento de lenguaje natural que permitan entender y procesar automáticamente la información contenida en las facturas, reduciendo la necesidad de intervención manual.
- Realizar pruebas piloto con retroalimentación continua para evaluar el rendimiento del clasificador automático con un grupo selecto de empresas de contabilidad, recopilando retroalimentación para ajustes y mejoras que aseguren su efectividad y eficiencia.
- Desarrollar un marco de actualización y escalabilidad para establecer un sistema que permita actualizaciones regulares del clasificador en función de nuevos aprendizajes, cambios en la legislación fiscal y las necesidades cambiantes de los usuarios, garantizando su escalabilidad y adaptabilidad a largo plazo.

CAPÍTULO 3

Justificación

La imperiosa necesidad de innovar en sectores tradicionales, especialmente en áreas tan fundamentales y complejas como la contabilidad, se encuentra en el corazón de este proyecto. La transformación del proceso manual de clasificación de facturas mediante la automatización no solo promete eliminar errores humanos y optimizar tiempos, sino también liberar a los profesionales contables para enfocarse en tareas de mayor valor estratégico. Esta iniciativa es crucial para mantener la competitividad y eficiencia en un entorno empresarial cada vez más dinámico y exigente.

Al centrarse en un mercado familiar y seguro para el autor, gracias a sus relaciones preexistentes con empresas de contabilidad, el proyecto se sitúa en una posición privilegiada para implementar y perfeccionar una innovación disruptiva. Esta cercanía al campo de aplicación asegura una adaptación precisa a las necesidades del sector, permitiendo una transición suave hacia prácticas más modernas y eficientes. El deseo de cambiar el mercado a través de un producto innovador subraya la visión de futuro del proyecto y su compromiso con la redefinición de la contabilidad.

La implementación de tecnologías avanzadas como la inteligencia artificial en el proceso de clasificación de facturas no solo establece un precedente para la digitalización del sector contable, sino que también sirve como catalizador para la transformación digital en áreas conservadoras. Este enfoque refleja una comprensión profunda de la importancia de la innovación tecnológica en la creación de soluciones eficientes y sostenibles, posicionando al proyecto como un líder en la evolución hacia prácticas empresariales más inteligentes y estratégicas.

CAPÍTULO 4

Marco Teórico

4.1. Contabilidad

Desde tiempos inmemoriales, la humanidad ha desarrollado mecanismos para controlar sus bienes. Es inherente al ser humano cuantificar y calificar los recursos que maneja. Desde antiguas tablas de barro hasta modernos softwares de gestión, la diversidad de métodos para alcanzar este objetivo es evidente en todos los aspectos de la vida.

Con la evolución de la economía mundial, han emergido estándares y métodos básicos para gestionar este control. En el contexto de este proyecto, la contabilidad se aborda desde una perspectiva económica, más que administrativa, pues es esencial centrarse en los fundamentos de los registros para cumplir los objetivos propuestos.

La contabilidad es un sistema de información destinado al registro, evaluación y comunicación de información económica y financiera fundamental para la toma de decisiones y para proporcionar transparencia ante los entes reguladores fiscales locales [9] (Alcarria, 2009).

José Alcarria [9] sintetiza el proceso contable en cuatro pasos principales:

- Obtención de la Información
- Análisis y Valoración
- Registro de Hechos Contables
- Elaboración de Informes

Como se puede observar, es un proceso teóricamente automatizable. La mayoría de los proyectos de software operan de manera similar: capturan datos para su uso, presentan vistas para observar y analizar esos datos, mantienen un registro claro e íntegro de la información y finalmente cumplen un objetivo con esa información.

4.2. Fundamentos de la Contabilidad Financiera

Según José Alcarria [9], los informes en contabilidad financiera se elaboran utilizando únicamente cinco elementos básicos. De estos elementos derivan todos los informes y reportes que se deben presentar a las entidades y personas interesadas. Cada elemento debe documentarse detalladamente para tomar decisiones correctas y gestionar las cuentas con transparencia.

4.2.1. Activos

Conjunto de todos los bienes y derechos con valor monetario que son propiedad de una empresa, institución o individuo. [11] (ASALE & RAE, 2023)

4.2.2. Pasivos

Valor monetario total de las deudas y compromisos que gravan a una empresa, institución o individuo, y que se reflejan en su contabilidad. [11] (ASALE & RAE, 2023)

4.2.3. Patrimonio Neto

Conjunto de los bienes y derechos propios adquiridos por cualquier título. [11] (ASALE & RAE, 2023)

4.2.4. Ingresos

Caudal que entra en poder de alguien, y que le es de cargo en las cuentas. [11] (ASALE & RAE, 2023)

4.2.5. Gastos

Consumo que se hace de un recurso que provoca el incremento de pérdidas o la disminución de beneficios. [23] (iAhorro, 2024)

4.3. Hecho Contable

Para poder cumplir con el proceso contable, se deben registrar eventos llamados "Hechos Contables"los cuales son cualquier acontecimiento que influye o puede influir de manera significativa, tanto cuantitativa como cualitativamente, en el patrimonio de una entidad y puede ser registrado contablemente [9] (Alcarria, 2009).

Para registrar estos hechos contables se utiliza el método de la partida doble.

4.3.1. Método de la Partida Doble

El método de la partida doble es una técnica contable que establece que cada transacción afecta al menos dos cuentas, y que los débitos deben ser iguales a los créditos. Este principio asegura que la ecuación contable básica (Activos = Pasivos + Patrimonio Neto) se mantenga siempre equilibrada [9] (Alcarria, 2009).

Para ello se utilizan dos conceptos básicos:

Debe

El debe se refiere a todos los ingresos que recibe una empresa y representan un cargo a la cuenta [10] (¿Qué Son El Debe Y Haber En Contabilidad? Diferencias, Ejemplos Y Más, 2024).

Haber

El haber registra las salidas y entregas de una cuenta [10] (¿Qué Son El Debe Y Haber En Contabilidad? Diferencias, Ejemplos Y Más, 2024).

4.4. Clasificación Contable

Todas estas operaciones deben registrarse y declararse a la entidad fiscal de cada país. Sin embargo, hay una gran cantidad de especificaciones que se deben incluir en estas declaraciones, principalmente en la justificación de cada gasto e ingreso. Para lograr este objetivo, cada factura debe clasificarse y describirse adecuadamente.

4.4.1. Definición de Clasificación Contable

La clasificación contable es la agrupación sistemática de los cargos y abonos relacionados a una persona o situación de la misma naturaleza, que se registran bajo un encabezamiento o título que los identifica [31] (Gonzalez, 2003).

4.4.2. Cuentas Contables

En el ámbito de la contabilidad, se hace referencia a los diferentes encabezados o agrupaciones bajo el término 'cuenta contable'. Según QuickBooks, un reconocido software de contabilidad a nivel mundial, las cuentas contables constituyen la base del sistema contable de una empresa. La ausencia de estas cuentas puede representar un desafío significativo para el registro adecuado de transacciones, el análisis financiero y la elaboración de informes financieros. [29] (QuickBooks, 2024)

La tarea de definir y clasificar estas cuentas recae en el contador y la dirección de la empresa, quienes deben asegurar que la estructura adoptada se alinee con las operaciones cotidianas de la organización. Una vez establecida esta estructura, es responsabilidad del contador llevar a cabo la clasificación correspondiente. [31] (Gonzalez, 2003)

4.5. Entorno Fiscal de Guatemala

El entorno fiscal en Guatemala se caracteriza por un sistema de impuestos y normativas contables diseñadas para regular las actividades económicas y garantizar el financiamiento de los servicios públicos. A continuación, se describen los componentes principales del sistema de impuestos y declaraciones, así como las reglas de contabilidad empresarial basadas en la información proporcionada por la Superintendencia de Administración Tributaria (SAT) y otras fuentes relevantes.

4.5.1. Sistema de Impuestos y Declaraciones

El sistema fiscal guatemalteco comprende varios impuestos que afectan distintas áreas económicas. Los principales impuestos son los siguientes:

Impuesto Sobre la Renta (ISR)

- **ISR de Actividades Lucrativas**: Grava los ingresos derivados de actividades comerciales, industriales, financieras y de servicios [35] (Superintendencia de Administración Tributaria [SAT], n.d.).
- **ISR de Actividades Agropecuarias**: Aplica a los ingresos de actividades agrícolas, ganaderas, forestales y similares [35] (SAT, n.d.).
- **ISR de Rentas del Trabajo**: Afecta a los ingresos obtenidos por personas físicas derivadas de relaciones laborales [35] (SAT, n.d.).

Impuesto al Valor Agregado (IVA)

Este impuesto grava el valor añadido en cada etapa de la producción y distribución de bienes y servicios. La tasa general es del 12 %, con ciertas excepciones y exoneraciones [35] (SAT, n.d.).

Impuesto Sobre Productos Financieros (ISPF)

Aplica sobre los rendimientos financieros obtenidos por personas y entidades, tales como intereses de depósitos bancarios y ganancias de inversiones [35] (SAT, n.d.).

Impuesto de Solidaridad (ISO)

Es un tributo temporal que grava el 1 % de los ingresos brutos de las empresas, con condiciones y deducciones específicas [35] (SAT, n.d.).

4.5.2. Reglas de Contabilidad Empresarial

Las reglas contables en Guatemala aseguran que las empresas mantengan registros precisos y transparentes de sus actividades financieras. Las principales normativas incluyen:

Registros Contables

Las empresas deben llevar libros contables autorizados por la SAT, incluyendo el libro diario, libro mayor, y libro de inventarios y balances. La contabilidad debe reflejar todas las operaciones financieras de manera veraz [15] (Colegio de Contadores Públicos y Auditores de Guatemala, 2007).

Estados Financieros

Al cierre de cada ejercicio fiscal, las empresas deben elaborar estados financieros que incluyan el balance general, estado de resultados, estado de cambios en el patrimonio neto y estado de flujos de efectivo. Estos

estados deben ser auditados por contadores públicos autorizados, garantizando su conformidad con las Normas Internacionales de Información Financiera (NIIF) [15] (Colegio de Contadores Públicos y Auditores de Guatemala, 2007).

Declaraciones y Pago de Impuestos

Las empresas deben presentar declaraciones mensuales y anuales según el tipo de impuesto aplicable, utilizando los formularios y medios electrónicos establecidos por la SAT. El pago de impuestos debe realizarse en los plazos determinados por la ley, para evitar sanciones y recargos [35] (SAT, n.d.).

Documentación de Soporte

Toda transacción debe estar respaldada por documentos válidos como facturas, recibos y contratos, los cuales deben ser conservados por un periodo mínimo de cinco años para posibles auditorías fiscales [15] (Colegio de Contadores Públicos y Auditores de Guatemala, 2007).

Estas disposiciones aseguran que el entorno fiscal en Guatemala se mantenga ordenado y eficiente, promoviendo la responsabilidad y transparencia de las actividades económicas en el país. La mayor parte de los conceptos mencionados anteriormente parten de registros tributarios llamados Facturas.

4.6. Concepto de Factura

Una factura es un documento de carácter mercantil que refleja la compraventa de un bien o la prestación de un servicio determinado [36] (Alejandro Donoso Sánchez, 2017).

4.6.1. Participantes de una Factura

A partir de esto, se pueden identificar dos participantes que conforman una factura: el comprador y el vendedor. Estos dos participantes, según las regulaciones fiscales de Guatemala, están obligados a mantener un sistema contable para registrar y declarar cada impuesto asociado a los productos o servicios proporcionados. La factura es un concepto que posee cierto nivel de abstracción, ya que documenta una transacción, y su representación puede variar ampliamente en función de diversas acciones u objetos. Por esta razón, se agrupa en dos categorías principales relacionadas con la factura: servicios y bienes. [15] (Colegio de Contadores Públicos y Auditores de Guatemala, 2007).

4.6.2. Elementos de una Factura

Como cualquier documento de identificación de una transacción hay varios elementos que se deben incluir en el mismo. De acuerdo con Alejandro Donoso [36], hay siete aspectos que cualquier factura debería poseer:

- Lugar y fecha de emisión
- Numeración de factura
- Identificación de comprador y vendedor
- Descripción de la operación
- Base imponible de la operación (o contraprestación sin impuestos)

- Impuestos indirectos que gravan la operación (IVA)
- Contraprestación total

4.6.3. Desafíos en la Clasificación Contable

En este contexto, surge un desafío particularmente relevante: "La automatización de la clasificación contable en un entorno completamente digital". La pregunta clave que se plantea es cómo un sistema computarizado puede identificar y asignar de manera precisa cada elemento de una factura a la cuenta contable adecuada. Esta cuestión aborda directamente la intersección entre la tecnología y la práctica contable, destacando la importancia de desarrollar soluciones sofisticadas que puedan manejar la complejidad de la contabilidad moderna.

4.7. Automatización de la Clasificación Contable

Dado el objetivo planteado, resulta esencial establecer una comprensión clara de varios conceptos fundamentales antes de abordar la automatización del proceso de clasificación contable. Este esfuerzo se enmarca dentro del desarrollo teórico que precede la aplicación práctica del proyecto. Al integrar estos conceptos en la automatización del proceso de clasificación contable, se pretende desarrollar un sistema capaz de entender y clasificar de forma autónoma los elementos de las facturas en las cuentas contables correspondientes. Este enfoque no solo apunta a mejorar la eficiencia y precisión del proceso contable, sino también a proporcionar una base teórica sólida para la implementación práctica del proyecto.

Existen innumerables formas de automatizar los sistemas, claramente todo depende del contexto y los problemas que se pretenden solucionar. Desde automatizaciones con maquinaria especializada hasta pipelines de datos. En el caso de este proyecto, dado que se encuentra bajo un sistema principalmente digital, donde los registros se generan de forma automática y se presentan en informes textuales, la automatización puede permanecer por completo de forma digital.

En el ámbito de la clasificación automática, se han planteado todo tipo de algoritmos capaces de encontrar patrones y señales en los datos que puedan dar pauta a un comportamiento esperado que permita a una computadora generar las clasificaciones correctas. Dado lo popular que se ha hecho la Inteligencia Artificial, este proyecto utilizará ese recurso para crear una solución innovadora al reto plantedo.

4.7.1. Herramientas

Python

Python es un lenguaje de programación muy popular, utilizado en aplicaciones web, desarrollo de software, ciencia de datos y machine learning. Su popularidad se debe a su eficiencia y facilidad de aprendizaje, además de ser compatible con múltiples plataformas. Python es de código abierto, lo que permite a los desarrolladores descargarlo gratuitamente e integrarlo en diversos sistemas, acelerando así el desarrollo [1] (Amazon Web Services, 2022).

Beneficios de Python

Entre los beneficios de Python, se destaca su sintaxis sencilla y similar al inglés, lo que facilita la lectura y comprensión del código. Esto permite a los desarrolladores ser más productivos, ya que pueden escribir programas con menos líneas de código en comparación con otros lenguajes. Además, Python cuenta con

una amplia biblioteca estándar con códigos reutilizables, lo que reduce la necesidad de programar desde cero. Python también se integra fácilmente con otros lenguajes de programación como Java, C y C++. La comunidad activa de desarrolladores de Python, que se extiende por todo el mundo, ofrece soporte rápido y abundantes recursos en línea para aprender el lenguaje, como videos, tutoriales y documentación [1] (Amazon Web Services, 2022).

Librerías de Python

Según Verne Academy (2022) [7], algunas de las librerías más utilizadas en Python para Data Science y Machine Learning incluyen Pandas, Numpy, Plotly, Scikit-learn, Category-encoders, Imbalance Learning, LightGBM/XGBoost, Keras/Tensorflow, Shap y AzureML-sdk.

4.7.2. Inteligencia Artificial (IA)

La Inteligencia Artificial se refiere a sistemas o máquinas que simulan la inteligencia humana para realizar tareas y pueden mejorar sus capacidades basándose en la información que recogen. La IA se aplica en una amplia gama de campos, desde el reconocimiento de voz hasta el análisis de datos. En el contexto de la contabilidad, la IA puede ser utilizada para automatizar la clasificación de datos contables, mejorando la eficiencia y la precisión [30] (Russell & Norvig, 2016).

4.7.3. Machine Learning (ML)

Machine Learning es un subcampo de la IA que permite a las máquinas aprender de los datos y mejorar su desempeño a través de la experiencia, sin estar explícitamente programadas para cada tarea. En la clasificación contable, los algoritmos de ML pueden identificar patrones en los datos financieros y aprender a clasificarlos de manera adecuada [13] (Bishop, 2006).

Los algoritmos de machine learning se dividen en cuatro categorías principales: supervisados, no supervisados, semisupervisados y de refuerzo. Cada tipo tiene sus propias ventajas y es elegido según las necesidades de velocidad, precisión y presupuesto disponible [24] (IBM, 2024).

Algoritmos de aprendizaje supervisado

En el aprendizaje supervisado, se entrenan modelos con datos etiquetados. Existen dos tipos principales de problemas: clasificación y regresión. Los algoritmos de clasificación, como los clasificadores lineales, máquinas de vectores de apoyo (SVM), árboles de decisión, k-vecinos más cercanos (KNN) y bosques aleatorios, se utilizan para asignar datos a categorías específicas. En cambio, los algoritmos de regresión, como la regresión lineal, logística y polinómica, se emplean para modelar la relación entre variables dependientes e independientes, permitiendo realizar proyecciones como los ingresos de ventas [24] (IBM, 2024).

Algoritmos de aprendizaje no supervisado

El aprendizaje no supervisado trabaja con datos no etiquetados para descubrir patrones y agrupaciones dentro del conjunto de datos. Los métodos más comunes incluyen el clustering jerárquico y K-means, que identifican y organizan datos similares sin etiquetas previas [24] (IBM, 2024).

Algoritmos de aprendizaje semisupervisado

Esta técnica combina elementos del aprendizaje supervisado y no supervisado. Se utiliza cuando solo una parte de los datos de entrada está etiquetada, lo que permite aprovechar tanto la precisión del aprendizaje supervisado como la flexibilidad del uso de datos no etiquetados [24] (IBM, 2024).

Algoritmos de refuerzo

Los algoritmos de refuerzo aprenden mediante un sistema de recompensas y penalizaciones, similar a cómo los humanos aprenden de la experiencia. Este enfoque es común en áreas como la gestión de recursos, la robótica y los videojuegos, donde las acciones se ajustan continuamente para maximizar las recompensas a largo plazo [24] (IBM, 2024).

4.7.4. Deep Learning (DL)

El Deep Learning es una técnica avanzada de ML que utiliza redes neuronales profundas para modelar complejidades. Es especialmente útil para procesar grandes volúmenes de datos y reconocer patrones complejos, lo que lo hace idóneo para tareas como la clasificación de elementos dentro de las facturas en categorías contables específicas [26] (LeCun, Bengio, & Hinton, 2015).

4.8. Aplicaciones del Procesamiento del Lenguaje Natural en la Clasificación Contable

Dada la naturaleza de la información que se maneja, principalment textual, vale por completo la pena incluír temas de procesamiento de lenguaje natural para procurar llegar a una solución robusta.

4.8.1. Procesamiento del Lenguaje Natural (PLN)

El Procesamiento del Lenguaje Natural es una rama de la Inteligencia Artificial que se enfoca en la interacción entre computadoras y humanos a través del lenguaje natural. En la clasificación contable, el PLN puede ser utilizado para interpretar y clasificar el texto de las facturas, extrayendo información relevante y asignándola a las cuentas contables adecuadas [25] (Jurafsky & Martin, 2019).

4.8.2. Aplicaciones Específicas del PLN en la Contabilidad

Análisis de Sentimientos

Aunque tradicionalmente es asociado con la evaluación de opiniones y emociones en textos, el análisis de sentimientos puede adaptarse para evaluar el tono y la intención detrás de las entradas contables. Esto puede ser útil para identificar transacciones inusuales o categorías específicas basadas en el contexto del texto.

Clasificación de Texto

La clasificación de texto es un proceso de Machine Learning que asigna etiquetas o categorías a textos no estructurados. Este proceso es crucial para analizar y categorizar eficazmente los datos contables,

identificando patrones que pueden no ser evidentes a primera vista [8] (Aggarwal & Zhai, 2012).

Minería de Datos

La minería de datos implica descubrir patrones y correlaciones en grandes conjuntos de datos. En el contexto contable, permite analizar y categorizar los datos, encontrando relaciones útiles para la toma de decisiones.

4.8.3. Modelos Avanzados en PLN

Modelos de Atención y Transformers

Los modelos de atención, especialmente los basados en arquitecturas Transformer, han revolucionado el campo del PLN. Estos modelos son capaces de ponderar la importancia de diferentes partes de un texto para entender su contexto completo, lo que los hace extremadamente útiles para tareas complejas de clasificación y comprensión de textos en documentos contables [38] (Vaswani et al., 2017).

BERT y Modelos de Lenguaje Preentrenados

BERT (Bidirectional Encoder Representations from Transformers) y otros modelos de lenguaje preentrenados han establecido nuevos estándares en el PLN. Al ser entrenados en vastos corpus de texto, estos modelos tienen una comprensión profunda del lenguaje natural, lo que les permite realizar tareas de clasificación y comprensión de textos con un alto grado de precisión, incluso en el dominio específico de la contabilidad [16] (Devlin et al., 2019).

GPT-3 y Modelos Generativos de Gran Escala

GPT-3 (Generative Pre-trained Transformer 3) es uno de los modelos de lenguaje más avanzados desarrollado por OpenAI. Con 175 mil millones de parámetros, GPT-3 es capaz de generar texto coherente y realizar una variedad de tareas en Procesamiento de Lenguaje Natural (PLN) sin necesidad de entrenamiento adicional en tareas específicas. Su capacidad para generar texto de alta calidad, comprender y continuar contextos complejos, lo convierte en una herramienta poderosa para la automatización de tareas de comprensión y generación de texto en dominios como la contabilidad y finanzas [38] (Brown et al., 2020).

4.9. Beneficios y Desafíos de la Automatización en la Clasificación Contable

4.9.1. Beneficios

- Eficiencia: La automatización reduce significativamente el tiempo necesario para clasificar facturas, permitiendo a los contadores enfocarse en tareas más estratégicas.
- Precisión: Los algoritmos de ML y DL pueden minimizar errores humanos en la clasificación, mejorando la precisión de los registros contables.
- Cumplimiento Fiscal: Un sistema automatizado puede asegurar que todas las facturas se clasifiquen y registren de acuerdo con las normativas fiscales, reduciendo el riesgo de sanciones.

4.9.2. Desafíos

- Calidad de Datos: La efectividad de los modelos de IA depende de la calidad y cantidad de datos disponibles para el entrenamiento.
- Integración: La implementación de un sistema automatizado debe considerar la integración con los sistemas contables y de gestión existentes.
- Adaptabilidad: Los sistemas deben ser adaptables a cambios en las regulaciones fiscales y en las estructuras contables de las empresas.

4.10. Evaluación y Validación del Sistema Automatizado

Para asegurar la efectividad del sistema automatizado de clasificación contable, se deben considerar las siguientes métricas de evaluación:

- Precisión y Recall: Medir la exactitud con la que el sistema clasifica las facturas en las cuentas contables correctas.
- Tiempo de Procesamiento: Evaluar el tiempo requerido para procesar y clasificar facturas en comparación con métodos manuales.
- Feedback de Usuarios: Recopilar opiniones y sugerencias de los contadores y usuarios del sistema para identificar áreas de mejora.
- Pruebas Piloto: Implementar el sistema en un entorno controlado antes de su despliegue completo para identificar posibles problemas y ajustar el sistema en consecuencia.

4.11. Infraestructura

Para asegurar la escalabilidad y robustez necesarias para recibir consultas constantes y mantenerse activo, el proyecto adoptará un enfoque serverless, utilizará una base de datos no relacional y una aplicación web como interfaz de usuario con React. A continuación, se describen las principales herramientas empleadas en el proyecto, cada una seleccionada por su capacidad para soportar cargas dinámicas y ofrecer alta disponibilidad. Estas herramientas permitirán un desarrollo ágil y una infraestructura eficiente, adecuada para los requerimientos del proyecto.

4.11.1. Infraestructura Serverless

A pesar de su nombre, la computación "serverless"no implica la ausencia de servidores. En cambio, se enfatiza la desvinculación del desarrollador de la gestión de servidores, permitiéndoles enfocarse exclusivamente en la funcionalidad de la aplicación. Este enfoque se basa en que los proveedores de servicios en la nube asignen y ejecuten dinámicamente el código de backend, facturando a los usuarios según los procesos de computación reales en lugar de la capacidad de servidor predeterminada [33] (Shafiei, Khonsari y Mousavi, 2022).

Toda esta infraestructura debe ser implementada sobre un sistema de herramientas adecuado y eficiente, capaz de escalar para soportar las necesidades y cumplir con los objetivos. Para eso se definió trabajar con el entorno Google.

Google Cloud

Google Cloud es una suite de servicios en la nube que ofrece una amplia gama de herramientas y recursos para empresas y desarrolladores. Estos servicios se pueden utilizar para crear y administrar aplicaciones, almacenar datos, analizar información y mucho más. Google Cloud es una de las plataformas de nube más populares del mercado, junto con AWS y Azure [5] (Google Cloud, 2024).

Características principales de Google Cloud

Google Cloud ofrece una amplia gama de servicios, que se pueden agrupar en las siguientes categorías:

- Infraestructura como servicio (IaaS): Proporciona recursos informáticos básicos, como máquinas virtuales, almacenamiento y redes.
- Plataforma como servicio (PaaS): Ofrece herramientas para desarrollar y desplegar aplicaciones en la nube.
- Software como servicio (SaaS): Incluye una amplia gama de aplicaciones empresariales, como Google Workspace y Google Maps.
- Inteligencia artificial y aprendizaje automático: Ofrece servicios para desarrollar e implementar aplicaciones de IA y ML.
- Herramientas de análisis de datos: Permite recopilar, procesar y analizar grandes conjuntos de datos.

Cloud Functions

Las Cloud Functions son una solución de computación sin servidor que permite a los desarrolladores ejecutar código en respuesta a eventos sin necesidad de gestionar servidores. Estas funciones se ejecutan en entornos gestionados y escalables, donde se pueden desencadenar mediante eventos HTTP, cambios en bases de datos, mensajes en colas, entre otros. [4] (Cloud Functions | Google Cloud, 2024)

Las Cloud Functions se escriben generalmente en lenguajes de programación populares como JavaScript, Python, y Go. Su principal ventaja es la capacidad de escalar automáticamente en función de la demanda y de integrarse fácilmente con otros servicios en la nube. [4] (Cloud Functions | Google Cloud, 2024)

Dada la escalabilidad de esta herramienta y el bajo costo que representa dado que es bajo demanda, fue seleccionada para representar el método de comunicación entre todo el sistema y la interfaz gráfica.

App Engine

Google App Engine es una plataforma de desarrollo sin servidor que permite a los desarrolladores crear y desplegar aplicaciones web y móviles en la infraestructura de Google. Proporciona servicios gestionados como bases de datos, almacenamiento y autenticación, lo que permite a los desarrolladores centrarse en escribir código sin preocuparse por la gestión de servidores. App Engine soporta varios lenguajes de programación como Python, Java, Go y PHP, y ofrece escalabilidad automática para manejar el tráfico de las aplicaciones. [6] (Documentación de App Engine | App Engine Documentation | Google Cloud, 2024)

4.11.2. Base de datos

Una base de datos es un conjunto organizado de información estructurada, almacenada y accesible electrónicamente. Las bases de datos son fundamentales en la gestión de información en sistemas informáticos

modernos, permitiendo el almacenamiento, recuperación y manipulación eficiente de datos [17] (Elmasri & Navathe, 2016). Las bases de datos se clasifican principalmente en dos tipos: relacionales y no relacionales. Las bases de datos relacionales han sido el estándar durante décadas, organizando datos en tablas con relaciones predefinidas. Sin embargo, con el aumento de datos no estructurados y la necesidad de escalabilidad, las bases de datos no relacionales han ganado popularidad [32] (Sadalage & Fowler, 2012).

Base de Datos no Relacionales

Las bases de datos no relacionales, también conocidas como NoSQL (Not Only SQL), son sistemas de gestión de datos que proporcionan un mecanismo flexible para almacenar y recuperar información. A diferencia de las bases de datos relacionales, no utilizan tablas con esquemas fijos y relaciones predefinidas [3] (MongoDB, 2023).

Características principales:

- Esquema flexible: Permiten almacenar datos sin una estructura rígida predefinida.
- Escalabilidad horizontal: Facilitan la distribución de datos en múltiples servidores.
- Alto rendimiento: Optimizadas para operaciones de lectura y escritura rápidas.
- Variedad de modelos de datos: Incluyen documentos, clave-valor, columnas anchas y grafos.

Las bases de datos no relacionales son especialmente útiles en aplicaciones web, móviles y de big data, donde la velocidad y la flexibilidad son cruciales [37] (Vaish, 2013).

Firestore

Firestore es una base de datos NoSQL flexible y escalable desarrollada por Google como parte de la plataforma Firebase. Diseñada para aplicaciones web y móviles, Firestore ofrece sincronización en tiempo real y soporte sin conexión, lo que la hace ideal para desarrollos modernos [2] (Firebase, 2023).

Características principales:

- Modelo de datos: Utiliza un modelo de datos jerárquico basado en documentos y colecciones.
- Consultas expresivas: Permite consultas complejas y eficientes sin necesidad de índices personalizados para cada consulta.
- Escalabilidad automática: Se ajusta automáticamente a la demanda de tráfico sin necesidad de configuración manual.
- Seguridad robusta: Ofrece reglas de seguridad flexibles a nivel de documento.

Firestore es particularmente útil en escenarios donde se requiere una estructura de datos flexible y actualizaciones en tiempo real, como aplicaciones colaborativas, juegos en línea o sistemas de mensajería [27] (Moroney, 2017).

Almacenamiento General

El almacenamiento general se refiere a los sistemas y tecnologías utilizados para guardar y gestionar datos digitales de manera persistente. En el contexto de la informática moderna, el almacenamiento general abarca una amplia gama de soluciones que van desde dispositivos físicos hasta servicios en la nube [22] (Hennessy & Patterson, 2019).

Características principales del almacenamiento general:

- Persistencia: Los datos se mantienen incluso cuando no hay alimentación eléctrica.
- Escalabilidad: Capacidad de aumentar el espacio de almacenamiento según las necesidades.
- Accesibilidad: Permite la recuperación y manipulación de datos almacenados.
- Seguridad: Implementa medidas para proteger los datos contra accesos no autorizados o pérdidas.

Cloud Storage

Cloud Storage es un modelo de almacenamiento de datos en línea donde la información se mantiene en servidores remotos gestionados por un proveedor de servicios en la nube. Este enfoque ofrece una solución escalable, accesible y económica para el almacenamiento de datos en la era digital [18] (Erl et al., 2013).

4.11.3. Seguridad

La seguridad en el ámbito de la informática se refiere a la protección de sistemas, redes y datos contra accesos no autorizados, ataques y daños. En un mundo cada vez más digitalizado, la seguridad de la información se ha vuelto crucial para individuos, empresas y organizaciones [34] (Stallings & Brown, 2018).

Elementos clave de la seguridad informática:

- Confidencialidad: Garantizar que la información sea accesible solo a personas autorizadas.
- Integridad: Asegurar que los datos no sean alterados de manera no autorizada.
- Disponibilidad: Garantizar el acceso a la información cuando se necesite.
- Autenticación: Verificar la identidad de los usuarios y sistemas.

Encriptación

La encriptación es un proceso fundamental en la seguridad de la información que convierte datos legibles en un formato codificado para protegerlos de accesos no autorizados. Es esencial para salvaguardar la confidencialidad de la información sensible [28] (Paar & Pelzl, 2010).

La encriptación se aplica en diversas áreas, como comunicaciones seguras, almacenamiento de contraseñas y protección de datos en reposo.

Secret Manager

Secret Manager es un servicio de gestión de secretos que permite almacenar, administrar y acceder de forma segura a información sensible como claves API, contraseñas y certificados. Estos servicios son cruciales para mantener la seguridad en aplicaciones y sistemas modernos [21] (Secret Manager | Google Cloud, 2024).

Ejemplos de Secret Managers incluyen AWS Secrets Manager, Google Cloud Secret Manager y HashiCorp Vault.

4.12. Repositorio

Un repositorio es un espacio centralizado donde se almacena, organiza, mantiene y difunde información digital, típicamente código fuente de programas informáticos. Los repositorios son fundamentales en el desarrollo de software moderno, facilitando la colaboración y el control de versiones [14] (Chacon & Straub, 2014).

4.12.1. Github

GitHub es una plataforma de desarrollo colaborativo que aloja repositorios de código utilizando el sistema de control de versiones Git. Fundada en 2008 y adquirida por Microsoft en 2018, GitHub se ha convertido en el hogar de millones de proyectos de código abierto y privados [12] (Bell & Beer, 2018). Características principales de GitHub:

Github Actions

GitHub Actions es un servicio de integración y entrega continua (CI/CD) integrado en GitHub. Permite automatizar flujos de trabajo de desarrollo de software directamente desde los repositorios de GitHub [20] (Acerca de La Integración Continua Con Acciones de GitHub - Documentación de GitHub, 2024).

Características clave de GitHub Actions:

- Automatización: Permite crear flujos de trabajo personalizados para build, test y deploy.
- Multiplataforma: Soporta diferentes sistemas operativos y lenguajes de programación.
- Matriz de trabajos: Facilita la ejecución de pruebas en múltiples configuraciones.
- Marketplace: Acceso a acciones predefinidas creadas por la comunidad.
- Integración nativa: Se integra perfectamente con otros servicios de GitHub.

GitHub Actions ha simplificado significativamente los procesos de CI/CD, permitiendo a los desarrolladores automatizar sus flujos de trabajo directamente desde sus repositorios.

CAPÍTULO 5

Metodología

5.1. Antecedentes

La contabilidad, pilar fundamental en la gestión empresarial, ha experimentado una notable evolución a lo largo del tiempo. Históricamente, los contadores realizaban sus labores de forma manual, manteniendo registros físicos en libros contables y efectuando cálculos sin asistencia tecnológica. Este método tradicional, aunque meticuloso, era propenso a errores y demandaba una cantidad considerable de tiempo y recursos.

En las últimas décadas, la introducción de herramientas tecnológicas ha transformado significativamente la práctica contable. La aparición de hojas de cálculo electrónicas como Microsoft Excel marcó el inicio de esta revolución. Posteriormente, el desarrollo de sistemas de Planificación de Recursos Empresariales (ERP) como SAP y ODOO ha permitido una gestión integrada de la contabilidad con otras áreas de la empresa, optimizando procesos y mejorando la eficiencia operativa. Estas soluciones han automatizado gran parte del proceso contable, permitiendo a los profesionales manejar volúmenes más grandes de información con mayor precisión y eficiencia.

Sin embargo, a pesar de estos avances, muchas tareas aún requieren intervención manual, como el registro y clasificación de documentos contables, en particular las facturas. Esta situación ha llevado a la búsqueda de soluciones más avanzadas que permitan automatizar aún más estos procesos.

Un avance reciente y significativo en Guatemala ha sido la implementación obligatoria de documentos tributarios electrónicos (DTE) por parte de la SAT. Esta medida ha agilizado numerosos procesos de registro y ha sentado las bases para una mayor automatización en el futuro.

En este contexto de evolución tecnológica y regulatoria, surge un proyecto desarrollado por el estudiante Juan Carlos Baján sobre el cual se cimenta esta tesis, que busca automatizar el proceso contable en su totalidad. Este proyecto ha logrado optimizar eficientemente una parte crucial de la contabilidad: el procesamiento de facturas. El sistema actual es capaz de procesar miles de facturas en cuestión de minutos, extrayendo toda la información relevante y almacenándola en una base de datos. Cabe mencionar que este proyecto mencionado no forma parte de la tesis, es fundamental para que se lleve a cabo pues es a trevés de este sistema que se recauda la información que alimentará la tesis, pero no pertenece a la tesis.

La plataforma utiliza Google Cloud Platform como infraestructura, permitiendo un acceso eficiente a la información a través de cloud functions. La interfaz de usuario facilita a los contadores la clasificación de las facturas, que posteriormente se conectan con el ERP de la empresa para su almacenamiento completo. Este sistema ha sido el resultado de meses de planificación, desarrollo e implementación, y actualmente está siendo utilizado por un bufete de contadores, quienes proporcionan retroalimentación continua para su mejora.

El desarrollo de sistemas de este tipo es fundamental para generar progreso en la sociedad. Permite a las personas dejar de enfocarse en actividades repetitivas y poco contribuyentes, como el registro de numerosos hechos contables, liberando tiempo y recursos para tareas de mayor valor añadido. Esto no solo mejora la eficiencia empresarial, sino que también contribuye al desarrollo profesional y personal de los contadores.

Además, los recientes avances en inteligencia artificial ofrecen nuevas posibilidades para el análisis de los hechos contables. Las capacidades de análisis de texto permiten una comprensión más profunda y compleja de la información contenida en los documentos contables, abriendo nuevas vías para la automatización y la toma de decisiones basada en datos.

Un aspecto crucial en el desarrollo e implementación de estos sistemas es la seguridad de la información. Es fundamental contar con la autorización de todos los involucrados y garantizar que la información se mantenga segura. Además, se deben implementar medidas para que, en caso de una filtración de datos, los entes de dicha información sean irrastreables y no identificables, protegiendo así la privacidad y confidencialidad de los individuos y las empresas.

La presente tesis se enfoca en desarrollar un sistema de clasificación automática de facturas, aprovechando las capacidades de la inteligencia artificial. Este sistema busca complementar y mejorar la plataforma existente, con el objetivo de disminuir significativamente la carga de trabajo manual de los contadores.

Aunque este avance representa solo una parte del proceso contable completo, constituye un paso importante hacia la automatización integral de la contabilidad. El éxito de este proyecto no solo beneficiará a los contadores al reducir sus tareas rutinarias, sino que también tiene el potencial de mejorar la precisión, eficiencia y velocidad de los procesos contables en general. Esto, a su vez, puede traducirse en beneficios significativos para las empresas en términos de gestión financiera, cumplimiento regulatorio y capacidad para centrarse en actividades estratégicas de mayor valor.

5.2. Descripción y Alcance

Esta tesis se centra en el desarrollo de un sistema basado en inteligencia artificial para automatizar la clasificación de facturas, un proceso que actualmente se realiza de manera manual. La finalidad de esta solución es asignar de forma automática los productos de una factura a las cuentas contables correspondientes, mejorando así la precisión y eficiencia del registro contable. Utilizando técnicas de inteligencia artificial, se busca reducir significativamente la carga de trabajo relacionada con la clasificación manual, liberando a los contadores de tareas repetitivas y susceptibles a errores. El sistema también se diseña con la capacidad de reentrenarse de manera sencilla, permitiendo una adaptación continua a cambios en la clasificación de cuentas. Además, se desplegará en la nube mediante Google Cloud Functions, garantizando un preprocesamiento de datos automatizado y una integración fluida dentro de la infraestructura contable.

El alcance de este trabajo no se limita simplemente a la automatización de un proceso, sino que aspira a ser un catalizador para la transformación de la profesión contable. La plataforma pretende ser un punto de apoyo desde el cual los contadores puedan impulsar y elevar la calidad de sus servicios. Al liberar a los profesionales de las tareas repetitivas y propensas a errores asociadas con la clasificación manual de facturas, se busca crear un espacio para que los contadores se enfoquen en actividades de mayor valor añadido.

Esta tesis tiene como objetivo facilitar un cambio de paradigma en la práctica contable, permitiendo a los profesionales dedicar más tiempo y recursos al análisis de datos financieros, la interpretación de tendencias económicas y la asesoría estratégica a sus clientes. La plataforma no busca reemplazar el juicio y la experiencia

de los contadores, sino potenciarlos, proporcionándoles herramientas que amplíen sus capacidades analíticas.

Además, el alcance de la tesis incluye el fomento de la estandarización de los procesos contables y los esquemas de cuentas. Si bien la plataforma no impondrá un sistema único, se espera que, a través de su uso, se genere una convergencia natural hacia mejores prácticas en la organización y clasificación de la información financiera.

Es importante señalar que el alcance de esta tesis no se extiende a la totalidad del proceso contable. No pretende automatizar la generación de estados financieros, la declaración de impuestos o la toma de decisiones estratégicas. Más bien, se centra en optimizar un aspecto fundamental: la clasificación de facturas, como punto de partida para una mejora integral de los servicios contables.

En última instancia, el alcance de este tesis trasciende la mera implementación tecnológica. Busca ser un instrumento de cambio que permita a los contadores ofrecer un servicio de mayor calidad, más preciso y con un valor agregado significativo para sus clientes. La visión es que esta plataforma contribuya a elevar el estándar de la profesión contable, permitiendo a los profesionales del área dedicar más tiempo a actividades que realmente aprovechan su experiencia y juicio experto.

5.3. Recolección de Datos

El proceso de recolección de datos para este estudio se fundamenta en la información generada a través de la plataforma ya anteriormente mencionada, dicha plataforma representa el punto de partida de esta tesis pero el tema central esta tesis es la clasificación contable automática para cada item dentro de la factura. Esta metodología de obtención de datos presenta una ventaja significativa: los datos se generan de forma orgánica a través del uso cotidiano del sistema por parte de los contadores.

Es importante destacar que los contadores únicamente cargan las facturas en formato XML a la plataforma. Este formato estandarizado, proporcionado por la Superintendencia de Administración Tributaria (SAT), contiene toda la información fiscal relevante. La plataforma, a través de una infraestructura en la nube, procesa estos archivos XML de manera segura, encriptando los datos sensibles y extrayendo la información de forma estandarizada.

El sistema está diseñado para que el contador solo necesite proporcionar cuatro elementos adicionales:

1. La descripción final de cada producto
2. La clasificación contable de cada producto
3. Una descripción general de la factura
4. El tipo de gasto general de la factura

Esta metodología de recolección de datos minimiza la entrada manual de información, reduciendo así la probabilidad de errores y aumentando significativamente la eficiencia del proceso.

La disposición de los usuarios para utilizar esta plataforma, y por ende generar los datos necesarios para nuestro estudio, se fundamenta en los beneficios operativos que obtienen. El sistema actual reduce drásticamente el tiempo requerido para el registro y clasificación de facturas, lo que representa un incentivo sustancial para su adopción.

Para contextualizar la magnitud de esta mejora en eficiencia, se han analizado dos escenarios representativos del proceso tradicional de registro de facturas:

5.3.1. Escenario 1: Procesamiento de Facturas Físicas

En este caso, un contador procesa una factura física relacionada con el consumo de energía eléctrica. El tiempo estimado para analizar la factura, extraer la información relevante, transcribirla al software contable y realizar las clasificaciones necesarias oscila entre 3 y 5 minutos por factura. Considerando el escenario más optimista, el procesamiento de 150 facturas requeriría aproximadamente 7.5 horas de trabajo ininterrumpido.

5.3.2. Escenario 2: Procesamiento de Facturas Electrónicas

En este escenario, el contador procesa una factura electrónica del mismo tipo. El tiempo estimado para completar el proceso, incluyendo la extracción de datos y su ingreso al software contable, varía entre 2 y 4 minutos por factura. En el caso más favorable, el procesamiento de 150 facturas consumiría alrededor de 5 horas.

En contraste, la plataforma desarrollada permite a los usuarios procesar cada factura en aproximadamente 25 segundos. Esto implica que la tarea de clasificar 150 facturas puede completarse en 62.5 minutos, lo que representa una reducción del 80 % en el tiempo dedicado a esta actividad.

Esta mejora sustancial en la eficiencia operativa constituye el principal incentivo para que los contadores adopten la plataforma, contribuyendo así a la generación de los datos necesarios para el desarrollo de modelos de clasificación automática.

Los datos recolectados a través de la plataforma para cada factura incluyen toda la información contenida en el archivo XML proporcionado por la SAT, además de los cuatro elementos adicionales ingresados por el contador. Esta rica variedad de datos proporciona una base sólida para el desarrollo de modelos predictivos de clasificación contable.

El siguiente paso en el proceso de investigación implica un análisis exploratorio detallado de estos datos. En esta fase, se identificarán y examinarán en profundidad las variables específicas extraídas de los archivos XML, así como las proporcionadas por los contadores. Este análisis permitirá identificar patrones, relaciones y características relevantes que puedan informar el desarrollo de modelos de clasificación automática eficaces.

5.4. Análisis Exploratorio

5.4.1. Información Extraída del XML

Tras la carga del archivo XML en la plataforma, el sistema extrae y almacena una serie de datos cruciales para el proceso contable. Esta información se puede categorizar de la siguiente manera:

1. **Datos Generales de la Factura:**
   - Tipo de documento
   - Código de moneda
   - Fecha y hora de emisión
   - Número
   - Serie
2. **Información del Emisor:**
   - NIT del emisor

- Nombre del emisor
- Nombre comercial
- Código del establecimiento
- Dirección completa (incluyendo municipio, departamento y país)

3. **Información del Receptor:**
   - ID del receptor
   - Nombre del receptor
   - Dirección completa
4. **Detalles de los Ítems:** Para cada ítem en la factura, se extrae:
   - Tipo
   - Cantidad
   - Unidad de medida
   - Descripción del producto o servicio
   - Precio unitario
   - Precio total
   - Descuento
   - Información de impuestos (tipo, monto gravable, monto del impuesto)
5. **Totales:**
   - Total de impuestos
   - Gran total de la factura
6. **Información de Certificación:**
   - NIT del certificador
   - Nombre del certificador
   - Número y serie de autorización
   - Fecha y hora de certificación

Esta estructura de datos proporciona una base sólida para el análisis contable y fiscal, permitiendo una categorización precisa y un procesamiento eficiente de la información de cada factura.

5.5. Generalidades del Dataset Utilizado

El dataset recolectado consta de 12,000 facturas de compra debidamente clasificadas. Contiene las variables previamente mencionadas, así como 4 nuevas variables que indican los rubros que los contadores deben rellenar. Este dataset fue proporcionado por la empresa de contabilidad, protegiendo la identidad de los participantes y para usos únicamente académicos. Con esta información se contaba con los elementos necesarios para comenzar a tomar decisiones.

5.6. Selección de Variables

Mucha de la información dentro del dataset no influye ni tiene relación con la clasificación contable, como por ejemplo, la información del certificador. Dado que la clasificación se da sobre cada ítem dentro de la factura, los montos realmente influyentes son los propios de cada ítem y no tanto los generales de la factura. Por otro lado, de acuerdo a la documentación provista por el SAT para emitir facturas, la dirección y el código postal son campos abiertos al usuario, por lo que pueden ser muy generales y no aportar información descriptiva. Por lo tanto, las variables que se tomaron en cuenta a partir de este punto fueron:

1. **Datos Generales de la Factura:**
   - Tipo de documento
   - Código de moneda
   - Fecha y hora de emisión
   - Número
   - Serie
2. **Información del Emisor:**
   - NIT del emisor
   - Nombre del emisor
   - Nombre comercial
   - Código del establecimiento
3. **Información del Receptor:**
   - ID del receptor
   - Nombre del receptor
4. **Detalles de los Ítems:** Para cada ítem en la factura:
   - Tipo
   - Cantidad
   - Unidad de medida
   - Descripción del producto o servicio
   - Precio unitario
   - Precio total
   - Descuento
   - Información de impuestos (tipo, monto gravable, monto del impuesto)

5.6.1. Análisis Estadístico

El análisis estadístico realizado proporcionó una visión más completa y detallada de la estructura del conjunto de datos (dataset). A continuación, se presenta un resumen estadístico exhaustivo de las variables categóricas, seguido de un análisis en profundidad de cada una de ellas.

Este resumen estadístico revela aspectos importantes sobre la distribución y diversidad de cada variable categórica en el dataset:

Tabla 5.1: Resumen estadístico de las variables categóricas

	Tipo de Documento	Moneda	Emisor	Establecimiento	Receptor
count	12000	12000	12000	12000	12000
unique	8	2	1959	2655	63
freq	10240	11849	514	463	1600

1. **Tipo de Documento**: Se identificaron 8 tipos únicos de documentos, con una frecuencia máxima de 10,240 para un tipo específico.
2. **Moneda**: Se utilizaron 2 tipos de moneda en las transacciones registradas.
3. **Emisor**: Se registraron 1,959 emisores únicos, lo que indica una gran diversidad en el origen de los documentos.
4. **Establecimiento**: Con 2,655 establecimientos únicos, se observa una variedad aún mayor que en los emisores.
5. **Receptor**: Se identificaron 63 receptores únicos, sugiriendo una concentración relativamente baja en comparación con otras categorías.

Para profundizar en la comprensión de estas variables, se realizó un análisis gráfico de las distribuciones más relevantes.

Análisis del Tipo de Documento

El dataset recolectado contiene múltiples tipos de documentos. Para visualizar mejor esta distribución, se generó un gráfico de barras de frecuencia.

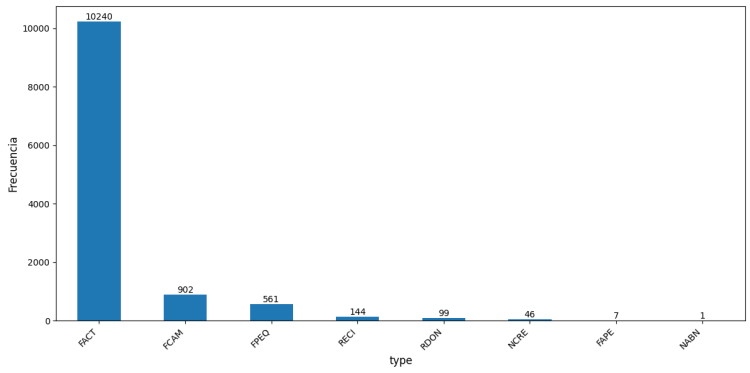

Figura 5.1: Frecuencia del Tipo de Factura

La Figura 5.1 revela una disparidad significativa en las frecuencias de cada tipo de factura. Las facturas etiquetadas como "FACT" predominan notablemente, representando 10,240 del total de 12,000 documentos (85.33 %). Esta distribución altamente sesgada podría tener implicaciones importantes para el análisis subsiguiente:

1. **Representatividad**: Los otros tipos de documentos están subrepresentados, lo que podría llevar a conclusiones sesgadas si se consideran todos los tipos en el análisis.
2. **Impacto en el modelado**: La inclusión de tipos de documentos poco frecuentes podría introducir ruido en los modelos predictivos, afectando potencialmente su rendimiento y generalización.

3. **Enfoque del estudio**: Dado que las facturas "FACT", "FCAM" y "FPEQ" constituyen la gran mayoría de los datos, centrar el análisis en este tipo de documento permitiría obtener resultados más robustos y representativos.

Considerando estos factores, se tomó la decisión metodológica de reducir el dataset a solo las facturas, en este grupo entran las Facturas (FACT), Facturas Cambiarias (FCAM) y Facturas de Pequeño Contribuyente (FPEQ), resultando en un conjunto de datos más homogéneo de 11,703 documentos. Esta decisión busca mejorar la consistencia interna del análisis y mitigar posibles sesgos introducidos por tipos de documentos poco frecuentes.

Análisis de la Moneda

Tras la reducción del dataset, se procedió a examinar la distribución de las monedas utilizadas en las transacciones.

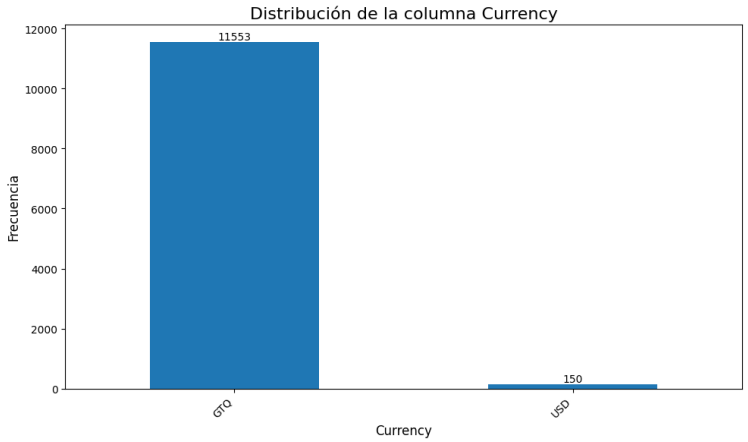

Figura 5.2: Frecuencia de Monedas

La Figura 5.2 muestra una clara predominancia del quetzal como moneda de transacción. Específicamente:

- El 98.7 % de las facturas están denominadas en quetzales (11,553 de 11,703).
- Solo el 1.3 % de las facturas utilizan dólares como moneda (150 de 11,703).

Esta distribución altamente asimétrica presenta tanto oportunidades como desafíos para el análisis:

1. **Simplificación del análisis**: Concentrarse en las transacciones en quetzales permitiría una interpretación más directa de los resultados financieros, sin la necesidad de considerar tasas de cambio o fluctuaciones monetarias.
2. **Representatividad regional**: La predominancia del quetzal refleja la naturaleza localizada de las transacciones.
3. **Decisión metodológica**: Dada la baja frecuencia de transacciones en dólares, se optó por excluirlas del análisis para mantener la coherencia y evitar la introducción de variabilidad innecesaria en los modelos posteriores.

Basándose en este análisis, se procedió a descartar las facturas en dólares, refinando aún más el conjunto de datos para los análisis subsiguientes.

Análisis de Fecha

Una vez sintetizada mejor la información, se procedió a realizar un análisis detallado de la distribución temporal de las facturas según su fecha de emisión. Este análisis reveló patrones interesantes y aspectos importantes del proceso de recolección de datos que no se habían considerado inicialmente.

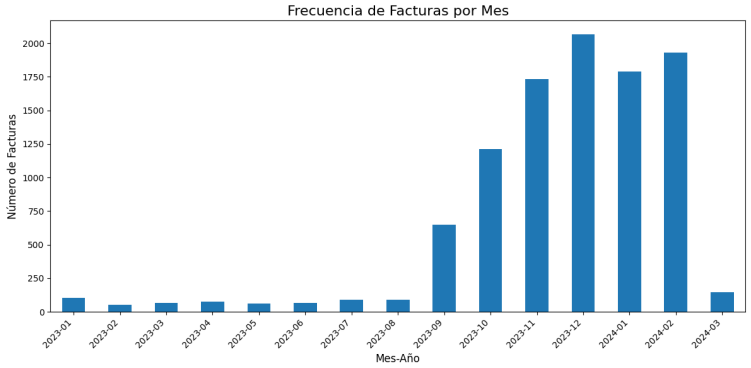

Figura 5.3: Frecuencia de Facturas Según su Fecha de Emisión

El histograma presentado en la Figura 5.3 muestra una distribución temporal de las facturas que merece un análisis más profundo. Se pueden observar las siguientes características notables:

- **Baja frecuencia inicial**: De enero a agosto de 2023, la frecuencia de facturas es notablemente baja.
- **Aumento significativo**: A partir de septiembre de 2023, se observa un incremento sustancial en el número de facturas registradas.
- **Variabilidad mensual**: Existen fluctuaciones considerables en la cantidad de facturas emitidas mes a mes, especialmente en los períodos más recientes.

Esta distribución inusual requería una explicación más detallada, la cual se obtuvo al examinar la relación entre las fechas de emisión y los receptores de las facturas.

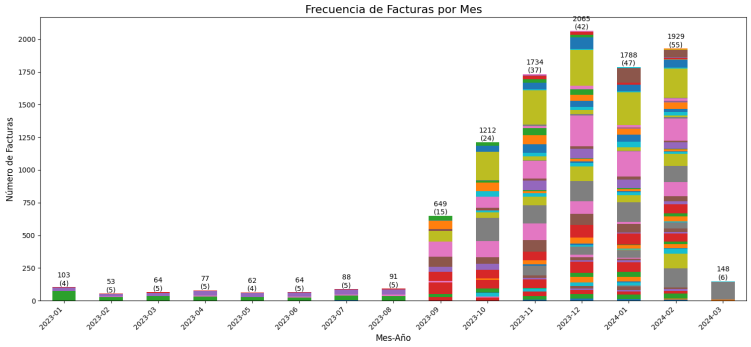

Figura 5.4: Frecuencia de Facturas según su Fecha de Emisión Agrupadas por Receptor

La Figura 5.4 proporciona una visión más clara de la distribución temporal, donde cada color representa un receptor distinto. Este gráfico revela información crucial sobre el proceso de recolección y actualización de datos:

1. **Período de baja frecuencia (enero 2023 - agosto 2023)**:

- Se observan entre 4 y 5 empresas receptoras activas.
- Este período refleja una actualización retrospectiva de la contabilidad para un grupo específico de empresas.

2. **Período de actualización (septiembre 2023 - febrero 2024)**:
   - Se evidencia un aumento significativo en el número de facturas y receptores.
   - Este incremento coincide con la actualización de la contabilidad para las empresas del período anterior.
3. **Implementación de la plataforma de recolección (desde septiembre 2023)**:
   - La plataforma de recolección de datos comenzó a funcionar, permitiendo a los contadores utilizarla para las contabilidades mensuales.
   - Esto explica el aumento sostenido en el número y diversidad de facturas registradas.
4. **Punto de finalización de datos (marzo 2024)**:
   - Marca el mes en que se realizó la descarga de los datos para este análisis.
   - Explica por qué la serie temporal termina en esta fecha.

Este análisis temporal reveló la importancia de considerar los procesos administrativos y tecnológicos subyacentes en la generación y recolección de datos contables. La implementación de la nueva plataforma de recolección en septiembre de 2023 marcó un punto de inflexión significativo, aumentando tanto la cantidad como la diversidad de los datos disponibles para el análisis.

A pesar de las variaciones significativas en las fechas, se consideró que reducir la cantidad de datos de acuerdo con su frecuencia temporal podría ser un error, por lo que se decidió permanecer con la misma cantidad de registros

Análisis de Emisores y Receptores

El análisis de las entidades involucradas en las transacciones, específicamente los emisores y receptores de las facturas, es fundamental para comprender la estructura y dinámica del conjunto de datos. En el contexto de esta investigación, que se centra primordialmente en las compras debido a las limitaciones de la plataforma de recolección de datos, es crucial distinguir entre estos dos roles:

- **Receptor**: Entidad cuya contabilidad se está registrando, representando al comprador en la transacción.
- **Emisor**: Contraparte del evento, representando al vendedor o proveedor del bien o servicio.

Es importante notar que las empresas típicamente se especializan en una gama limitada de productos o servicios. Por ejemplo, las empresas zapateras generalmente se limitan a la venta de calzado y productos relacionados, las empresas de plástico se concentran en productos plásticos, y los restaurantes se especializan en alimentos. Esta especialización sugiere la posibilidad de una correlación significativa entre el emisor de la factura y la cuenta contable asignada, una hipótesis que será examinada en detalle más adelante en este estudio.

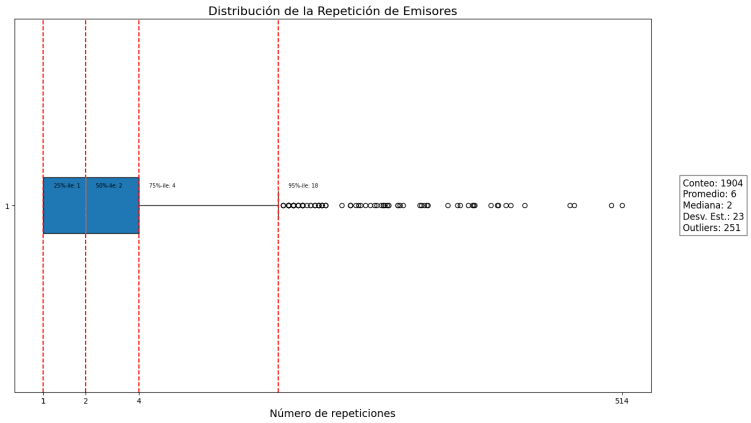

Figura 5.5: Distribución de la repetición de Emisores

Análisis de Emisores Para evaluar la distribución y frecuencia de los emisores en el conjunto de datos, se realizó un análisis estadístico visual utilizando un diagrama de caja (boxplot).

La Figura 5.5 revela una dispersión considerable en la frecuencia de aparición de los emisores. Las observaciones clave son:

- La mayoría de los emisores aparecen con una frecuencia baja, entre 1 y 4 ocurrencias aproximadamente.
- Existe una cola larga en la distribución, indicando que algunos emisores aparecen con una frecuencia significativamente mayor.
- La mediana de repeticiones es baja, lo que sugiere que más del 50 % de los emisores tienen muy pocas apariciones en el dataset.

Esta distribución altamente sesgada presenta un desafío significativo para el desarrollo de modelos predictivos robustos. La baja representatividad de la mayoría de los emisores podría llevar a:

1. Dificultades en la generalización del modelo para emisores poco frecuentes.
2. Posible sobreajuste (overfitting) hacia los emisores más frecuentes.
3. Incertidumbre en la predicción para transacciones con emisores nuevos o poco comunes.

La solución óptima para mitigar este problema sería aumentar sustancialmente el volumen de datos, especialmente para los emisores menos frecuentes. Sin embargo, dadas las limitaciones actuales del conjunto de datos, esta queda como una recomendación prioritaria para futuros esfuerzos de recolección de datos.

Para enriquecer este análisis, es esencial considerar otra variable en la factura: el establecimiento. Este representa una sub-sección de la contraparte involucrada. En términos prácticos, si un cliente realiza una compra en una sección, podría también hacerlo en otra. Esta información podría ser valiosa para la clasificación posterior, ya que podría indicar grupos de productos específicos. Por ejemplo, una empresa de soluciones financieras podría utilizar el establecimiento 1 para servicios de administración y el establecimiento 2 para asesorías, lo que aportaría una visión más detallada de las transacciones realizadas. Adicionalmente, el establecimiento podría ofrecer información geográfica, dado que algunas empresas podrían asignar un establecimiento a cada sucursal, permitiendo identificar en qué área del país se ubica cada una. Aunque no es una regla estricta, esta variable podría proporcionar datos adicionales relevantes. Sin embargo, también es posible que su inclusión incremente el sesgo y disminuya la representatividad para la mayoría de los emisores,

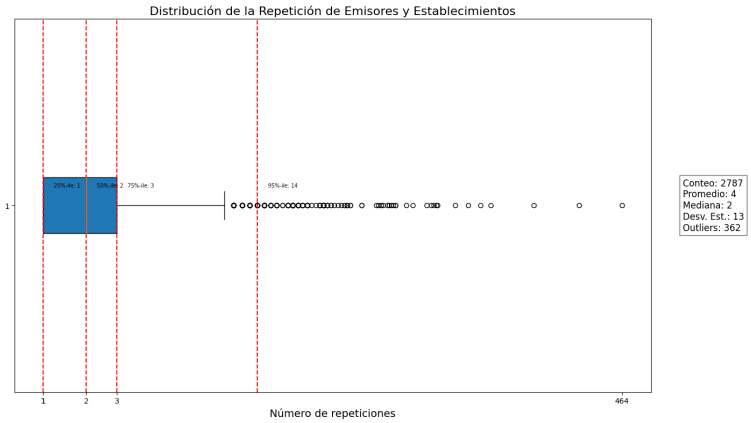

Figura 5.6: Distribución de la Repetición de Emisores y Establecimientos

como se observa en la figura 5.6. Los datos muestran un aumento en la dispersión, así como en los valores únicos y atípicos, mientras que tanto el promedio como la mediana disminuyeron, lo que refleja una mayor variabilidad en los datos. Esto podría plantear serios desafíos para futuros análisis.

Análisis de Receptores En contraste con los emisores, el análisis de los receptores revela una dinámica diferente, como se evidencia en la Figura 5.7.

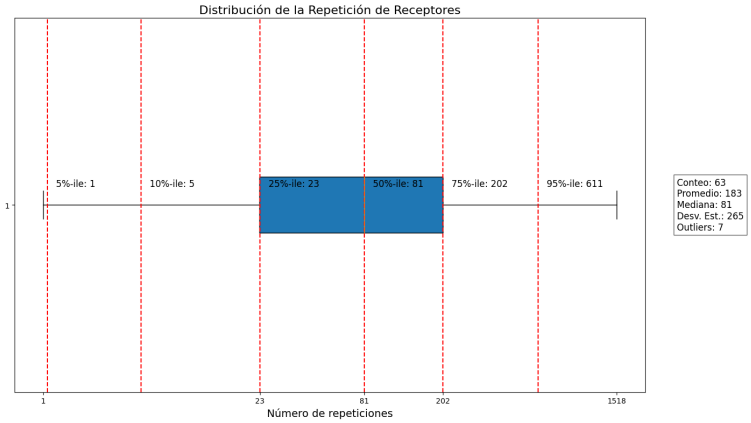

Figura 5.7: Distribución de la Repetición de Receptores Inicial

Las observaciones clave del análisis inicial de receptores son:

- La dispersión en la frecuencia de aparición de los receptores es considerable, abarcando un rango desde 23 hasta 202 apariciones por receptor.
- El cuartil inferior (25 % de los receptores) muestra una frecuencia relativamente baja de apariciones, lo que podría comprometer la representatividad de estos casos en el análisis global.
- Existe una variabilidad significativa en la cantidad de facturas asociadas a cada receptor, lo que podría indicar diferencias en el volumen de transacciones o en la completitud de los datos por receptor.

Considerando estas observaciones, se tomó la decisión metodológica de implementar un criterio de exclusión para mejorar la robustez del análisis:

- Se estableció un umbral mínimo de 25 facturas por receptor.
- Los receptores con menos de 25 facturas fueron excluidos del análisis subsiguiente.
- Esta decisión resultó en una reducción del 4.7 % en la cantidad total de registros en el conjunto de datos.

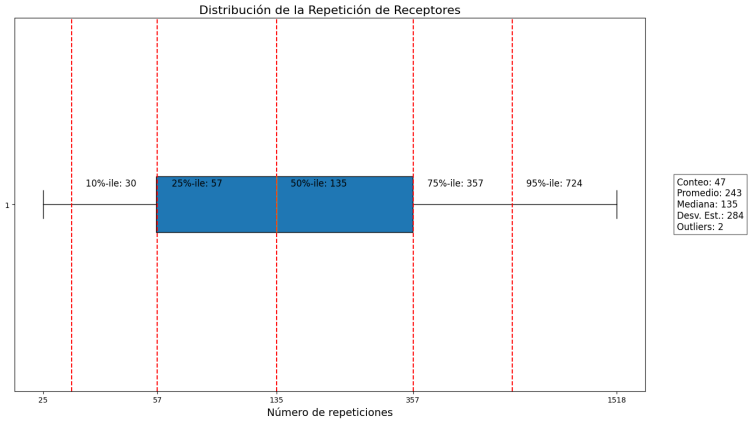

Figura 5.8: Distribución de la Repetición de Receptores Final

La aplicación de este filtro produjo cambios significativos en la distribución de los receptores, como se muestra en la Figura 5.8.Los resultados post-filtrado revelan:

- Una reducción en el número total de receptores a 47.
- Una distribución más homogénea de la frecuencia de aparición entre los receptores remanentes.
- Un aumento en el límite inferior de apariciones por receptor, mejorando la representatividad de cada caso en el análisis.

Es importante señalar que, aunque se consideró la posibilidad de elevar el umbral mínimo a 100 facturas por receptor, lo cual habría resultado en una reducción del 11.3 % del conjunto de datos, se optó por mantener el umbral en 25 por las siguientes razones:

1. Preservar un tamaño de muestra estadísticamente significativo, manteniendo la reducción por debajo del 5 % del conjunto de datos original.
2. Balancear la necesidad de representatividad por receptor con la conservación de la diversidad en el conjunto de datos.
3. Evitar una pérdida excesiva de información que podría ser valiosa para el análisis global y la generalización de los resultados.

Esta decisión metodológica busca optimizar el equilibrio entre la calidad y la cantidad de los datos, proporcionando una base más sólida para los análisis subsiguientes y el desarrollo de modelos predictivos, mientras se mantiene un nivel aceptable de diversidad en el conjunto de datos.

Análisis del Total

Aunque inicialmente no se prestó demasiada atención a esta variable, ya que los valores de interés primario se encuentran en los ítems de cada factura, es pertinente mostrar su distribución para resaltar la diversidad presente en el conjunto de datos. En la figura 5.9 se observa que la mayoría de las facturas presentan un total que oscila entre 1 y 2,500 quetzales. Este comportamiento puede atribuirse a que, aunque algunas facturas contienen ítems de menor valor unitario, la cantidad de ítems incluidos incrementa el valor total de la factura. Este análisis preliminar permitió identificar la necesidad de explorar en mayor profundidad la distribución de los ítems dentro de las facturas para obtener una comprensión más integral de la composición de los totales.

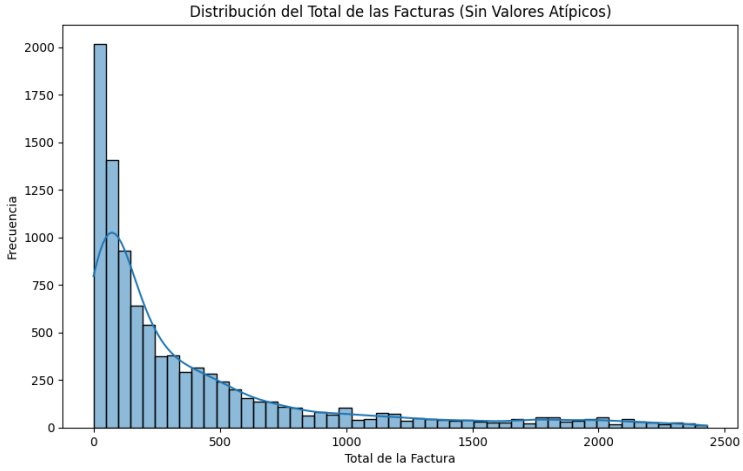

Figura 5.9: Distribución del total de las facturas sin valores atípicos

Análisis General de Ítems

Dado que el objetivo principal de este proyecto es clasificar automáticamente la mayor cantidad posible de ítems presentes en una factura, realizar un análisis exhaustivo de esta variable es esencial como siguiente paso. Este análisis preliminar proporciona una base sólida antes de proceder a un examen más detallado de las características específicas de los ítems.

En la figura 5.10 se presenta la distribución del número de ítems por factura. La mayoría de las facturas incluyen únicamente un ítem, lo que sugiere que las transacciones registradas en este conjunto de datos tienden a ser operaciones simples, involucrando un solo producto o servicio. Esta tendencia es relevante, ya que puede influir en la complejidad de la clasificación automática que se pretende realizar, dado que las transacciones más simples podrían requerir estrategias de clasificación diferentes en comparación con las más complejas.

Para complementar este análisis, se examinó la relación entre el número de ítems por factura y el total monetario de la misma, como se ilustra en la figura 5.11. Los resultados indican que las facturas con un total entre 1000 y 5000 quetzales tienen, en promedio, 4 ítems. Esta observación sugiere una distribución en forma de pirámide, donde el incremento en el valor total de la factura se asocia con un aumento en el número de ítems, aunque de manera no lineal. Este patrón es consistente con la expectativa de que las compras de mayor valor generalmente incluyen una mayor cantidad de productos o servicios.

Este análisis inicial reveló patrones significativos que se debieron considerar en etapas posteriores del proyecto, donde se examinaron las características específicas de los ítems con mayor detalle. Comprender la frecuencia y la distribución de los ítems en relación con el valor total de la factura pudo proporcionar información valiosa para desarrollar estrategias más efectivas de clasificación.

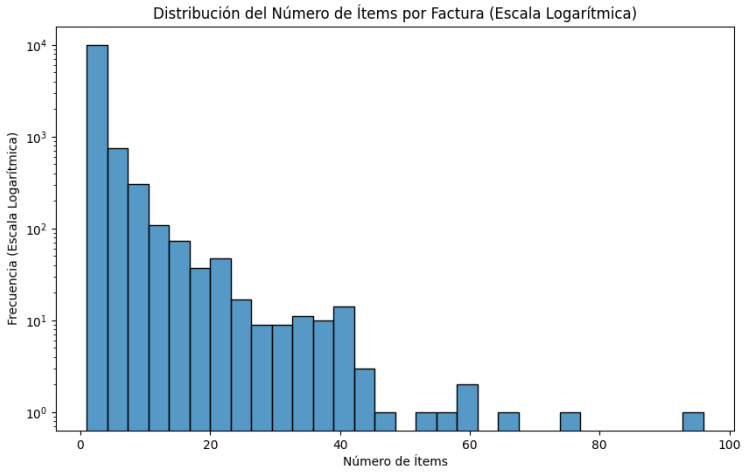

Figura 5.10: Distribución del número de ítems por factura

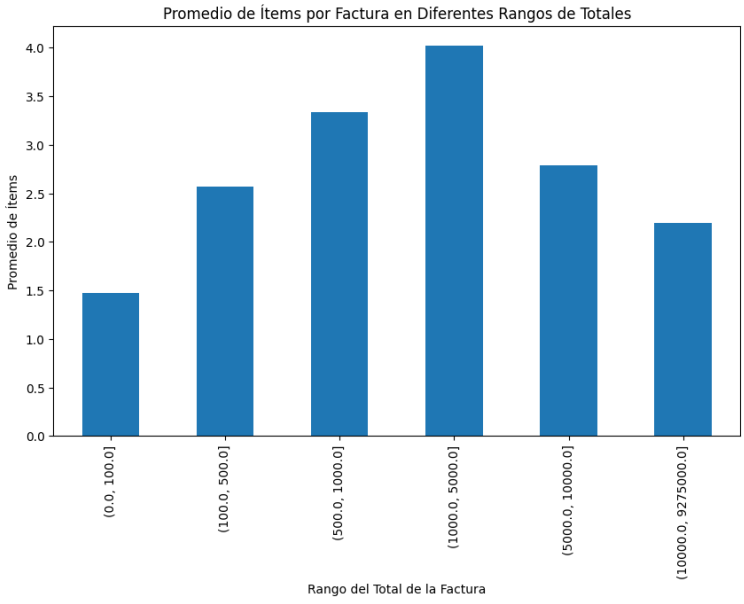

Figura 5.11: Promedio de ítems por factura en diferentes rangos de totales

Después de realizar un análisis general de los ítems, se procedió a crear un nuevo conjunto de datos en el que cada registro representa un solo producto, lo que incrementó el número de registros de 11,437 a 28,858. Este aumento en la granularidad permitió realizar un análisis más profundo y detallado, que es crucial para la predicción de la cuenta contable asociada a cada ítem. A continuación, se describen las siete variables más importantes consideradas en este análisis:

Item-Cantidad Esta variable ofrece una gran cantidad de información relevante. Como se mencionó anteriormente, los datos muestran una clara tendencia hacia la unicidad, es decir, la mayoría de las facturas se refieren a un solo ítem. Sin embargo, es posible que ese ítem se subdivida en múltiples unidades del mismo producto. Este fenómeno es evidente en la figura 5.12, que ilustra la distribución de la cantidad de ítems dentro de las facturas.

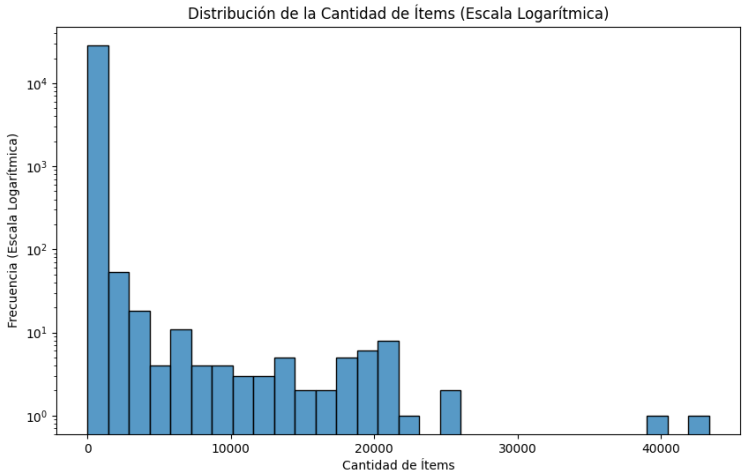

Figura 5.12: Distribución de la variable cantidad dentro de ítems

Item-Total Esta variable está fuertemente correlacionada con el total de la factura, respaldado con un coeficiente de correlación de 0.99. Al comparar las figuras 5.13 y 5.9, se observa que la mayoría de los valores de están concentrados entre 0 y 600 quetzales, mientras que los valores del total de la factura se concentran entre 0 y 2500 quetzales. Esta diferencia resalta la importancia de analizar los totales a nivel de ítem, lo que proporciona una visión más detallada de la composición del valor de las facturas y puede revelar patrones que no son evidentes al considerar solo el total general de la factura.

Item-Total Unitario Aunque las variables anteriores (Cantidad y Total) son de gran interés para el proyecto, se decidió hacer el análisis más granular unificando ambas en una sola variable llamada precio unitario. Esta variable, que ya estaba presente en el XML de la factura, no estaba incluida en el conjunto de datos original, por lo que se calculó y se presenta en la figura 5.14. Los datos mostraron una concentración aún mayor en el rango de 0 a 250 quetzales, lo que se consideró beneficioso para el proyecto, dado que reduce el rango de selección, facilitando así el análisis.

Item-Tipo La variable `Item-Tipo` clasifica los ítems en dos categorías: `B` para Bienes y `S` para Servicios. Este análisis reveló que los ítems clasificados como Bienes (`B`) constituyen el 75 % del total, mientras que los Servicios (`S`) representan el 25 %. La figura 5.15 muestra esta distribución, indicando una mayor prevalencia de transacciones relacionadas con bienes físicos en comparación con servicios. Esta distinción fue importante para entender la naturaleza de las transacciones en el dataset, y podría influir en cómo se interpretan y gestionan los datos en análisis futuros.

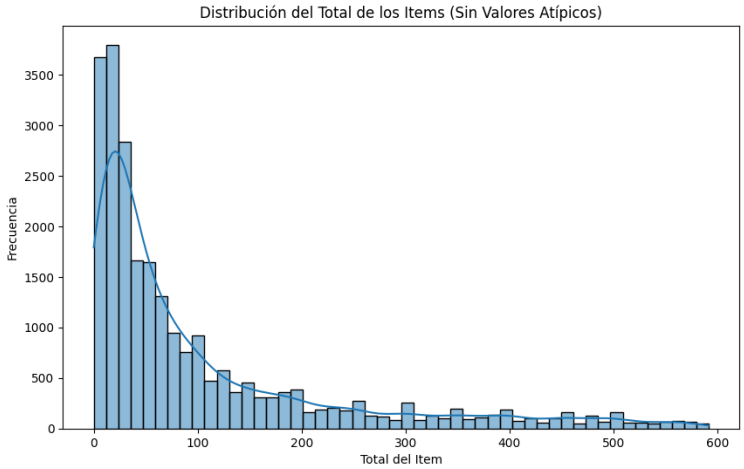

Figura 5.13: Distribución del total por ítem

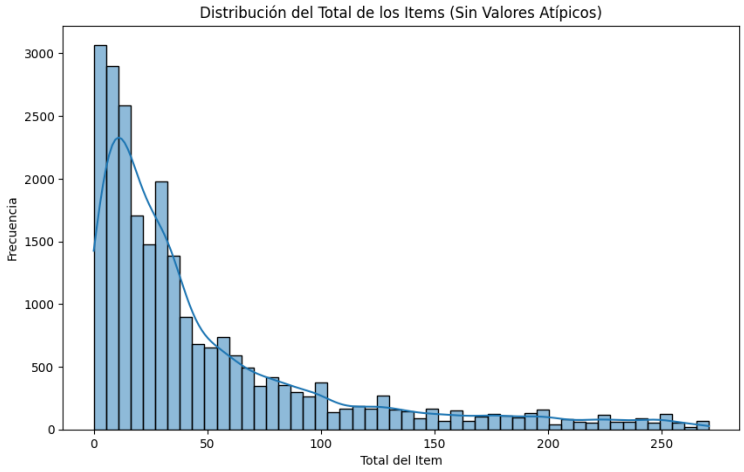

Figura 5.14: Distribución del precio unitario por ítem

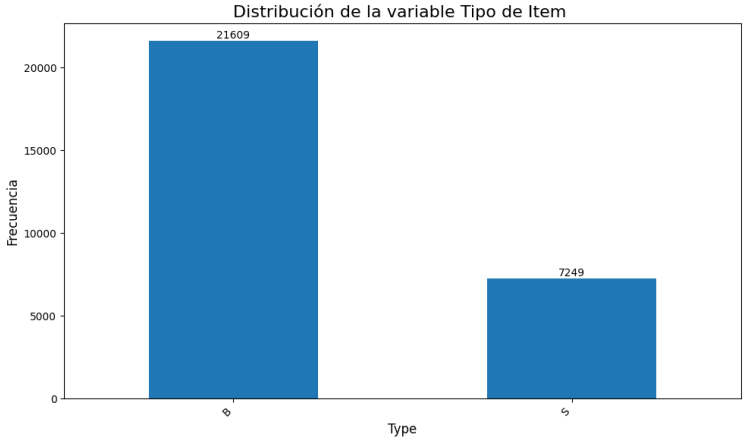

Figura 5.15: Distribución de la variable Tipo de Item

Item-Unidad de Medida Se decidió excluir completamente esta variable del análisis debido a su naturaleza inconsistente. Este campo es de libre entrada por parte de los emisores de las facturas, lo que ha resultado en 143 valores únicos, muchos de ellos con errores ortográficos, incoherencias y falta de uniformidad. La ausencia de estandarización en este campo lo hace poco confiable para cualquier análisis significativo.

Item-Descripción Inicial y Final Al igual que la unidad de medida, este campo es de libre entrada por el usuario. Sin embargo, a diferencia del caso anterior, esta variable tiene el potencial de contener información valiosa que justifica un análisis más profundo. Por esta razón, se aplicaron técnicas de Procesamiento de Lenguaje Natural (PLN) para extraer y analizar el contenido de estas descripciones, lo cual se detallará en secciones posteriores.

Item-Clasificación Contable Este campo representó uno de los mayores desafíos para el sistema debido al marcado desbalance presente en el dataset desde el principio. En muchas empresas, el número mínimo de facturas a analizar es de 25, pero en este caso se encontró que la frecuencia de muchas clasificaciones contables era inferior a 10, como se ilustra en la figura 5.16. Según el análisis de distribución, más del 50 % de las clasificaciones contables registraron una frecuencia menor a 7. Esto llevó a la necesidad de establecer un umbral mínimo de 6, lo que redujo el número de clasificaciones contables de 655 a 405, una disminución de casi el 40 %. Esta reducción significativa se consideró una limitación importante para el modelo futuro, ya que restringió considerablemente el número de clasificaciones contables disponibles para el análisis. Se espera que esta limitación se mitigue con el tiempo, a medida que aumente el volumen de datos y, por ende, la representatividad de las clasificaciones contables.

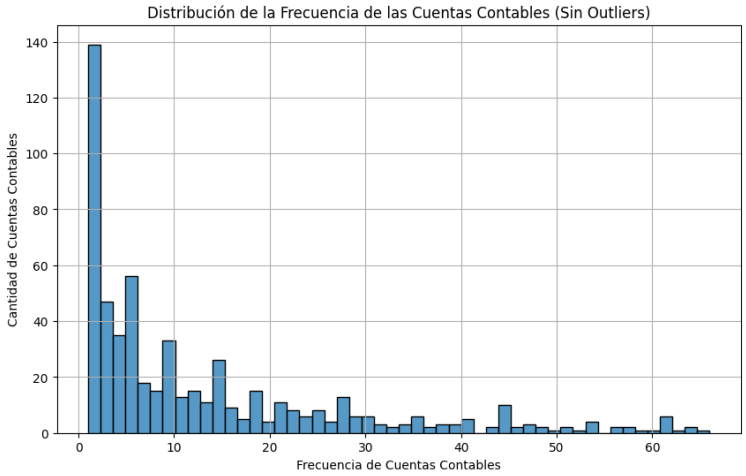

Figura 5.16: Frecuencia de las Clasificaciones Contables

Por otro lado, se observa una sobre-representación en ciertas clasificaciones contables, como se muestra en la figura 5.17. Algunas de estas clasificaciones superan las 1,400 ocurrencias, mientras que la curva de frecuencia desciende abruptamente hasta menos de 200 en la clasificación número 32 más frecuente. Esta distribución desigual representa un problema significativo, ya que es común que algunas empresas generen un volumen de facturas considerablemente mayor que otras, exacerbando el desbalance en las clasificaciones y limitando la efectividad de los futuros modelos de clasificación. Para ilustrar mejor este desbalance, se generaron dos gráficas (5.18 y 5.19)que muestran cómo 15 empresas concentran más del 75 % de todos los registros. Para abordar esta problemática, se diseñó un sistema de Machine Learning Engineering que tenía como objetivo escalar la información de manera que este desbalance se reduzca con el tiempo. Esta solución será detallada en secciones posteriores.

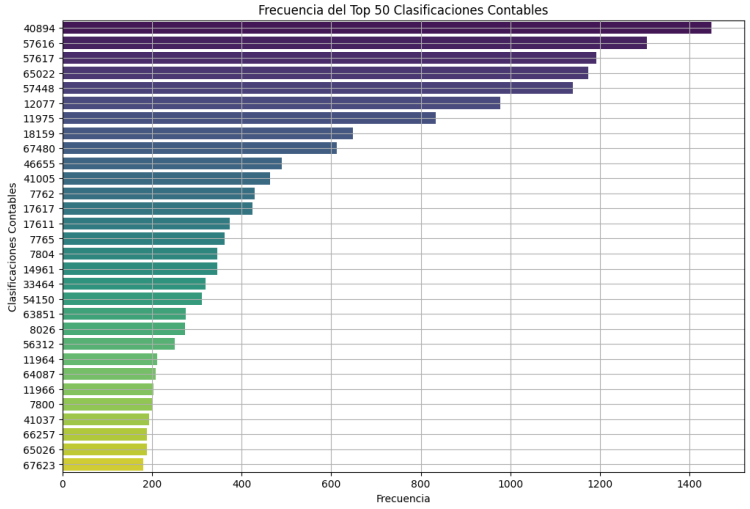

Figura 5.17: Distribución de la Frecuencia de las Clasificaciones Contables

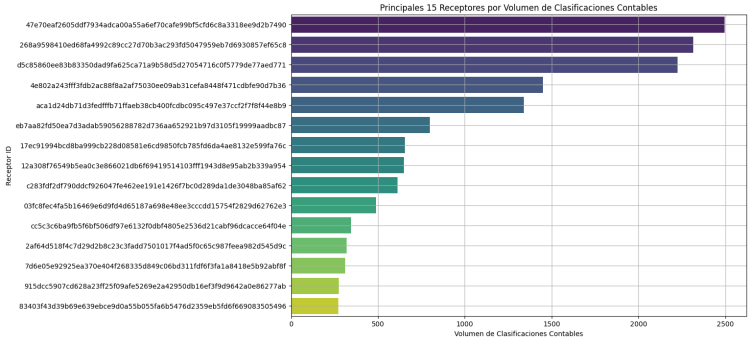

Figura 5.18: Principales 15 Receptores por Volumen de Clasificaciones Contables

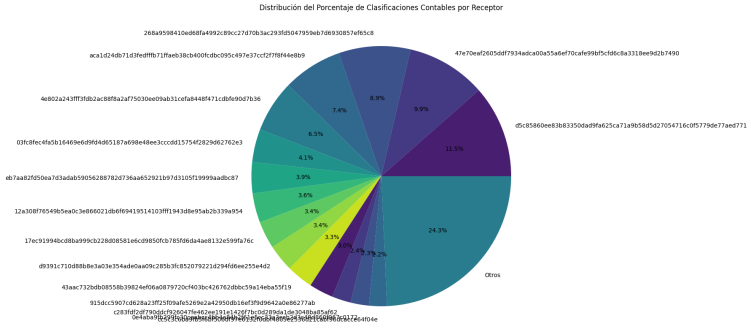

Figura 5.19: Distribución del Porcentaje de Clasificaciones Contables por Receptor

5.6.2. Filtro Final de Variables

El proceso de análisis y expansión del dataset de facturas a nivel de ítems permitió realizar una selección más precisa de las variables relevantes para el estudio. Esta etapa de filtrado inicial fue crucial para optimizar el manejo de datos y enfocar el análisis en los aspectos más significativos de la información disponible. Tras un exhaustivo examen de todas las variables presentes en el dataset original, se identificaron diez variables clave. Estas variables fueron seleccionadas de acuerdo con su relevancia para la comprensión de las transacciones, su potencial para revelar patrones y su utilidad en la clasificación contable. A continuación, se presenta la lista de variables seleccionadas junto con una breve justificación de su importancia:

1. **Fecha**: Se podría considerar para capturar posibles patrones temporales en la clasificación contable. Sin embargo, en este caso los rangos temporales son limitados impidiendo encontrar patrones útiles.
2. **Emisor**: Se incluyó para identificar si ciertos emisores están asociados con clasificaciones contables específicas.
3. **Establecimiento**: Se contempló para determinar si el lugar de la transacción influye en la clasificación contable.
4. **Receptor**: Se consideró porque en definitiva el destinatario de la factura tiene relación con la clasificación contable asignada.
5. **Tipo**: Se incluyó para examinar cómo la naturaleza de la transacción podría determinar su clasificación contable.
6. **Total Unitario**: Se consideró para analizar si el valor monetario tiene impacto en la clasificación contable.
7. **Descripción Inicial**: Se incluyó como fuente principal de información textual para predecir la clasificación contable.
8. **Descripción Final**: Se contempló para proporcionar detalles adicionales que podrían mejorar la precisión en la predicción de la clasificación contable.
9. **Clasificación Contable**: Se incluyó como la variable objetivo a predecir basándose en las demás variables.

5.7. Selección de Modelo de Predicción

El objetivo principal era implementar un sistema de clasificación que permitiera automatizar la asignación de productos y servicios a cuentas contables específicas, garantizando una alta precisión y eficiencia operativa. Para seleccionar el modelo de predicción adecuado, se establecieron ciertos criterios clave que guiaron el análisis y la elección del modelo:

1. Eficiencia computacional y tamaño del modelo: El modelo seleccionado debía ser lo suficientemente pequeño para consumir pocos recursos computacionales, ya que esto garantizaría un procesamiento eficiente sin sobrecargar los sistemas donde se implementara.
2. Almacenabilidad: El modelo debía ser fácil de almacenar y mantener, lo que permitía que se guardaran versiones entrenadas y se cargaran rápidamente cuando fuera necesario. Este criterio era clave para asegurar que se pudieran realizar actualizaciones periódicas sin complicaciones técnicas.
3. Escalabilidad: Dado que el número de facturas y el tamaño de los datos podían crecer con el tiempo, el modelo debía ser escalable. Esto aseguraría que el rendimiento no se viera afectado significativamente con el aumento de datos.

5.7.1. Modelos Evaluados

Con estos criterios en mente, se procedió a seleccionar y evaluar varios modelos de predicción comúnmente utilizados en problemas de clasificación. Cada uno de ellos fue evaluado por su capacidad de clasificar correctamente las facturas y por su alineación con los criterios previamente definidos.

1. Logistic Regression: Un modelo sencillo y eficiente en términos computacionales, ideal para problemas de clasificación binaria y multiclase. Aunque su desempeño puede verse limitado si las relaciones entre las variables no son lineales, se seleccionó debido a su simplicidad y rapidez en el entrenamiento. Los hiperparámetros utilizados fueron:
   - **Penalty:** L2
   - **Multi-class:** auto
   - **Tolerance (tol):** 1×10^{-4}
   - **Solver:** lbfgs
   - **Max iterations:** 1000
2. Random Forest: Un modelo de ensamble basado en árboles de decisión, conocido por mejorar la precisión en problemas de clasificación más complejos. Sin embargo, su uso puede ser más intensivo en términos computacionales, y su implementación requiere una mayor capacidad de procesamiento. Los hiperparámetros utilizados fueron:
   - **Number of estimators (n_estimators):** 100
   - **Max features:** sqrt
   - **Max leaf nodes:** none
   - **Number of jobs (n_jobs):** 1
   - **Bootstrap:** False
3. Decision Tree: Evaluado por su fácil interpretación y capacidad para clasificar datos complejos sin necesidad de un preprocesamiento intensivo. Aunque propenso al sobreajuste, su simplicidad lo hacía una opción atractiva para escenarios menos complejos. Los hiperparámetros utilizados fueron:
   - **Criterion:** Gini
   - **Splitter:** best
   - **Min samples split:** 2
   - **Min samples leaf:** 1
   - **Class weight:** equal
4. SVM (Support Vector Machine): Este modelo se destacó por su capacidad para manejar problemas de clasificación no lineales. Sin embargo, su alto costo computacional podría limitar su uso con grandes volúmenes de datos.
5. K-Nearest Neighbors (KNN): Este modelo, aunque simple, puede volverse computacionalmente costoso a medida que crece el conjunto de datos, lo que puede afectar su escalabilidad.
6. Naive Bayes: Seleccionado por su bajo costo computacional y su capacidad para manejar problemas de clasificación con distribuciones probabilísticas simples. Aunque puede no ser tan preciso como otros modelos, su eficiencia lo convierte en una opción viable cuando la simplicidad es prioritaria.

5.7.2. Evaluación de los Modelos

Cada modelo fue entrenado y evaluado utilizando conjuntos de datos históricos de facturas con etiquetas correspondientes a sus cuentas contables. Se consideraron diversas métricas, incluidas precisión (accuracy), recall, F1-score y tiempo de entrenamiento, para medir la capacidad predictiva de cada modelo.

5.8. Integración de Procesamiento de Lenguaje Natural (NLP) en la Clasificación de Facturas

Con el objetivo de mejorar la precisión del modelo de clasificación de facturas, se implementaron técnicas de Procesamiento de Lenguaje Natural (NLP) para procesar y analizar el valor "descripción"de los productos o servicios incluidos en las facturas. Este campo, proporcionado por la factura original, contenía información textual que podía ser aprovechada para realizar una doble clasificación: una basada en la descripción del producto y otra basada en las demás variables del dataset. La combinación de ambas se planteó como un mecanismo para aumentar la precisión del modelo final.

5.8.1. Limpieza y Normalización de las Descripciones

El primer paso fue procesar las descripciones del dataset, que en muchos casos presentaban datos parcialmente ilegibles, como códigos, números y referencias internas utilizadas por las empresas. Estas imperfecciones dificultaban el análisis, ya que el texto no era fácilmente interpretable por el modelo de aprendizaje automático. Para abordar este desafío:

Se desarrolló un proceso de limpieza que eliminó códigos innecesarios, números y otros caracteres irrelevantes que podían interferir con la clasificación. Este paso se completó con éxito, lo que permitió transformar las descripciones en datos más consistentes y estructurados.

El siguiente obstáculo se presentó en forma de errores ortográficos dentro de las descripciones. Para garantizar que el modelo pudiera procesar el texto de manera efectiva, era necesario corregir estos errores. Se diseñó un algoritmo de corrección ortográfica que, utilizando un amplio conjunto de palabras en español extraído de la RAE, comparaba cada palabra de la descripción con las palabras más cercanas en el pool y realizaba las correcciones automáticamente. Este enfoque permitió una estandarización del lenguaje dentro de las descripciones, facilitando el trabajo del modelo.

5.8.2. Reducción de la Longitud de las Descripciones

Una vez limpiadas las descripciones, se identificó otro problema: la longitud excesiva del texto en algunas de ellas, lo que podría impactar negativamente el rendimiento del modelo. Para reducir esta longitud sin perder información relevante, se implementó un proceso de reducción de características basado en la frecuencia de las palabras.

1. Se recopiló un corpus con todas las descripciones y se analizaron las palabras más repetidas.
2. Se comparó este análisis con un documento proporcionado por la RAE que lista las palabras más frecuentes del español, con una frecuencia normalizada.
3. Con base en ambas frecuencias, se seleccionaron las cuatro palabras más relevantes de cada descripción, lo que permitió reducir la cantidad de información procesada por el modelo sin comprometer su capacidad para identificar las características clave de cada producto o servicio.

5.8.3. Clasificación Global utilizando un Modelo de Lenguaje (BERT)

Después de reducir y limpiar las descripciones, se utilizó el modelo BERT [19] (Dccuchile/Bert-Base-Spanish-Wwm-Cased · Hugging Face, 2019) para generar una clasificación general de cada descripción. El propósito de esta clasificación era asignar cada producto o servicio a una de las siguientes categorías predefinidas, que engloban una gran variedad de sectores:

1. Electrónica
2. Alimentos y Bebidas
3. Ropa y Calzado
4. Hogar y Jardín
5. Salud y Belleza
6. Juguetes y Juegos
7. Automotriz
8. Consultoría y Asesoría
9. Servicios Médicos
10. Servicios Educativos
11. Servicios de Tecnología
12. Servicios de Mantenimiento y Reparación
13. Servicios de Transporte
14. Entretenimiento y Eventos
15. Deportes y Recreación
16. Productos para Mascotas
17. Papelería y Oficina
18. Herramientas y Equipamiento
19. Servicios Financieros
20. Viajes y Turismo
21. Muebles y Decoración
22. Construcción y Remodelación
23. Servicios de Limpieza
24. Servicios de Seguridad
25. Publicidad y Marketing
26. Servicios de Belleza y Spa
27. Logística y Almacenamiento
28. Servicios Legales
29. Agricultura y Ganadería
30. Energía y Suministros
31. Servicios de Alimentos y Catering
32. Servicios de Bienestar y Fitness
33. Servicios de Telecomunicaciones

34. Arte y Artesanías

Estas categorías fueron definidas en la plataforma de origen de los datos y representaban una clasificación global de los productos o servicios facturados. El modelo BERTO fue entrenado para procesar las descripciones y asignar automáticamente cada una a la clasificación más adecuada.

5.8.4. Integración de la Clasificación Global en el Modelo de Machine Learning

Una vez que cada descripción fue clasificada de manera global, se añadió esta nueva variable de clasificación global al conjunto de datos como un campo adicional. La idea detrás de esta integración era que la clasificación general del producto o servicio, en combinación con las otras características del dataset, podría mejorar significativamente la capacidad del modelo de machine learning para realizar una clasificación precisa.

Se entrenó el modelo de machine learning final, incorporando tanto las variables originales como la nueva clasificación global generada a partir de las descripciones de los productos. El objetivo de esta combinación era aumentar la precisión del sistema, aprovechando tanto el contenido textual de las facturas como las demás variables estructuradas en el dataset.

5.9. Implementación de un Sistema de Preprocesamiento, Entrenamiento y Despliegue del Modelo

Para garantizar la escalabilidad y eficiencia en el procesamiento de grandes volúmenes de facturas, se implementó un sistema automatizado basado en *cloud functions*. Este sistema se diseñó para llevar a cabo las tareas de preprocesamiento de datos, entrenamiento del modelo y despliegue de forma eficiente y modular, permitiendo la clasificación automatizada de facturas con un enfoque adaptable a las necesidades de cada empresa. A continuación, se detalla el flujo de trabajo y los componentes clave del sistema.

5.9.1. Preprocesamiento de Facturas en Batches

El primer paso en el proceso es la *obtención y preprocesamiento de las facturas* almacenadas en la base de datos del sistema general. Para manejar esto, se diseñó una *cloud function* que se activa de forma reactiva, a través de un botón en la interfaz del sistema, lo que permite al usuario iniciar el procesamiento en cualquier momento. Este sistema es escalable y eficiente, dado que las funciones en la nube están limitadas en tiempo de ejecución, y por ello, se introdujo una lógica de procesamiento por *batches*:

- **Obtención de Facturas:** La *cloud function* consulta la base de datos y obtiene todas las referencias a las facturas que deben ser procesadas. Si el número de facturas excede las 250, el sistema divide automáticamente el total en *batches* de 250 facturas. Esta segmentación se realiza para evitar sobrecargar la ejecución de la función, que tiene un límite de 9 minutos por ejecución. Esto también facilita el manejo de grandes volúmenes de facturas, de hasta 50,000 o más.
- **Procesamiento por Batches:** Una vez creados los *batches*, se activa una segunda *cloud function* para procesar cada lote de 250 facturas. Esta función realiza el *preprocesamiento de las descripciones y demás variables* relacionadas con cada ítem de las facturas, aplicando las técnicas de limpieza y transformación descritas anteriormente. El resultado de este preprocesamiento es almacenado en archivos JSON, cada uno nombrado con un índice correspondiente al *batch* procesado, lo que permite mantener un orden y seguimiento en el procesamiento.

5.9.2. Almacenamiento Estructurado y Preparación de Datos

Los archivos JSON generados contienen las 250 facturas preprocesadas, con los datos estructurados de tal manera que puedan ser utilizados en el entrenamiento del modelo. En esta fase se asegura que:

- Los datos se normalizan y reorganizan para que el modelo de *machine learning* pueda ser entrenado correctamente.
- Se incluyen tanto las variables numéricas como las no numéricas, con el preprocesamiento adecuado para preparar la información para la fase de entrenamiento.

Este enfoque modular permite que cada *batch* de facturas sea procesado de forma independiente, lo que facilita la escalabilidad del sistema al poder manejar grandes volúmenes de datos en paralelo, sin interrupciones.

5.9.3. Entrenamiento del Modelo

Una vez que todas las facturas han sido preprocesadas, una nueva *cloud function* se encarga de *entrenar el modelo* utilizando los datos previamente almacenados. El flujo de entrenamiento sigue los siguientes pasos:

- **Recogida de Datos:** La *cloud function* encargada del entrenamiento recopila todos los archivos JSON generados durante el preprocesamiento y los combina para crear el conjunto de datos final.
- **Codificación y Creación de Variables Dummies:** Los datos no numéricos son codificados mediante técnicas de *encoding*, y se generan las variables dummies necesarias para que el modelo pueda procesar las variables categóricas de manera efectiva. Esto incluye el tratamiento de las descripciones de productos que se han transformado mediante NLP.
- **Entrenamiento y Métricas:** Con el conjunto de datos preparado, el modelo se entrena utilizando el algoritmo previamente seleccionado. Durante esta fase, se generan métricas de rendimiento, como precisión, *recall*, F1-score, entre otras, que permiten evaluar la calidad del modelo en función de la clasificación de las facturas.
- **Almacenamiento del Modelo Entrenado:** Una vez finalizado el entrenamiento, el modelo resultante se guarda en una *base de datos especializada* para el almacenamiento de archivos de modelos, lo que permite que sea accesible para futuras predicciones por parte de otras *cloud functions*. Esto asegura que el sistema sea eficiente, al permitir el acceso a los modelos previamente entrenados de manera rápida y organizada.

5.9.4. Modelos por Empresa para Evitar Sesgos y Mejorar Escalabilidad

Un aspecto clave de este sistema es que **se genera un modelo específico por empresa**. Cada empresa dentro de la plataforma cuenta con dos modelos de clasificación: uno para las compras y otro para las ventas. Esta decisión se tomó por varias razones:

- **Reducir los Grupos y Errores de Codificación:** Al generar modelos individuales por empresa, se evita la mezcla de nomenclaturas contables entre diferentes empresas, lo que podría introducir errores en la clasificación y en el procesamiento de datos.
- **Escalabilidad a Largo Plazo:** La creación de modelos separados permite que el sistema crezca de forma escalable, ya que cada empresa puede tener volúmenes de datos y características particulares que pueden diferir de otras empresas.

- **Minimizar el Sesgo por Volumen de Facturas:** Al tratar a cada empresa de manera individual, se previenen sesgos que podrían ocurrir debido a diferencias en la cantidad de facturas procesadas entre empresas de distintos tamaños. Esto asegura que el modelo se entrene correctamente sin influencias externas.

Además, este enfoque se centra en las facturas que ya han sido *verificadas y clasificadas correctamente*, lo que asegura que solo se utilicen datos de alta calidad para entrenar el modelo, evitando errores o inconsistencias en las predicciones futuras.

5.9.5. Despliegue y Uso del Modelo en Predicciones

Una vez que los modelos han sido entrenados y almacenados, están listos para ser utilizados en la clasificación automática de nuevas facturas. Las predicciones se realizan a través de otra *cloud function*, que accede al modelo correspondiente de cada empresa y aplica la clasificación de manera automática. Este proceso asegura que el sistema continúe siendo eficiente, escalable y preciso, cumpliendo con los objetivos planteados de reducir la intervención manual y optimizar el flujo contable.

CAPÍTULO 6

Resultados

6.1. Desempeño del Modelo General

6.1.1. Regresión Logística

El modelo de Regresión Logística mostró un desempeño satisfactorio en la tarea de clasificación general de facturas, alcanzando una exactitud de 0.8373. Este resultado sugiere que la Regresión Logística es capaz de captar de manera efectiva los patrones generales presentes en el conjunto de datos, logrando una alta capacidad de predicción en la mayoría de las clases.

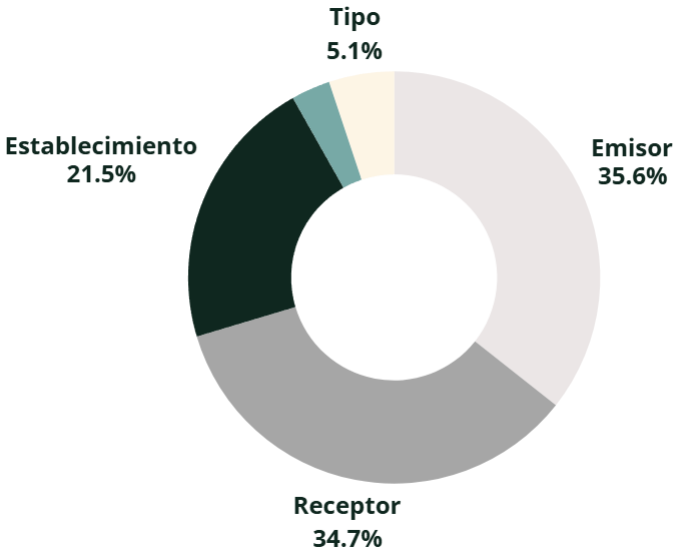

Figura 6.1: Feature Importance de la Regresión Logística

	Precision	Recall	F1-Score	Support
Accuracy			**0.84**	**7375**
Macro avg	0.50	0.50	0.48	7375
Weighted avg	0.81	0.84	0.81	7375

Tabla 6.1: Resultados de clasificación para el modelo Regresión Logística

El modelo de Regresión Logística demostró ser especialmente eficiente en clases que tienen una representación significativa en el conjunto de datos, como la clase 30 y la clase 290, que corresponden a tipos de facturas comunes. Sin embargo, en clases con menos representaciones, el desempeño fue más limitado, lo cual se abordará en la discusión posterior.

6.1.2. Máquinas de Soporte Vectorial (SVM)

El desempeño del modelo SVM fue considerablemente inferior en comparación con el de la Regresión Logística. La exactitud global obtenida fue de apenas 0.1200, lo que indica que este modelo no fue capaz de aprender patrones adecuados para la clasificación de facturas en un contexto general que abarque múltiples empresas.

	Precision	Recall	F1-Score	Support
Accuracy			**0.12**	**7375**
Macro avg	0.00	0.01	0.00	7375
Weighted avg	0.02	0.12	0.04	7375

Tabla 6.2: Resultados de clasificación para el modelo SVM

6.1.3. K-Nearest Neighbors (KNN)

El modelo K-Nearest Neighbors (KNN) obtuvo una exactitud de 0.6353, lo que indica un desempeño moderado. En términos generales, el KNN demostró ser capaz de capturar algunas relaciones de cercanía entre las facturas, aunque su precisión disminuyó en las clases con menor cantidad de datos o con patrones menos claros.

	Precision	Recall	F1-Score	Support
Accuracy			**0.64**	**7375**
Macro avg	0.37	0.37	0.35	7375
Weighted avg	0.62	0.64	0.61	7375

Tabla 6.3: Resultados de clasificación para el modelo KNN

6.1.4. Naive Bayes

El modelo de Naive Bayes mostró un rendimiento aceptable con una exactitud de 0.8350. Este resultado era esperado dado que este tipo de modelo se basa en la independencia de características, lo que puede limitar su capacidad para captar relaciones más complejas en los datos de facturas, pero aún así muestra un rendimiento competitivo en tareas de clasificación general.

	Precision	Recall	F1-Score	Support
Accuracy			**0.83**	**7375**
Macro avg	0.64	0.69	0.63	7375
Weighted avg	0.89	0.83	0.84	7375

Tabla 6.4: Resultados de clasificación para el modelo Naive Bayes

6.1.5. Árbol de Decisión

El modelo de Árboles de Decisión presentó una exactitud de 0.8816, posicionándose como una opción efectiva para la tarea de clasificación. Este modelo fue capaz de manejar de manera efectiva las clases con mayor representación y mostró robustez en las clases más pequeñas, aunque en estas últimas el desempeño fue inferior.

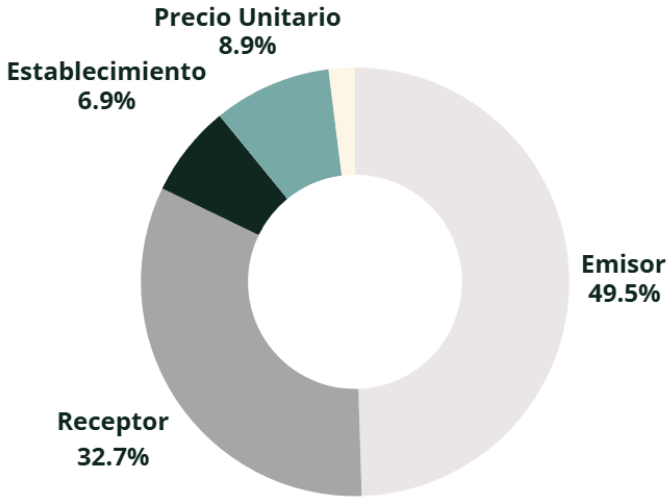

Figura 6.2: Feature Importance del Árbol de Decisión

	Precision	Recall	F1-Score	Support
Accuracy			**0.88**	**7375**
Macro avg	0.69	0.67	0.67	7375
Weighted avg	0.88	0.88	0.88	7375

Tabla 6.5: Resultados de clasificación para el modelo Árbol de Decisión

6.1.6. Bosques Aleatorios

El modelo de Bosques Aleatorios fue el segundo más preciso, con una exactitud de 0.9. Este modelo, al combinar múltiples árboles de decisión, logró un equilibrio entre precisión y recall en la mayoría de las clases. Esto lo convierte en una opción robusta para la tarea de clasificación general, aunque no superó al modelo de Regresión Logística.

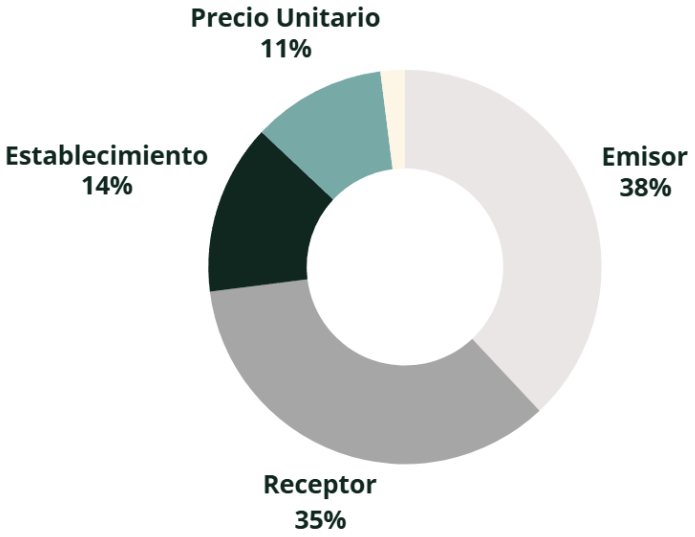

Figura 6.3: Feature Importance de Bosques Aleatorios

	Precision	Recall	F1-Score	Support
Accuracy			**0.90**	**7375**
Macro avg	0.72	0.71	0.70	7375
Weighted avg	0.89	0.90	0.89	7375

Tabla 6.6: Resultados de clasificación para el modelo Bosques Aleatorios

6.2. Comparación General de los Modelos

En la Tabla 6.7, se presenta un resumen de la exactitud alcanzada por cada modelo en la tarea de clasificación general de facturas. En la Tabla 6.8 se muestra un resumen del promedio de cada métrica para cada modelo según una muestra representativa de los datos (10 %). Una tabla por métrica que detalla los resultados por clase se presenta en los anexos 1, 2 y 3 .

Modelo	Exactitud
Bosques Aleatorios	0.8969
Árboles de Decisión	0.8816
Regresión Logística	0.8373
KNN	0.8300
Naive Bayes	0.6400
SVM	0.1200

Tabla 6.7: Comparación de la exactitud de los modelos

Métrica	**Bosque Aleatorio**	**Árbol de Decisión**	**Naïve Bayes**	**Regresión Logística**	**KNN**	**SVM**
Precisión	0.7692	0.7135	0.6706	0.4384	0.3816	0
Recall	0.7403	0.7155	0.7752	0.4461	0.3861	0
F1-Score	0.7480	0.7035	0.6933	0.4338	0.3544	0

Tabla 6.8: Desempeño promedio de los modelos en la muestra representativa.

En la Tabla 6.7 se observa un panorama general del desempeño de los modelos en términos de exactitud, destacándose los modelos de Bosque Aleatorio y Árbol de Decisión con valores de 84 % y 82 %, respectivamente. En contraste, se aprecia un desempeño considerablemente bajo para el modelo SVM, el cual obtuvo una exactitud de apenas 12 %, lo que anticipa un rendimiento poco efectivo al momento de clasificar un conjunto de datos con una distribución compleja y de múltiples clases.

Al analizar el promedio de los resultados específicos de la muestra representativa del 10 % de las clases en la Tabla 6.8, se puede observar que el Bosque Aleatorio demuestra ser el modelo más efectivo, alcanzando una precisión de 0.769, un recall de 0.740 y un F1-Score de 0.748. Este modelo mantiene un equilibrio adecuado entre la precisión y la sensibilidad, lo cual refleja su capacidad para generalizar correctamente en la muestra aleatoria de clases seleccionadas.

En comparación, el Árbol de Decisión muestra métricas ligeramente inferiores con una precisión de 0.713 y un recall de 0.715, resultando en un F1-Score de 0.703. Aunque estos valores son menores en comparación con el Bosque Aleatorio, siguen demostrando una capacidad aceptable para la clasificación de las clases en la muestra. Por otro lado, Naïve Bayes, si bien presenta una precisión de 0.670, destaca por su recall de 0.775, lo cual evidencia una buena sensibilidad del modelo para detectar correctamente las clases, aunque comprometiendo en cierta medida su precisión, como se refleja en un F1-Score de 0.693.

En el caso de los modelos lineales, los resultados son más limitados. La Regresión Logística presenta una precisión de 0.438 y un recall de 0.446, con un F1-Score de 0.433, lo que sugiere una menor capacidad para clasificar correctamente en este contexto específico. El modelo KNN, por su parte, evidencia aún mayores dificultades con una precisión de 0.381 y un F1-Score de 0.354, lo que sugiere que la técnica basada en vecinos más cercanos no logra manejar adecuadamente la complejidad del dataset utilizado.

Por último, el modelo SVM no obtuvo resultados satisfactorios, presentando un valor de 0 en todas las métricas evaluadas. Esto puede deberse a la baja representatividad de algunas clases en el conjunto de datos, una característica común en problemas de clasificación con una distribución de clases muy sesgada. Este comportamiento, junto con los resultados de la Tabla 6.7 , resalta las dificultades de este modelo para adaptarse al problema específico tratado en este estudio.

Es importante considerar que, debido a la naturaleza del dataset utilizado, con una representación desigual de las clases, algunos modelos se enfrentaron a limitaciones claras. Sin embargo, el desempeño de los modelos de Bosque Aleatorio, Árbol de Decisión y Naïve Bayes muestra una capacidad significativa para clasificar correctamente una proporción considerable de los registros en la muestra, como se evidencia en los resultados detallados de la Tabla 6.8.

6.3. Pruebas adicionales utilizando análisis de lenguaje natural (NLP)

Luego de haber identificado el modelo adecuado para este trabajo, se procedió a realizar pruebas adicionales utilizando una empresa que obtuvo las peores métricas de desempeño en el modelo general. Esta empresa particular presentó una exactitud de 0.64, lo que, aunque no óptimo, representa un avance significativo en la clasificación automática de facturas. Clasificar correctamente el 64 % de las facturas de manera automática es una mejora sustancial para los equipos de contabilidad, pues reduce la carga de trabajo manual. Sin embargo, dado que la precisión aún está lejos de ser ideal, se optó por complementar los datos con análisis de lenguaje natural (NLP).

Como se explicó en el capítulo anterior, se emplearon estrategias de reducción inteligente de texto y extracción de características fundamentales. Utilizando modelos de lenguaje de gran tamaño (LLMs), se clasificaron las descripciones de las facturas dentro de las categorías establecidas previamente. Inicialmente, se utilizó un modelo llamado BERT, el cual proporcionó resultados satisfactorios, pero posteriormente se optó por GPT-3 de OpenAI, debido a su mejor desempeño y eficiencia. El modelo GPT-3 logró clasificar correctamente el 80 % de las descripciones de las facturas, con un uso mínimo de tokens, lo que también mantuvo bajo el costo económico del proceso. Sin embargo, un inconveniente identificado fue el tiempo

adicional que este proceso introduce al pipeline de preprocesamiento.

6.3.1. Resultados obtenidos utilizando análisis NLP

A pesar de los avances obtenidos mediante la implementación de técnicas de NLP, los resultados no fueron tan significativos como se esperaba. Después de incluir las descripciones clasificadas como una variable adicional en el modelo de predicción, el incremento en la exactitud fue de solo un 5 %, alcanzando un 69 %. Aunque este aumento podría parecer atractivo, se debe considerar el impacto del tiempo y los recursos adicionales que implica incorporar el análisis de lenguaje natural en el pipeline. En consecuencia, la mejora marginal en la exactitud podría no justificar el costo añadido en este escenario específico.

Resultados sin utilizar NLP

	Precision	**Recall**	**F1-Score**	**Soporte**
Accuracy	0.64			330
Macro avg	0.58	0.57	0.54	330
Weighted avg	0.64	0.64	0.63	330

Tabla 6.9: Resultados de precisión, recall, F1-Score y soporte para la empresa con menor rendimiento sin técnicas de NLP

Los resultados obtenidos sin la utilización de técnicas de procesamiento de lenguaje natural (NLP) revelan un desempeño moderado del modelo, con una exactitud general de 0.64. Este nivel de precisión sugiere que el modelo es capaz de realizar predicciones correctas en un 64 % de los casos, aunque existen indicios de desequilibrios en el comportamiento hacia las diferentes clases. En los anexos se encuentran una tabla más detalladas para el rendimiento de este modelo (Bosque Aleatorio) 4.

Al observar las métricas promedio macro, los valores de precisión y recall se encuentran en 0.58 y 0.57 respectivamente, lo que indica que el modelo presenta una capacidad limitada para identificar correctamente cada clase de manera uniforme. El F1-Score macro de 0.54 refleja que, al promediar sin ponderación, la relación entre precisión y recall no está completamente balanceada, lo cual puede deberse a una inadecuada diferenciación de ciertas clases.

En cuanto a los valores ponderados, se alcanza una precisión y recall de 0.64, mientras que el F1-Score ponderado es de 0.63. Esto sugiere que, al considerar el peso de cada clase en el conjunto de datos, el modelo logra mantener un rendimiento coherente con la métrica de exactitud general. Sin embargo, la ligera disminución en la métrica F1-Score indica que existen casos en los que la precisión y la sensibilidad del modelo no están completamente alineadas.

Resultados utilizando NLP

Al incluir las técnicas de NLP para el procesamiento de las descripciones de las facturas, se observó un incremento en la exactitud, alcanzando un 0.69. El reporte de clasificación modificado se presenta en la Tabla 6.10. En los anexos se encuentran una tabla más detalladas para el rendimiento de este modelo (Bosque Aleatorio) 5.

Los resultados obtenidos tras la implementación de técnicas de procesamiento de lenguaje natural (NLP) indican una mejora en el desempeño del modelo. Con una exactitud incrementada a 0.69, se evidencia un avance respecto a los resultados obtenidos previamente sin el uso de NLP. Esta mejora de aproximadamente un 5 % en la exactitud sugiere que el procesamiento y análisis de las descripciones mediante técnicas NLP contribuyó a una mejor identificación de las clases.

	Precision	Recall	F1-Score	Soporte
Accuracy	0.69			330
Macro avg	0.70	0.70	0.68	330
Weighted avg	0.69	0.69	0.68	330

Tabla 6.10: Resultados de precisión, recall, F1-Score y soporte para la empresa con menor rendimiento con técnicas de NLP

Al analizar el reporte de clasificación, se observa un incremento tanto en la precisión como en el recall promedio macro, alcanzando ambos valores de 0.70, lo que demuestra una mayor capacidad del modelo para identificar correctamente las clases de forma uniforme. Además, el F1-Score macro de 0.68 refleja un equilibrio más consistente entre la precisión y la sensibilidad, lo cual es crucial para la correcta clasificación de todas las clases.

El análisis de los valores ponderados también indica una mejora, con métricas de precisión y recall de 0.69 y un F1-Score ponderado de 0.68. Estas cifras reflejan un avance uniforme al considerar la importancia de cada clase en el conjunto de datos. Sin embargo, la diferencia marginal entre las métricas macro y ponderadas sugiere que, si bien la inclusión de técnicas NLP mejoró el rendimiento global, el impacto no fue sustancialmente diferente para clases menos representadas.

En particular, se destaca el desempeño en ciertas clases específicas, como la clase 7803, donde se observó un incremento notable en la precisión y el F1-Score. Este aumento en clases específicas indica que las técnicas NLP permitieron al modelo captar mejor los patrones textuales relevantes para dichas clases. Sin embargo, este impacto positivo no se replicó con la misma magnitud en todas las clases, lo cual limita parcialmente los beneficios observados en la mejora global del modelo.

6.4. Resultados del Sistema de Preprocesamiento y Entrenamiento Automático

El sistema de preprocesamiento y entrenamiento automático se implementó utilizando una arquitectura basada en *cloud functions* y almacenamiento en la nube. A continuación, se presentan los resultados obtenidos durante la ejecución de estas tareas.

6.4.1. Preprocesamiento de Facturas

Luego de desplegar las *cloud functions* encargadas del preprocesamiento, se observó que los tiempos de procesamiento podían extenderse hasta 15 minutos en algunos casos. El sistema generaba archivos JSON con un mínimo de 250 ítems, que luego se almacenaban en Google Cloud Storage. A pesar de los tiempos prolongados, el sistema fue capaz de procesar exitosamente todas las facturas solicitadas, dividiéndolas en los lotes establecidos.

El preprocesamiento se probó con diferentes volúmenes de facturas, y en todos los casos el sistema logró completar el proceso sin interrupciones ni errores, generando los archivos necesarios para el entrenamiento.

6.4.2. Entrenamiento Automático de Modelos

Tras el preprocesamiento, el sistema procedía al entrenamiento automático de los modelos. Se realizaron pruebas con diferentes cantidades de facturas: 1500, 900 y 300. En todos los casos, el tiempo de entrenamiento

Figura 6.4: Flujo de Preprocesamiento

no superó un minuto, demostrando la eficiencia del sistema al aplicar el *encoding* adecuado para todas las variables involucradas.

Además, una vez completado el entrenamiento, los modelos y sus *encoders* se almacenaban automáticamente, permitiendo su reutilización para futuras consultas de clasificación.

Figura 6.5: Flujo de Entrenamiento

6.4.3. Tiempo de Respuesta del Sistema de Consultas

El sistema de consultas, que permitía clasificar facturas utilizando los modelos entrenados, mostró tiempos de respuesta variables. En condiciones de *coldstart*, el tiempo máximo de respuesta fue de 15 segundos. Sin embargo, una vez que el sistema estaba activo, el tiempo promedio de respuesta se redujo a 2 segundos.

Este comportamiento fue consistente en todas las pruebas realizadas, lo que asegura la disponibilidad continua del sistema para consultas, incluso cuando se intercambiaban modelos entrenados sin interrupción del servicio.

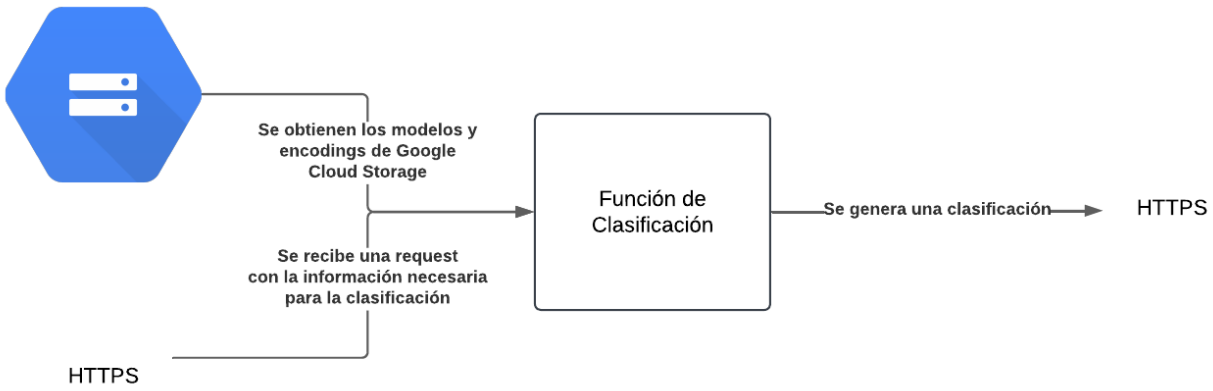

Figura 6.6: Flujo de Clasificación

CAPÍTULO 7

Discusión

Los resultados obtenidos en esta investigación proporcionan una base sólida para reflexionar sobre el potencial y los desafíos de la automatización de la clasificación de facturas mediante algoritmos de aprendizaje automático (ML) y procesamiento de lenguaje natural (NLP). La variedad de modelos implementados y evaluados no solo reveló diferencias significativas en cuanto a su desempeño técnico, sino también importantes limitaciones impuestas por el contexto organizacional en el que estas soluciones se aplican. En esta discusión, se profundiza en el análisis de los resultados obtenidos, así como en las reflexiones que emergen sobre el papel del aprendizaje automático en el procesamiento de documentos empresariales. Además, se aborda el impacto del análisis NLP y las lecciones derivadas de su implementación.

7.1. Desempeño Comparativo de los Modelos de Aprendizaje Automático

A lo largo de este trabajo, se evaluaron diversos modelos de aprendizaje automático con el objetivo de determinar cuál se adaptaba mejor a la clasificación automática de facturas empresariales. Los resultados indicaron claramente que el Árbol de Decisión y los Bosques Aleatorios sobresalieron como las opciones más efectivas, alcanzando exactitudes de 0.88 y 0.9, respectivamente. Estos hallazgos resaltan que, incluso en problemas complejos como la clasificación de facturas, donde las clases son heterogéneas y los datos están desbalanceados, estos enfoques tradicionales siguen siendo robustos y eficaces.

Una de las razones principales detrás del buen desempeño de estos modelos radica en la alta correlación entre el emisor de las facturas y su clasificación contable. Muchas de las empresas evaluadas tienden a operar dentro de los mismos sectores y, por tanto, emiten facturas relacionadas con productos o servicios similares. Esta relación directa facilita la identificación de patrones consistentes en los datos, haciendo más predecible la clasificación contable. De hecho, en la mayoría de los casos, el tipo de factura emitida refleja de manera clara la actividad comercial del emisor, lo que permite a los modelos capturar con precisión las regularidades en los datos.

La Regresión Logística, con su simplicidad y capacidad para identificar patrones lineales en los datos, es particularmente efectiva en este contexto, donde las características clave como el emisor y ciertos atributos

específicos de las facturas son predictores fuertes. Por otro lado, los Bosques Aleatorios, al combinar múltiples árboles de decisión, maximizan su capacidad de capturar relaciones no lineales, lo que les permite manejar mejor las variaciones entre clases y datos más ruidosos, brindando estabilidad y precisión adicional frente a escenarios desbalanceados.

En contraposición, el modelo de Máquinas de Soporte Vectorial (SVM) mostró un rendimiento muy inferior, con una exactitud de solo 0.1200. Este bajo desempeño se debe en gran medida a que el SVM no fue capaz de captar adecuadamente las correlaciones claras entre emisores y clasificaciones que dominaban el conjunto de datos. Además, su dificultad para manejar datos altamente dimensionales y desbalanceados complicó la optimización de los hiperplanos necesarios para una correcta clasificación. Este resultado pone de relieve que la sofisticación de un modelo no garantiza necesariamente un mejor rendimiento, y que es crucial seleccionar el algoritmo adecuado en función de las características del conjunto de datos y la naturaleza del problema.

El modelo de Regresión Logística, con una exactitud de 0.84, también fue una opción competitiva, aunque no tan efectiva como los Bosques Aleatorios. Su simplicidad y facilidad para interpretar las decisiones de clasificación lo convierten en una herramienta valiosa en entornos donde la transparencia es esencial, aunque, al igual que otros modelos individuales, no logró superar el desempeño del enfoque combinado de los Bosques Aleatorios.

Finalmente, los resultados de K-Nearest Neighbors (KNN) y Naive Bayes, con exactitudes de 0.64 y 0.83 respectivamente, aunque adecuados, revelan las limitaciones de estos enfoques en su capacidad de generalización y manejo de clases desbalanceadas. Aunque ofrecen una solución aceptable, no son tan efectivos en este contexto específico, donde la correlación entre el emisor y la clasificación es un factor clave, lo que favorece a modelos que puedan aprovechar mejor estas dependencias estructurales en los datos.

7.2. Desafíos de la Clasificación Multiclase en Datos Empresariales

Uno de los principales retos identificados durante la implementación de los modelos fue el desbalance de clases en el conjunto de datos. Las clases más representadas obtuvieron métricas notablemente altas en la mayoría de los modelos, lo que sugiere que una mayor cantidad de datos facilita que los algoritmos aprendan patrones claros y consistentes. Sin embargo, en clases con menos representaciones el desempeño fue considerablemente inferior. Este problema es inherente a los problemas de clasificación multiclase en los que ciertas categorías son minoritarias, y a menudo son más difíciles de predecir debido a la falta de información suficiente para que los modelos capturen sus características.

Este desafío resalta la importancia de utilizar técnicas para manejar el desbalance de clases, como el remuestreo o la ponderación de clases, estrategias que pueden ayudar a mitigar la falta de representatividad en clases minoritarias. Sin embargo, también pone en evidencia que, aunque estas técnicas son útiles, no son una solución definitiva, y los modelos pueden seguir enfrentando dificultades en estas situaciones.

Otro aspecto importante fue la variabilidad dentro de las clases. El hecho de que una misma clase pueda agrupar facturas de naturaleza diversa complica el proceso de clasificación, ya que el modelo debe ser capaz de identificar una amplia gama de patrones para predecir correctamente la clase de una factura. En este sentido, los modelos de aprendizaje automático, aunque efectivos en la mayoría de los casos, tienen sus limitaciones cuando los datos no contienen suficiente información estructurada o cuando existe una alta variabilidad dentro de las clases.

7.3. El Impacto y las Limitaciones del Procesamiento de Lenguaje Natural (NLP)

Uno de los esfuerzos más importantes de este trabajo fue la implementación de técnicas de procesamiento de lenguaje natural (NLP) para mejorar la capacidad de clasificación de las facturas. Sin embargo, los resultados obtenidos no fueron tan satisfactorios como se esperaba. A pesar de que el uso de modelos avanzados como GPT-3 permitió mejorar la exactitud de la clasificación en un 5 % para la compañía con menor rendimiento sin técnicas de NLP, este incremento resultó marginal frente a las expectativas iniciales y los recursos computacionales empleados.

La principal limitación del análisis NLP fue que la información contenida en las descripciones de las facturas, aunque útil, no fue suficiente para captar completamente el contexto necesario para una clasificación precisa. Este hallazgo reveló un punto crítico en el diseño del sistema: el contenido de las facturas no es la única ni la mejor fuente de información para determinar su clasificación. En realidad, el funcionamiento interno de la empresa —incluyendo el contexto organizacional, el departamento solicitante y las políticas específicas de la compañía— juega un papel fundamental en la correcta clasificación de las facturas.

De hecho, se descubrió que la misma factura podía tener clasificaciones diferentes dependiendo del departamento que la solicitaba, lo que no se podía deducir únicamente a partir de los datos presentes en la factura. Este tipo de información contextual no está explícita en los datos textuales y, por lo tanto, es inalcanzable incluso para modelos de NLP avanzados. Aunque las técnicas NLP permiten analizar y procesar texto no estructurado de manera eficiente, no pueden inferir el conocimiento tácito sobre los procesos internos de la empresa. Este descubrimiento subraya una limitación fundamental de los sistemas automatizados en la clasificación de documentos empresariales: sin un entendimiento profundo del contexto en el que se producen y gestionan estos documentos, la capacidad predictiva del sistema se ve considerablemente reducida.

7.4. Lecciones Aprendidas: El Valor del Contexto Organizacional

Una de las lecciones más valiosas de esta investigación es el reconocimiento de la importancia del contexto empresarial en la clasificación de facturas. A menudo, en proyectos de automatización, se asume que la calidad de los datos disponibles es el único factor determinante para el éxito de un modelo de aprendizaje automático. Sin embargo, en este caso, quedó claro que la comprensión de cómo opera la empresa es tan o más importante que los datos en sí mismos. La misma factura, con exactamente la misma información, puede tener diferentes clasificaciones dependiendo de quién la solicitó y con qué propósito. Este tipo de variabilidad organizacional es un factor clave que ningún algoritmo puede capturar sin acceso a información adicional sobre los procesos internos.

Esta revelación llevó a la decisión de no seguir forzando la inclusión de NLP en el sistema, ya que, si bien aportaba ciertas mejoras, no resolvía el problema fundamental: la falta de contexto sobre cómo y por qué se emiten las facturas dentro de la empresa. Este límite técnico y conceptual sugiere que, aunque la tecnología es una herramienta poderosa, no puede reemplazar el conocimiento humano especializado en el funcionamiento de una organización.

Por lo tanto, es fundamental que en futuras implementaciones se considere una integración más estrecha entre los sistemas automatizados y los conocimientos humanos. En lugar de depender completamente de los modelos de aprendizaje automático o de NLP, es probable que los mejores resultados se logren a través de sistemas híbridos, donde la inteligencia artificial se combine con la intervención humana para gestionar las situaciones más complejas que requieren un entendimiento profundo del contexto organizacional.

7.5. Impacto del Sistema Automático de Preprocesamiento y Entrenamiento

El sistema de preprocesamiento y entrenamiento automático implementado en este proyecto demostró ser una solución eficiente y escalable para manejar grandes volúmenes de datos. A pesar de los tiempos prolongados de procesamiento en algunos casos (hasta 15 minutos), el sistema fue capaz de gestionar adecuadamente la ingesta, el preprocesamiento y el almacenamiento de las facturas para su posterior análisis. Este enfoque no solo optimizó el manejo de los datos, sino que también proporcionó una arquitectura flexible que puede adaptarse a diferentes necesidades empresariales.

La automatización del entrenamiento de modelos y el almacenamiento de estos, junto con sus *encoders*, garantiza que los modelos puedan ser reutilizados y mejorados de manera continua sin intervención manual significativa. Este tipo de infraestructura, basada en *cloud functions* y almacenamiento en la nube, representa un paso hacia la creación de sistemas empresariales más eficientes y escalables, capaces de manejar grandes volúmenes de datos sin comprometer la precisión o la disponibilidad del sistema.

CAPÍTULO 8

Conclusiones

1. La automatización de la clasificación contable mediante inteligencia artificial es factible y puede transformar los procesos contables tradicionales. El desarrollo de un sistema basado en modelos de aprendizaje automático, como la Regresión Logística y los Bosques Aleatorios, ha demostrado ser altamente efectivo para asignar productos y servicios a las cuentas contables correspondientes, alcanzando una alta precisión en la clasificación. Esto cumple con el objetivo general de este trabajo, al mostrar que la inteligencia artificial puede aumentar significativamente la eficiencia operativa en los procesos contables al reducir la carga manual.

2. La correlación entre el emisor de las facturas y su clasificación contable es clave para el éxito del sistema. A lo largo del proyecto, se ha evidenciado que la principal razón por la cual los modelos de aprendizaje automático logran un alto desempeño es la fuerte relación entre el emisor de la factura y el tipo de productos o servicios ofrecidos. Las empresas suelen especializarse en tipos similares de productos, lo que facilita la predicción automática de la clasificación contable. Esta correlación reduce la complejidad del problema, lo que permite a los modelos funcionar con mayor precisión y eficiencia, cumpliendo con el objetivo específico de implementar tecnologías avanzadas en el clasificador.

3. El procesamiento de lenguaje natural (NLP) aporta mejoras marginales, pero está limitado por la falta de contexto organizacional. Aunque se implementaron técnicas de NLP para analizar las descripciones de las facturas, su impacto fue limitado debido a que no pueden capturar la información contextual relevante, como el departamento de la empresa que solicita la factura. Si bien las técnicas NLP mejoraron el rendimiento del modelo en un 5 %, esto no justifica su integración generalizada, debido a su limitada capacidad para comprender el funcionamiento específico de cada empresa. Este hallazgo destaca la importancia de una retroalimentación continua en pruebas piloto, cumpliendo parcialmente con el objetivo de evaluar y ajustar la efectividad del clasificador.

4. Es necesaria una mayor integración entre el conocimiento humano y los sistemas automatizados para garantizar un desempeño óptimo. Aunque los modelos de aprendizaje automático han demostrado ser efectivos para muchas empresas, la investigación ha revelado que algunos casos de clasificación requieren un entendimiento más profundo de los procesos internos de la empresa, información que no está contenida en las facturas. La intervención humana sigue siendo necesaria en situaciones donde el contexto organizacional es crucial, lo que indica que, para maximizar la eficiencia del sistema, este debe operar de manera híbrida, donde los sistemas automáticos sean complementados por el conocimiento humano.

5. El sistema tiene un potencial de escalabilidad y actualización, pero su éxito depende de la adaptabilidad a cambios organizacionales y legales. El diseño del sistema con un enfoque modular, capaz de

integrar nuevas fuentes de datos y actualizaciones continuas, asegura su escalabilidad a largo plazo. Sin embargo, para que este sistema se mantenga relevante y efectivo, es crucial que incorpore cambios en las legislaciones fiscales y se ajuste a las necesidades específicas de los usuarios empresariales. Esto cumplirá con el objetivo de garantizar un marco de actualización y adaptabilidad del clasificador, manteniendo su capacidad de evolución frente a los cambios regulatorios y del entorno empresarial.

CAPÍTULO 9

Recomendaciones

1. **Desarrollar un enfoque híbrido entre inteligencia artificial y conocimiento humano**: Si bien los modelos de aprendizaje automático y las técnicas de NLP demostraron ser eficaces en muchos casos, la naturaleza contextual de las empresas requiere que se considere un enfoque híbrido en el que la intervención humana complementa el sistema automático. Esto será particularmente útil para aquellas facturas que no puedan clasificarse adecuadamente sin un conocimiento profundo de los procesos internos de la empresa. Se recomienda la implementación de sistemas que permitan la intervención humana en casos complejos, asegurando así una mayor precisión global.

2. **Profundizar en técnicas de aprendizaje automático que manejen mejor la variabilidad organizacional**: Dado que la misma factura puede ser clasificada de manera diferente según el departamento que la solicite, se recomienda explorar modelos de aprendizaje automático que incorporen datos contextuales adicionales. Por ejemplo, el desarrollo de un sistema que tenga acceso a la estructura organizacional de la empresa podría ayudar a mejorar la clasificación en función de qué departamento está involucrado. Este enfoque podría incluir la integración de datos históricos del comportamiento de clasificación por departamento.

3. **Mejorar el sistema de retroalimentación y pruebas piloto con las empresas**: Para garantizar la efectividad y adaptabilidad del sistema, es crucial realizar pruebas continuas en entornos reales. Se recomienda establecer un marco de retroalimentación constante con las empresas usuarias para ajustar los modelos en tiempo real, basándose en cambios en las operaciones comerciales o la legislación. Este ciclo continuo de retroalimentación permitirá que el sistema siga siendo relevante y se adapte a las necesidades cambiantes del entorno empresarial.

4. **Optimizar el uso de técnicas NLP en dominios específicos**: Aunque el análisis de lenguaje natural tuvo un impacto limitado en este caso, se recomienda seguir investigando en técnicas NLP especializadas para sectores empresariales concretos. En algunos casos, la integración de NLP mejorada y optimizada para industrias específicas podría tener un mayor impacto en la precisión del sistema. Esto podría incluir modelos de lenguaje entrenados específicamente en vocabularios contables y financieros que mejoren la clasificación automática de facturas en contextos muy particulares.

5. **Asegurar la actualización continua del sistema frente a cambios legislativos y tecnológicos**: Dado el dinamismo del entorno regulatorio y fiscal, se recomienda establecer un mecanismo automático de actualización del sistema que incorpore cambios en la legislación de manera fluida y sin interrupciones en el servicio. También es importante mantener la infraestructura tecnológica actualizada, aprovechando nuevas técnicas de aprendizaje automático y NLP a medida que se desarrollen, para asegurar que el sistema permanezca en la vanguardia de la innovación tecnológica.

Bibliografía

[1] *¿Qué es Python? - Explicación del lenguaje Python - AWS*, 2022. https://aws.amazon.com/es/what-is/python/, Recuperado de Amazon Web Services, Inc.

[2] *Cloud Firestore | Firebase Documentation*, 2023. https://firebase.google.com/docs/firestore, Recuperado de Google Developers.

[3] *What is a Non-Relational Database?*, 2023. https://www.mongodb.com/databases/non-relational, Recuperado de MongoDB.

[4] *Cloud Functions | Google Cloud*, 2024. https://cloud.google.com/functions?hl=es_419, Recuperado de Google Cloud.

[5] *Descripción general de Google Cloud*, 2024. https://cloud.google.com/docs/overview?hl=es-419, Recuperado de Google Cloud.

[6] *Documentación de App Engine | App Engine Documentation | Google Cloud*, 2024. https://cloud.google.com/appengine/docs?hl=es-419, Recuperado de Google Cloud.

[7] Academy, Verne: *10 Librerías Python para Data Science y Machine Learning*, 2022. https://verneacademy.com/blog/articulos-ia/10-librerias-python-data-science-machine-learning/, Publicado el 12 de septiembre de 2022.

[8] Aggarwal, Charu y Chengxiang Zhai: *Mining Text Data*. Enero 2012.

[9] Alcarria Jaime, J. J.: *Contabilidad financiera I*. Publicacions de la Universitat Jaume I, España, 2008. https://www.google.com.gt/books/edition/Contabilidad_financiera_I/6m42LTDkhzoC?hl=es&gbpv=1&dq=contabilidad+pdf&printsec=frontcover.

[10] Appvizer.es: *¿Qué son el debe y haber en contabilidad ? Diferencias, ejemplos y más*, 2024. https://www.appvizer.es/revista/contabilidad-finanzas/contabilidad/debe-y-haber, Publicado el 7 de junio de 2024.

[11] ASALE, R. y RAE: *Diccionario de la lengua española RAE - ASALE*. 2023. https://dle.rae.es/pasivo?m=form, Edición del Tricentenario.

[12] Bell, P. y B. Beer: *Introducing GitHub: A Non-Technical Guide*. O'Reilly Media, 2nd edición, 2018.

[13] Bishop, Christopher M.: *Pattern Recognition and Machine Learning*. Springer Science+Business Media, New York, NY, 2006, ISBN 978-0387-31073-2. https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

[14] Chacon, Scott y Ben Straub: *Pro Git*. Apress, Springer Nature, New York, NY, 2014, ISBN 9781484200766. http://library.oapen.org/handle/20.500.12657/28155.

[15] Colegio de Contadores Públicos y Auditores de Guatemala: *Adopción de la Norma Internacional de Información Financiera para Pequeñas y Medianas Entidades*. Diario de Centro América, 2007.

[16] Devlin, Jacob, Ming Wei Chang, Kenton Lee y Kristina Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. En *North American Chapter of the Association for Computational Linguistics*, 2019. https://api.semanticscholar.org/CorpusID:52967399.

[17] Elmasri, R. y S. B. Navathe: *Fundamentals of Database Systems*. Pearson, 7th edición, 2016.

[18] Erl, T., Z. Mahmood y R. Puttini: *Cloud Computing: Concepts, Technology & Architecture*. Prentice Hall, 2013.

[19] Face, Hugging: *dccuchile/bert-base-spanish-wwm-cased*, 2019. https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased, Accedido: 18 de octubre de 2024.

[20] GitHub Docs: *Acerca de la integración continua con Acciones de GitHub*, 2024. https://docs.github.com/es/actions/about-github-actions/about-continuous-integration-with-github-actions, GitHub Docs.

[21] Google Cloud: *Secret Manager*, 2024. https://cloud.google.com/secret-manager?hl=es, Google Cloud.

[22] Hennessy, J. L. y D. A. Patterson: *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann, 6th edición, 2019.

[23] iAhorro: *iAhorro*, 2024. https://www.iahorro.com/diccionario/g/gastos, Recuperado de iAhorro.

[24] IBM: *¿Qué es un algoritmo de machine learning?*, 2024. https://www.ibm.com/es-es/topics/machine-learning-algorithms#:~:text=Un%20algoritmo%20de%20machine%20learning%20o%20machine%20learning%20es%20un,determinado%20de%20variables%20de%20entrada, Recuperado el 22 de enero de 2024.

[25] Jurafsky, Daniel y James H. Martin: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Stanford University and University of Colorado at Boulder, 3rd edición, 2024. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf, Draft of August 20, 2024.

[26] LeCun, Yann, Y. Bengio y Geoffrey Hinton: *Deep Learning*. Nature, 521:436–44, Mayo 2015.

[27] Moroney, L.: *The Definitive Guide to Firebase: Build Android Apps on Google's Mobile Platform*. Apress, 2017.

[28] Paar, Christof y Jan Pelzl: *Understanding Cryptography: A Textbook for Students and Practitioners*. Springer, Berlin, Heidelberg, 2010. https://doi.org/10.1007/978-3-642-04101-3.

[29] QuickBooks: *Cuentas contables: qué son y cómo se clasifican - QuickBooks*, sep 2024. https://quickbooks.intuit.com/global/resources/es/contabilidad/clasificacion-de-cuentas-contables/, Accedido: 18 de octubre de 2024.

[30] Russell, Stuart J. y Peter Norvig: *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, 3rd edición, 2016.

[31] Saavedra, Guillermo González: *Contabilidad general*. Universidad Politécnica de Gómez Palacio, 2003. https://www.upg.mx/wp-content/uploads/2015/10/LIBRO-37-Contabilidad-General.pdf.

[32] Sadalage, P. J. y M. Fowler: *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley Professional, 2012.

[33] Shafiei, H., A. Khonsari y P. Mousavi: *Serverless Computing: A Survey of Opportunities, Challenges, and Applications*. ACM Computing Surveys, 54(11s):239, 2022. `https://ar5iv.labs.arxiv.org/html/1911.01296`.

[34] Stallings, William y Lawrie Brown: *Computer Security: Principles and Practice*. Pearson Education, Upper Saddle River, NJ, USA, 3rd edición, 2014, ISBN 978-0-13-377392-7.

[35] Superintendencia de Administración Tributaria: *Sistema de impuestos y declaraciones*. Superintendencia de Administración Tributaria, n.d. Guatemala.

[36] Sánchez, Alejandro Donoso: *Factura - Qué es y su Importancia en el Comercio*, May 2017. `https://economipedia.com/definiciones/factura.html`, Economipedia.

[37] Vaish, G.: *Getting Started with NoSQL*. Packt Publishing, 2013.

[38] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser y Illia Polosukhin: *Attention Is All You Need*. En *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA, 2017. `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Anexos

Trabajo Final de Graduación 2024. Juan Carlos Bajan. Disponible en: `https://github.com/JuanCarlosBajan/Trabajo-Final-de-Graduacion-2024`

Clase	Bosque Aleatorio	Árbol de Decisión	Naïve Bayes	Regresión Logística	KNN	SVM
3	0.000	0.250	0.200	0.000	0.000	0.000
20	1.000	1.000	1.000	0.000	0.000	0.000
29	1.000	1.000	1.000	0.000	1.000	0.000
34	1.000	1.000	1.000	0.000	0.000	0.000
44	1.000	0.333	0.500	1.000	0.200	0.000
59	0.000	0.000	0.000	0.000	0.000	0.000
61	1.000	0.833	0.625	0.500	0.833	0.000
71	1.000	1.000	1.000	0.000	0.000	0.000
86	0.933	0.933	0.933	0.875	0.520	0.000
106	1.000	1.000	1.000	0.800	0.400	0.000
109	1.000	1.000	1.000	0.000	1.000	0.000
119	0.920	0.920	1.000	0.904	0.674	0.000
122	1.000	1.000	1.000	0.000	1.000	0.000
125	0.900	1.000	0.643	0.750	0.571	0.000
138	0.000	0.000	0.000	0.000	0.000	0.000
147	0.828	0.767	1.000	0.533	0.419	0.000
154	1.000	1.000	0.800	0.000	0.000	0.000
168	0.706	0.632	0.818	0.563	0.467	0.000
183	0.778	0.778	0.875	0.778	0.636	0.000
186	0.692	0.686	0.660	0.682	0.375	0.000
207	1.000	1.000	1.000	0.000	0.000	0.000
208	1.000	1.000	1.000	1.000	0.000	0.000
241	1.000	1.000	0.500	0.000	0.250	0.000
283	0.000	0.000	0.000	0.000	0.000	0.000
286	1.000	0.000	0.333	0.000	0.000	0.000
294	1.000	1.000	1.000	0.000	0.667	0.000
308	1.000	1.000	0.429	1.000	0.000	0.000
335	1.000	1.000	1.000	1.000	1.000	0.000
350	1.000	1.000	1.000	0.982	0.846	0.000
369	1.000	0.977	1.000	1.000	0.842	0.000
375	1.000	1.000	1.000	1.000	0.435	0.000
393	0.000	0.000	0.000	0.000	0.000	0.000
397	1.000	0.667	0.400	0.750	1.000	0.000
412	0.000	0.000	0.000	0.000	0.000	0.000
417	0.909	0.765	1.000	0.625	0.714	0.000
424	1.000	1.000	1.000	1.000	0.143	0.000
431	1.000	1.000	0.286	0.200	0.000	0.000
458	0.933	0.875	0.667	0.929	0.273	0.000
460	0.167	0.125	0.154	0.667	1.000	0.000
473	0.000	0.000	0.000	0.000	0.000	0.000
Total	0.769	0.714	0.671	0.438	0.382	0.000

Tabla 1: Comparación de Precisión entre modelos de clasificación para muestra significativa

Clase	Bosque Aleatorio	Árbol de Decisión	Naïve Bayes	Regresión Logística	KNN	SVM
3	0.000	0.167	0.167	0.000	0.000	0.000
20	0.500	0.500	1.000	0.000	0.000	0.000
29	1.000	1.000	1.000	0.000	1.000	0.000
34	1.000	1.000	1.000	0.000	0.000	0.000
44	1.000	1.000	1.000	1.000	0.500	0.000
59	0.000	0.000	0.000	0.000	0.000	0.000
61	0.833	0.833	0.833	0.333	0.833	0.000
71	1.000	1.000	1.000	0.000	0.000	0.000
86	1.000	1.000	1.000	1.000	0.929	0.000
106	1.000	1.000	1.000	1.000	0.500	0.000
109	1.000	1.000	1.000	0.000	1.000	0.000
119	0.990	0.990	0.705	0.981	0.848	0.000
122	1.000	1.000	1.000	0.000	1.000	0.000
125	1.000	1.000	1.000	1.000	0.444	0.000
138	0.000	0.000	0.000	0.000	0.000	0.000
147	0.828	0.793	0.724	0.828	0.448	0.000
154	0.500	0.500	1.000	0.000	0.000	0.000
168	0.857	0.857	0.643	0.643	0.500	0.000
183	1.000	1.000	1.000	1.000	1.000	0.000
186	0.771	0.686	0.886	0.857	0.343	0.000
207	1.000	1.000	1.000	0.000	0.000	0.000
208	1.000	1.000	1.000	1.000	0.000	0.000
241	1.000	1.000	1.000	0.000	1.000	0.000
283	0.000	0.000	0.000	0.000	0.000	0.000
286	1.000	0.000	1.000	0.000	0.000	0.000
294	1.000	1.000	1.000	0.000	1.000	0.000
308	0.667	0.667	1.000	0.500	0.000	0.000
335	1.000	1.000	1.000	1.000	0.667	0.000
350	1.000	1.000	1.000	0.965	0.965	0.000
369	1.000	1.000	1.000	1.000	0.744	0.000
375	1.000	1.000	1.000	1.000	0.625	0.000
393	0.000	0.000	0.000	0.000	0.000	0.000
397	0.750	0.500	0.750	0.750	0.125	0.000
412	0.000	0.000	0.000	0.000	0.000	0.000
417	0.714	0.929	0.500	0.357	0.357	0.000
424	1.000	1.000	1.000	0.800	0.200	0.000
431	1.000	1.000	1.000	0.500	0.000	0.000
458	1.000	1.000	1.000	0.929	0.214	0.000
460	0.200	0.200	0.800	0.400	0.200	0.000
473	0.000	0.000	0.000	0.000	0.000	0.000
Total	0.740	0.716	0.775	0.446	0.386	0.000

Tabla 2: Comparación de Recall entre modelos de clasificación para muestra significativa

Clase	Bosque Aleatorio	Árbol de Decisión	Naïve Bayes	Regresión Logística	KNN	SVM
3	0.000	0.200	0.182	0.000	0.000	0.000
20	0.667	0.667	1.000	0.000	0.000	0.000
29	1.000	1.000	1.000	0.000	1.000	0.000
34	1.000	1.000	1.000	0.000	0.000	0.000
44	1.000	0.500	0.667	1.000	0.286	0.000
59	0.000	0.000	0.000	0.000	0.000	0.000
61	0.909	0.833	0.714	0.400	0.833	0.000
71	1.000	1.000	1.000	0.000	0.000	0.000
86	0.966	0.966	0.966	0.933	0.667	0.000
106	1.000	1.000	1.000	0.889	0.444	0.000
109	1.000	1.000	1.000	0.000	1.000	0.000
119	0.954	0.954	0.827	0.941	0.751	0.000
122	1.000	1.000	1.000	0.000	1.000	0.000
125	0.947	1.000	0.783	0.857	0.500	0.000
138	0.000	0.000	0.000	0.000	0.000	0.000
147	0.828	0.780	0.840	0.649	0.433	0.000
154	0.667	0.667	0.889	0.000	0.000	0.000
168	0.774	0.727	0.720	0.600	0.483	0.000
183	0.875	0.875	0.933	0.875	0.778	0.000
186	0.730	0.686	0.756	0.759	0.358	0.000
207	1.000	1.000	1.000	0.000	0.000	0.000
208	1.000	1.000	1.000	1.000	0.000	0.000
241	1.000	1.000	0.667	0.000	0.400	0.000
283	0.000	0.000	0.000	0.000	0.000	0.000
286	1.000	0.000	0.500	0.000	0.000	0.000
294	1.000	1.000	1.000	0.000	0.800	0.000
308	0.800	0.800	0.600	0.667	0.000	0.000
335	1.000	1.000	1.000	1.000	0.800	0.000
350	1.000	1.000	1.000	0.973	0.902	0.000
369	1.000	0.989	1.000	1.000	0.790	0.000
375	1.000	1.000	1.000	1.000	0.513	0.000
393	0.000	0.000	0.000	0.000	0.000	0.000
397	0.857	0.571	0.522	0.750	0.222	0.000
412	0.000	0.000	0.000	0.000	0.000	0.000
417	0.800	0.839	0.667	0.455	0.476	0.000
424	1.000	1.000	1.000	0.889	0.167	0.000
431	1.000	1.000	0.444	0.286	0.000	0.000
458	0.966	0.933	0.800	0.929	0.240	0.000
460	0.182	0.154	0.258	0.500	0.333	0.000
473	0.000	0.000	0.000	0.000	0.000	0.000
Total	0.748	0.703	0.693	0.434	0.354	0.000

Tabla 3: Comparación de F1-Score entre modelos de clasificación para muestra significativa

Class	**Precision**	**Recall**	**F1-score**	**Support**
7758	1.00	0.94	0.97	18
7762	0.57	0.59	0.58	68
7763	0.50	0.38	0.43	13
7765	0.85	0.70	0.77	50
7771	0.00	0.00	0.00	10
7800	0.55	0.71	0.62	24
7801	0.53	0.71	0.61	14
7803	0.00	0.00	0.00	6
7804	0.51	0.56	0.53	68
7821	0.67	0.50	0.57	4
7825	0.12	0.17	0.14	6
7830	0.59	0.67	0.62	15
7833	0.71	0.50	0.59	20
57663	0.44	0.80	0.57	10
57671	1.00	1.00	1.00	4

Tabla 4: Resultados del modelo de clasificación Bosques Aleatorios sin técnicas de NLP

Class	**Precision**	**Recall**	**F1-score**	**Support**
7758	1.00	1.00	1.00	18
7762	0.60	0.63	0.61	68
7763	0.43	0.23	0.30	13
7765	0.93	0.76	0.84	50
7771	0.33	0.10	0.15	10
7800	0.75	0.88	0.81	24
7801	0.50	0.86	0.63	14
7803	1.00	0.17	0.29	6
7804	0.56	0.59	0.58	68
7821	0.80	1.00	0.89	4
7825	0.29	0.33	0.31	6
7830	0.60	0.80	0.69	15
7833	0.80	0.40	0.53	20
57663	0.50	0.90	0.64	10
57671	0.80	1.00	0.89	4

Tabla 5: Resultados del modelo de clasificación Bosques Aleatorios con técnicas de NLP