Detección automatizada de anomalías en contratos públicos del Estado peruano mediante técnicas de aprendizaje automático

Bravo Lazaro, Beatriz
Universidad ESAN
Lima, Perú
22200118@ue.edu.pe

Garay Santos, Jefferson
Universidad ESAN
Lima, Perú
22200127@ue.edu.pe

***Abstract*—Corruption in public procurement represents one of the most critical challenges for government management in Peru, generating substantial economic losses and hindering oversight and control processes. Manual auditing of hundreds of thousands of contracts and administrative documents is insufficient to identify irregular patterns in a timely manner. This study proposes an automated anomaly detection system for Peruvian public procurement contracts by integrating advanced machine learning techniques and natural language processing. A consolidated dataset of 365,341 contracts from SEACE was constructed and enriched with an additional dataset of administrative penalties, enabling the creation of derived variables such as penalty frequency and the binary label *anomaly*.**

The methodological framework consists of two processing routes: structured data and unstructured data. For structured data, procedures such as data cleaning, monetary standardization, and feature engineering were applied. For unstructured data, contractual documents were extracted through OCR, and semantic embeddings were generated using the *paraphrase-multilingual-MiniLM-L12-v2* model, complemented by thematic clustering with K-Means. The final integrated dataset was used to train supervised models including Logistic Regression, KNN, Decision Tree, AdaBoost, and neural networks, which were evaluated through precision, recall, and F1-score metrics.

The results demonstrate that the proposed approach effectively identifies patterns associated with contractual risks and improves the early detection of irregular behaviors. This framework provides a valuable contribution to digital auditing and strengthens transparency and accountability within Peru's public procurement system.

***Index Terms*—public procurement, anomaly detection, machine learning, natural language processing, OCR, semantic embeddings.**

I. Introducción

La corrupción en las contrataciones públicas representa uno de los problemas más graves para el desarrollo económico y social del Perú. Como señala Martínez Huamán (2023), la corrupción ha sido percibida como el principal problema del país entre 2019 y 2022, superando ampliamente a la delincuencia, la pobreza y la falta de empleo [13]. Esta situación se agrava por la inefectividad del sistema penal, evidenciada en la baja tasa de condenas efectivas, que alcanzó solo el 0.0031% en 2013, a pesar del incremento del 63% en denuncias policiales por corrupción entre 2018 y 2021.

El impacto económico de esta problemática es considerable. La Contraloría General de la República estimó que, solo en 2023, las pérdidas derivadas de la corrupción ascendieron a aproximadamente S/ 24,268 millones [15]. Casos emblemáticos como el de Odebrecht, donde los tramos 2 y 3 de la carretera Interoceánica Sur generaron un daño patrimonial de alrededor de S/ 1,400 millones, ilustran la magnitud del problema [13]. Asimismo, se registran 367 obras del sector educación paralizadas, valorizadas en S/ 1,932 millones, afectando directamente el derecho fundamental a la educación [14].

Desde una perspectiva técnica, la detección manual de irregularidades en el vasto volumen de contrataciones públicas resulta ineficiente e insuficiente. Según el mapeo sistemático de Santos et al. (2025), los métodos tradicionales de auditoría solo logran examinar un porcentaje mínimo de los contratos, mientras que las técnicas basadas en aprendizaje automático permiten analizar conjuntos completos de datos e identificar patrones complejos que escapan al análisis humano convencional [19]. Esta capacidad es crucial, ya que —como demostraron Aldana et al. (2024)— las relaciones sistemáticas entre compradores y proveedores son mejores predictores de corrupción que las características individuales de los contratos, patrones que solo pueden ser detectados mediante algoritmos avanzados de *machine learning*.

Un desafío clave radica en la naturaleza no estructurada de la información relevante. Mientras las bases de datos tabulares contienen información básica de las contrataciones, los detalles críticos —como cláusulas específicas, modificatorias, anexos técnicos y términos de referencia— se encuentran usualmente en documentos PDF que escapan al análisis automatizado tradicional. Kim et al. (2022), en su estudio sobre el modelo Donut, señalan que la dependencia de sistemas OCR convencionales introduce errores de segmentación y transcripción, limitando la extracción confiable de información contractual.

Frente a este escenario, las técnicas de aprendizaje automático surgen como una herramienta prometedora para fortalecer los sistemas de control y detección temprana de irregularidades. Como reportó la Conferencia Anual Internacional por la Integridad (CAII, 2023), el uso de inteligencia artificial

permitió identificar a 2,200 proveedores que contrataron con el Estado por S/ 584 millones pese a estar impedidos de hacerlo [16]. Este hallazgo demuestra el potencial transformador de estas tecnologías en la lucha contra la corrupción.

En este contexto, el presente estudio empleará los datos abiertos proporcionados por el Gobierno del Perú a través del Organismo Supervisor de las Contrataciones del Estado (OSCE), aplicando técnicas avanzadas de procesamiento de datos y modelos de *machine learning* para identificar patrones sospechosos y posibles casos de corrupción. Esta aproximación permitirá optimizar los recursos destinados a la auditoría y control, además de generar alertas tempranas que prevengan el desvío de recursos públicos.

La principal contribución de esta investigación radica en adaptar y validar metodologías de vanguardia en el contexto específico del sistema de contrataciones peruano, considerando sus particularidades normativas y operativas. Los resultados esperados incluyen la identificación de variables predictoras de corrupción relevantes para el caso peruano, así como el desarrollo de un modelo con alta capacidad de detección de anomalías.

La principal contribución de esta investigación radica en adaptar y validar metodologías de vanguardia en el contexto específico del sistema de contrataciones peruano, considerando sus particularidades normativas y operativas. Los resultados esperados incluyen la identificación de variables predictoras de corrupción relevantes para el caso peruano, así como el desarrollo de un modelo con alta capacidad de detección de anomalías.

No obstante, este estudio busca ir más allá del análisis tradicional de datos estructurados,ya que se plantea la incorporación de técnicas de Reconocimiento Óptico de Caracteres (OCR) y minería de texto aplicadas a los documentos contractuales disponibles en los portales institucionales del Estado, con el fin de extraer información contenida en cláusulas, adendas y anexos técnicos. Este enfoque permitirá complementar las variables tabulares con representaciones semánticas del contenido contractual, posibilitando la detección de patrones discursivos o terminológicos asociados a posibles irregularidades.

De esta manera, la investigación propone una integración innovadora entre datos estructurados y no estructurados mediante un pipeline de aprendizaje automático, que incluye el procesamiento de texto, la vectorización semántica y la detección de anomalías a nivel contractual. Con ello, se busca contribuir no solo al avance metodológico en la aplicación de machine learning para el control público, sino también al fortalecimiento de la transparencia y la eficiencia en la gestión estatal del Perú.

II. Antecedentes

En los últimos años, el uso de técnicas de *machine learning* se ha extendido en el análisis de datos públicos, especialmente para detectar irregularidades y patrones anómalos en procesos administrativos. Estas herramientas permiten identificar comportamientos atípicos y posibles indicios de corrupción en la adjudicación de contratos estatales.

A pesar de los avances logrados, aún existen retos relacionados con la calidad de los datos, la interpretabilidad de los modelos y su integración en los sistemas de control público. En este apartado se presentan los antecedentes más relevantes que sustentan la presente investigación.

A machine learning model to identify corruption in México's public procurement contracts

El trabajo titulado *"A machine learning model to identify corruption in Mexico's public procurement contracts"* [1], presentado por Andres Aldana, Andrea Falcon-Corte y Hernan Larralde, fue publicado en marzo de 2024, y surge de la motivación por abordar el problema de la corrupción en las contrataciones públicas, la cual genera pérdidas económicas sustanciales (se estima que casi el **2%** del presupuesto federal mexicano se pierde en prácticas corruptas) y afecta el bienestar ciudadano, la competencia de mercado y la distribución de recursos.

Los autores plantean que la **detección automatizada de corrupción** en contrataciones públicas es crucial debido a la complejidad de las transacciones, la participación de múltiples actores y la dificultad de identificar patrones corruptos mediante métodos tradicionales, su hipótesis central sostiene que las relaciones sistemáticas entre compradores y proveedores constituyen mejores predictores de corrupción que las características individuales de los contratos.

Para validar esta hipótesis, los investigadores desarrollaron un enfoque de aprendizaje automático denominado **HyperForest**, que consiste en un **ensamble de clasificadores Random Forest** combinado con técnicas especializadas para manejar el desbalanceo extremo de clases (proporción 45:1 entre contratos no corruptos y corruptos).

Base de Datos

Para este estudio, los autores utilizaron el dataset de **CompraNet - Contratos Públicos** [2], que contiene aproximadamente **1.6 millones de contratos** públicos mexicanos del período 2013-2020. Tras un proceso de tratamiento de datos que incluyó la homogenización de variables categóricas y eliminación de registros incompletos, el conjunto final quedó compuesto por **1.54 millones de contratos** válidos. Adicionalmente, integraron los listados oficiales de proveedores sancionados de la **Secretaría de Administración Tributaria (SAT)** [3] y del portal de **Datos Abiertos Mexicanos** [4] para etiquetar los contratos, resultando en **33,494 contratos etiquetados como corruptos** y **1,506,892 como no corruptos**. Por lo que, cada instancia fue caracterizada mediante **19 variables** organizadas en cuatro tipos: características individuales del contrato, relaciones comprador-proveedor, características del comprador y factores de riesgo de corrupción.

Metodología

Preparación de Datos

El estudio utilizó el mismo conjunto de datos analizado previamente en Falcon-Cortes et al. (2022) [5], con la modificación de que se mantuvieron las variables categóricas en su

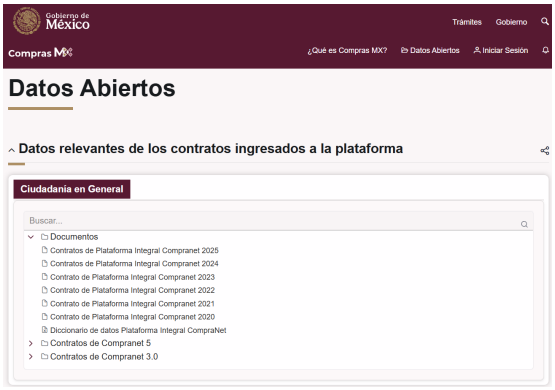

Fig. 1: Página web de CompraNet – Portal de Contrataciones Públicas del Gobierno Mexicano.

Fig. 2: Página web del SAT – Portal de Datos Abiertos de la Secretaría de Administración Tributaria.

formato original en lugar de convertirlas en variables dummy. Los datos se obtuvieron del sistema electrónico mexicano de información gubernamental **CompraNet** ([2]), que cubre todos los contratos públicos de 2013 a 2020.

1) Procesamiento de Datos:

- **Limpieza inicial**: De los 1.6 millones de contratos originales, se eliminaron aproximadamente 60,000 registros (4% del total) que presentaban variables importantes omitidas.
- **Homogenización**: Estandarización de todas las variables de texto para evitar problemas con caracteres especiales o espaciado.
- **Etiquetado**: Los contratos se etiquetaron como corruptos (**C**) si el proveedor aparecía en las listas oficiales de empresas sancionadas, y como no corruptos (**NC**) en caso contrario.

Variables Utilizadas

El conjunto final incluyó **19 variables** organizadas en cuatro tipos:

- **Tipo i) Características del contrato**: 5 variables categóricas y 4 numéricas que describen propiedades individuales del contrato.
- **Tipo ii) Relación comprador-proveedor**: 4 variables que cuantifican la relación entre las partes.
- **Tipo iii) Características del comprador**: 2 variables descriptivas de los compradores.
- **Tipo iv) Factores de riesgo**: 4 variables que aproximan los factores de riesgo de corrupción propuestos en la literatura.

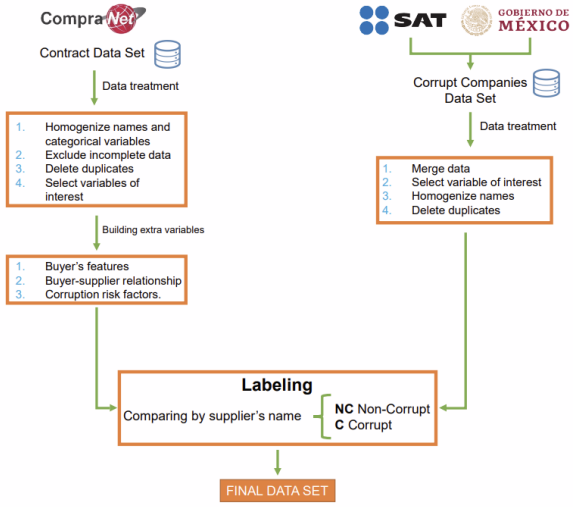

Fig. 3: Diagrama del proceso de preparación y etiquetado de datos.

Preparación y Balanceo de Datos

Para garantizar una evaluación robusta del modelo, se implementó una estrategia integral de particionamiento y balanceo de datos. El dataset completo se dividió en tres conjuntos independientes: **Conjunto de Entrenamiento** (50% de los datos), **Conjunto de Calibración** (20%) y **Conjunto de Prueba** (30%). Esta separación permitió un entrenamiento adecuado, una calibración objetiva de los parámetros del modelo y una evaluación final imparcial del rendimiento predictivo.

Dado el desbalance extremo en los datos, con una proporción de 45:1 entre contratos no corruptos (NC) y corruptos (C), se implementó la técnica de **Repeated Random Sub-sampling** para mitigar los sesgos en el entrenamiento. Específicamente, se generaron $|NC|/|C| = 45$ submuestras balanceadas, donde cada una contenía la totalidad de los contratos corruptos disponibles en el conjunto de entrenamiento y una cantidad igual de contratos no corruptos, seleccionados aleatoriamente sin reemplazo. Este enfoque permitió crear múltiples conjuntos de entrenamiento equilibrados, facilitando que el modelo aprendiera patrones representativos de ambas clases sin favorecer artificialmente la clase mayoritaria.

Arquitectura y Funcionamiento del Modelo Hyper-Forest

El modelo **Hyper-Forest** consiste en un sistema de votación basado en múltiples algoritmos de Random Forest que trabajan en conjunto. La arquitectura comprende **45 Random Forests independientes**, cada uno entrenado con una de las submuestras balanceadas obtenidas mediante el proceso de

Repeated Random Sub-sampling. Cada Random Forest se configuró con parámetros estándar: **500 árboles de decisión** por forest, selección de $\sqrt{p}$ características en cada división (donde p representa el número total de características), y nodos terminales con tamaño mínimo de 1.

El proceso de clasificación sigue tres pasos fundamentales:

1) **Votación individual**: Cada uno de los 45 Random Forests analiza el contrato y emite su voto sobre si corresponde a la clase No Corrupto (NC) o Corrupto (C)
2) **Cálculo de probabilidad**: Se determina la probabilidad de que un contrato sea no corrupto mediante la fórmula $\mathrm{P}(NC|x) = V(NC)/T$, donde $V(NC)$ representa el número de votos a favor de NC y $T = 45$ el total de Random Forests
3) **Decisión final**: Se aplica un umbral de clasificación θ donde si $\mathrm{P}(NC|x) > \theta$, el contrato se clasifica como no corrupto; en caso contrario, se clasifica como corrupto

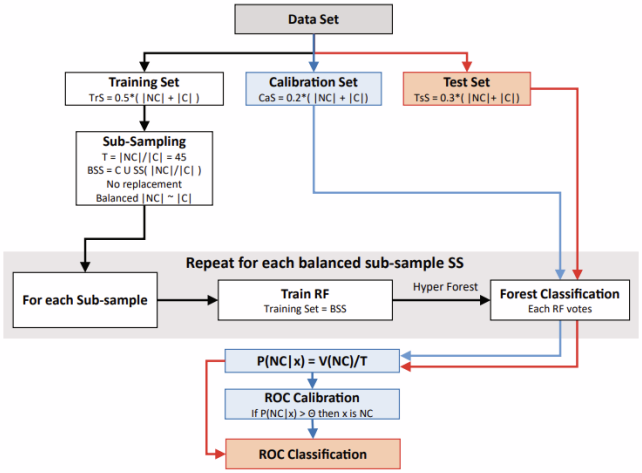

Fig. 4: Diagrama del flujo de entrenamiento, calibración y evaluación del modelo de Random Forest para la clasificación de contratos públicos.

Para determinar el valor óptimo del umbral θ, se utilizó el **Conjunto de Calibración** junto con la técnica de **Curva ROC**, que permite visualizar el balance entre la capacidad de detectar verdaderos no corruptos (TPR) y el riesgo de clasificar incorrectamente contratos corruptos como no corruptos (FPR). El análisis identificó $\theta = 0.61$ como el punto óptimo que maximiza ambos objetivos.

Finalmente, se implementó **Recursive Feature Elimination (RFE)** para optimizar el conjunto de variables predictoras. Esta técnica elimina iterativamente la variable menos importante según el análisis de importancia del Random Forest, reentrena el modelo con las variables restantes, y evalúa el rendimiento en cada etapa. El proceso continúa hasta agotar todas las variables, seleccionando finalmente el subconjunto que produce el mejor desempeño predictivo.

Resultados

El modelo Hyper-Forest demostró un rendimiento sobresaliente en la identificación de contratos corruptos, alcanzando una **precisión balanceada del 91%** tras aplicar selección de características recursiva. El análisis detallado por clase reveló que el modelo logra una **exactitud del 94%** en contratos no corruptos (NC) y del **88%** en contratos corruptos (C), con un **área bajo la curva ROC (AUC) de 0.94**, indicando una capacidad discriminativa excelente.

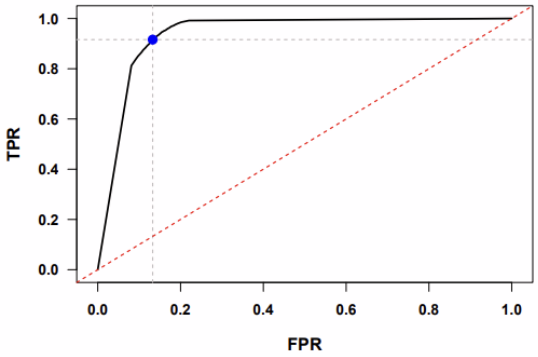

Fig. 5: Curva ROC del modelo Hyper-Forest, mostrando el umbral óptimo de clasificación (punto azul) con TPR = 0.92 y FPR = 0.1

La calibración óptima del modelo se obtuvo con un umbral $\theta = 0.61$, punto donde se consigue una **tasa de verdaderos positivos del 92%** (contratos NC correctamente identificados) y una **tasa de falsos positivos del 13%** (contratos C erróneamente clasificados como NC).

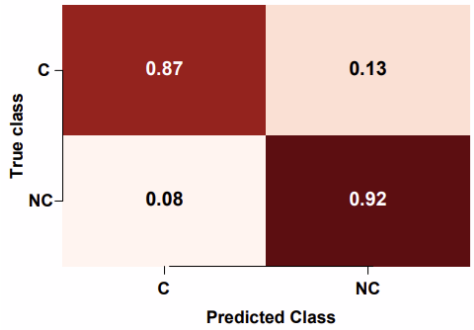

Fig. 6: Matriz de confusión del modelo Hyper-Forest sobre el conjunto de prueba.

En términos prácticos, la implementación del modelo representaría una **reducción del 90% en la carga investigativa**, ya que al aplicarse sobre 1,000 contratos nuevos, solo **97 serían señalados como sospechosos**, de los cuales **19 corresponderían a verdaderos casos de corrupción**, con apenas **3 contratos corruptos no detectados**. Estos resultados evidencian que el Hyper-Forest constituye una herramienta efectiva para priorizar esfuerzos investigativos en la detección de corrupción en contrataciones públicas, optimizando significativamente la asignación de recursos auditores.

Predicción de la corrupción vía red neuronal

El trabajo titulado *"Predicción de la corrupción vía red neuronal"* [7], presentado por Roberto Sánchez Fernández bajo la tutoría de David Renato Domínguez Carreta en la Universidad Autónoma de Madrid, fue desarrollado como un Trabajo Fin de Grado en 2017. Surge de la necesidad de analizar la evolución de la corrupción a nivel mundial y predecir su comportamiento futuro a partir de técnicas de **aprendizaje automático**.

El autor plantea que la **corrupción**, entendida como uno de los principales problemas estructurales en las sociedades modernas, puede estudiarse empíricamente mediante el análisis de datos provenientes de indicadores internacionales, como el *Barómetro Global de Corrupción* de Transparencia Internacional. Su hipótesis central sostiene que el comportamiento de la corrupción en los países sigue patrones cuantificables, y que estos pueden ser modelados a través de redes neuronales capaces de identificar tendencias subyacentes en los datos históricos.

Para validar esta hipótesis, se emplearon datos de los años 2004, 2005, 2006, 2010 y 2013, que fueron procesados mediante la herramienta **Weka**, aplicando algoritmos de **machine learning** como **vecinos más próximos**, **árboles de decisión** y **perceptrón multicapa**. Además, se incluyeron análisis de **entropía**, **exponentes de Lyapunov** y **clusterización**, con el fin de evaluar la estabilidad del sistema y la calidad de las predicciones generadas.

Este trabajo destaca por integrar técnicas estadísticas y de inteligencia artificial para abordar un problema social complejo. Su principal contribución radica en demostrar que las **redes neuronales multicapa**, combinadas con métricas de entropía y estabilidad, pueden ofrecer una herramienta útil para anticipar la evolución de la corrupción y orientar la toma de decisiones en políticas públicas.

Base de Datos

Para este estudio, el autor utilizó un conjunto de datos procedente de indicadores internacionales de corrupción, entre ellos el **Barómetro Global de Corrupción** y el **Índice de Percepción de la Corrupción** de Transparencia Internacional, junto con bases estadísticas complementarias del **Banco Mundial** y la **OCDE**. El periodo de análisis comprendió los años **2004, 2005, 2006, 2010 y 2013**, con el objetivo de identificar patrones temporales y espaciales en la evolución de la corrupción a nivel mundial. Los datos fueron preprocesados para eliminar duplicidades, valores faltantes y variables irrelevantes, obteniendo un conjunto final equilibrado de indicadores numéricos y categóricos que reflejan aspectos políticos, económicos y sociales.

Metodología

Preparación de Datos

El procesamiento de datos incluyó técnicas de normalización y reducción de dimensionalidad para optimizar la calidad de las entradas del modelo. Se evaluaron métricas estadísticas de dispersión y correlación para detectar multicolinealidad y redundancias. Posteriormente, se definieron

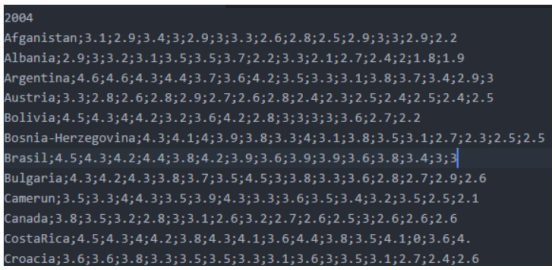

Fig. 7: Archivos de datos original

los conjuntos de entrenamiento y prueba, manteniendo un equilibrio entre países con niveles de corrupción altos y bajos, con el fin de evitar sesgos en el aprendizaje.

2) Procesamiento de Datos:

- **Limpieza inicial**: Se eliminaron observaciones con más del 10% de datos faltantes y se homogenizaron los formatos numéricos y textuales.
- **Normalización**: Todas las variables fueron escaladas entre 0 y 1 mediante la técnica *Min-Max Scaling*.
- **Selección de variables**: Se seleccionaron únicamente aquellas con correlaciones superiores al 0.5 con respecto al índice de corrupción.

Arquitectura del Modelo

El autor implementó una red neuronal **Perceptrón Multicapa (MLP)** con una arquitectura de tres capas: una capa de entrada con 15 neuronas correspondientes a los indicadores seleccionados, una capa oculta con 10 neuronas y una capa de salida con una sola neurona para predecir el nivel estimado de corrupción. Se utilizó la función de activación **sigmoide**, el optimizador **backpropagation** y una tasa de aprendizaje adaptativa. El modelo fue implementado con la herramienta **Weka**, junto con otros algoritmos comparativos como **Árboles de decisión** y **K-vecinos más cercanos (KNN)**.

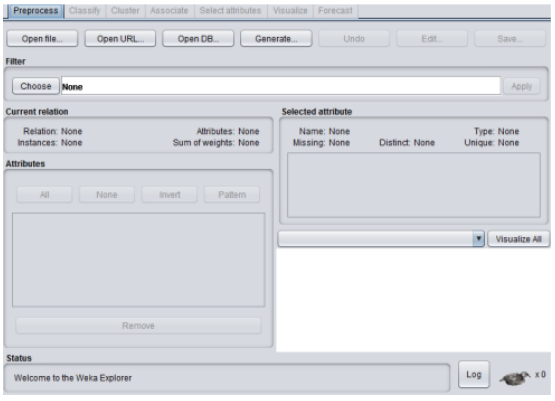

Fig. 8: Interfaz de Weka utilizada para la implementación del modelo Perceptrón Multicapa (MLP). Fuente: Sánchez Fernández (2017) [7].

Resultados El modelo de red neuronal alcanzó una **precisión superior al 90%** en la predicción de los niveles de corrupción por país. El análisis comparativo mostró que la red neuronal superó en rendimiento a los algoritmos clásicos

(Árboles de Decisión y KNN). Además, el autor aplicó métricas de **entropía** y **exponentes de Lyapunov** para evaluar la estabilidad y la predictibilidad del sistema, concluyendo que la corrupción presenta patrones caóticos pero parcialmente predecibles.

En términos prácticos, el estudio demostró que es posible anticipar la evolución de la corrupción en diferentes regiones del mundo a partir de datos históricos, proporcionando una herramienta útil para la formulación de políticas públicas orientadas a la transparencia.

A Machine Learning-based Anomaly Detection Framework in Life Insurance Contracts

El artículo *"A Machine Learning-based Anomaly Detection Framework in Life Insurance Contracts"* [8], desarrollado por Andreas Groll, Akshat Khanna y Leonid Zeldin (2024), propone un marco de detección de anomalías basado en **aprendizaje automático** para identificar irregularidades en contratos de seguros de vida. La motivación del estudio surge de la necesidad de garantizar la integridad de los datos en las bases de información del sector asegurador, donde las anomalías pueden indicar **errores, fraudes o comportamientos atípicos** que afecten la confianza y estabilidad del sistema financiero.

Los autores destacan que la principal dificultad en este contexto es la **escasez de datos etiquetados**, lo que dificulta el uso de métodos supervisados tradicionales. Por ello, se recurre a enfoques **no supervisados**, que permiten detectar desviaciones en los datos sin necesidad de etiquetas previas. La hipótesis central sostiene que las técnicas de **detección automática de anomalías**, especialmente las basadas en aprendizaje profundo, pueden superar las limitaciones de los métodos clásicos y ofrecer resultados más precisos y escalables.

La metodología propuesta combina enfoques de **machine learning clásico** (como **k-means**, **DBSCAN**, **HDBSCAN**, y **One-Class SVM**) con modelos de **deep learning**, entre ellos los **autoencoders** y **autoencoders variacionales (VAE)**. Estos últimos se entrenan para reconstruir los datos de entrada y detectar anomalías a partir del **error de reconstrucción**, proporcionando un marco adaptable a distintos conjuntos de datos. Para la validación se emplearon dos bases de datos abiertas de seguros de salud, donde se introdujeron manualmente anomalías simuladas para evaluar la capacidad de detección de los modelos.

Base de Datos

Para este estudio, los autores utilizaron dos conjuntos de datos abiertos relacionados con contratos de seguros de salud y de vida, que fueron adaptados al contexto del proyecto para validar el marco propuesto de detección de anomalías. Los datos contienen **miles de registros de contratos**, con variables que incluyen información demográfica, monto asegurado, duración del contrato, tipo de prima y siniestralidad. Con el fin de evaluar la capacidad de detección de los modelos, se introdujeron manualmente **anomalías simuladas** en las bases de datos, imitando patrones de fraude o errores típicos de registro.

Variable name	Type	Description
ID	Categorical	Unique identifier for each entry
Age	Continuous	Age of the individual
Diabetes	Categorical	Indicates if the individual has diabetes
BloodPressureProblems	Categorical	Indicates if the individual has blood pressure problems
AnyTransplants	Categorical	Indicates if the individual has had any transplants
AnyChronicDiseases	Categorical	Indicates if the individual has any chronic diseases
Height	Continuous	Height of the individual
Weight	Continuous	Weight of the individual
KnownAllergies	Categorical	Indicates if the individual has known allergies
HistoryOfCancerInFamily	Categorical	Indicates if there is a history of cancer in the family
NumberOfMajorSurgeries	Continuous	Number of major surgeries the individual has had
PremiumPrice	Continuous	Price of the insurance premium

TABLE I: Estructura general de uno de los datasets utilizados en el marco de detección de anomalías.

Metodología

Preparación de Datos

El proceso de preparación de datos incluyó técnicas de **estandarización** para las variables numéricas y **codificación one-hot** para las categóricas. Posteriormente, los datos fueron particionados en conjuntos de **entrenamiento (70%)**, **validación (15%)** y **prueba (15%)**, garantizando representatividad entre los diferentes tipos de contratos.

Modelos Implementados: Los autores desarrollaron un marco híbrido que combina algoritmos de aprendizaje automático clásico con modelos de aprendizaje profundo, estructurado de la siguiente forma:

- **Modelos clásicos**: K-means, DBSCAN, HDBSCAN y One-Class SVM.
- **Modelos de aprendizaje profundo**: Autoencoders y Autoencoders Variacionales (VAE), entrenados para reconstruir los datos de entrada y detectar desviaciones significativas.

Arquitectura del Modelo

El modelo central se basa en un **Autoencoder Variacional (VAE)** con tres capas ocultas de 64, 32 y 16 neuronas, empleando funciones de activación ReLU y una pérdida basada en **error de reconstrucción**. El objetivo del modelo es detectar patrones no lineales y relaciones ocultas que caracterizan anomalías en contratos de seguros. Para la validación se utilizó la métrica **Area Under the ROC Curve (AUC)** y la **precisión de detección de anomalías**.

Resultados

Los resultados mostraron que el **Autoencoder Variacional (VAE)** alcanzó una **precisión del 96%** en la identificación de anomalías simuladas, superando ampliamente a los modelos clásicos, especialmente en entornos con alta dimensionalidad. El análisis del **error de reconstrucción** permitió identificar las instancias más anómalas y visualizar su distribución mediante proyecciones en el espacio latente.

modelo	dataset 1		dataset 2	
	tiempo	mods	tiempo	mods
autoencoder	2:22	3	6:00	3
vae	4:35	3	20:00	3

TABLE II: Resultados compactos de detección de anomalías para autoencoder y VAE en ambos datasets.

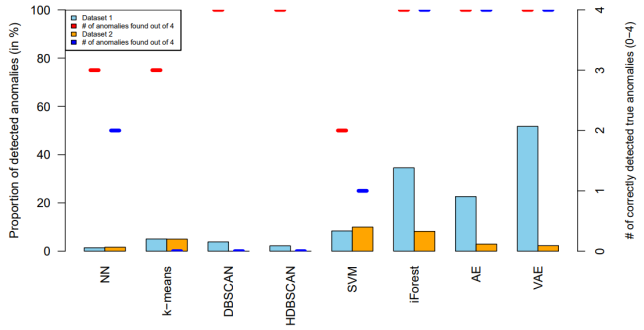

Fig. 9: Proporción de anomalías y anomalías detectadas manualmente (eje y).

En conjunto, los hallazgos demostraron que los modelos basados en **aprendizaje profundo** ofrecen una solución robusta, escalable y adaptable a distintas fuentes de datos del sector asegurador, mejorando la eficiencia en la detección de fraudes, errores administrativos y desviaciones en contratos reales.

Detection of fraud in public procurement using data-driven methods: a systematic mapping study

El trabajo titulado *"Detection of fraud in public procurement using data-driven methods: a systematic mapping study"* [9], desarrollado por Everton Schneider dos Santos, Matheus Machado dos Santos, Márcio Castro y Jônata Tyska Carvalho, fue publicado en el año **2025**. Surge de la necesidad de comprender cómo las **técnicas basadas en datos** se aplican en la detección de fraude y corrupción dentro de los procesos de contratación pública, un tema de creciente relevancia para la gobernanza y la transparencia institucional.

Los autores plantean que la detección de irregularidades en las contrataciones públicas requiere enfoques sistemáticos que integren **métodos de análisis de datos, aprendizaje automático y técnicas estadísticas**, ya que los procesos de adquisición gubernamental involucran grandes volúmenes de información y múltiples actores. Su objetivo central es realizar un **mapeo sistemático de la literatura** para identificar qué tipos de fraude son más estudiados, qué fuentes de datos se utilizan y qué metodologías se aplican.

Para alcanzar este propósito, los investigadores analizaron más de **6000 publicaciones científicas**, de las cuales seleccionaron **93 estudios primarios** utilizando el protocolo *PRISMA* y el procedimiento de *snowballing*. Clasificaron los trabajos en tres grandes campos metodológicos: **aprendizaje automático**, **ciencia de redes** y **análisis estadístico**.

Base de Datos

Para la recopilación de estudios, los autores aplicaron el protocolo **PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)**, analizando inicialmente más de **6000 publicaciones científicas** relacionadas con la detección de fraude en adquisiciones gubernamentales. Luego de un riguroso proceso de selección y depuración, el conjunto final quedó conformado por **93 estudios primarios** publicados entre 2006 y 2024, los cuales fueron analizados y clasificados según su enfoque metodológico, tipo de fraude y técnicas de análisis empleadas.

Las fuentes de datos de los artículos incluidos provienen principalmente de **bases de contratación pública**, **portales de transparencia**, y **registros administrativos de gobiernos nacionales y municipales**. Además, algunos estudios complementaron sus conjuntos de datos con información proveniente de **fuentes externas**, tales como indicadores socioeconómicos, datos de desempeño de proveedores o auditorías de organismos de control.

Field	RQ	Description
ID	–	Identification number assigned to each paper
Year	–	Year of publication
Language	–	Language of publication
Publication Type	–	Type of publication (journal or conference)
Journal	–	Title of the publication
JIF	–	Journal Impact Factor
Title	–	Title of the paper
Method	RQ1	Method used in the paper
Method Group	RQ1	Type of method used in the paper
Field	RQ1	Field of study of the paper
Hybrid Strategy	RQ1	Paper used methods from different fields
Fraud Type	RQ2	Type of fraud studied in the paper
Fraud Group	RQ2	Group of the fraud studied
Data Type	RQ3	Type of data used in the paper
Country	RQ3	Country of origin of the data
Samples	RQ3	Number of samples in the data
Economic Sector	RQ3	Economic sector of the data
Available Data?	RQ3	Authors published the data used in the paper
Data Period	RQ3	Range of dates of the data (start and end)
Feature	RQ4	Names of the features used in the paper
Feature Group	RQ4	Group of the feature used
Red Flag?	RQ4	Feature used is a red flag for corruption
Target Type	RQ4	Group of the target used
Target	RQ4	Names of the target used in the paper
Measure Type	RQ5	Type of measure used to describe the results
Result	RQ5	Value or textual synthesis of the results

TABLE III: Campos del dataset y su relación con las preguntas de investigación (RQ).

Metodología

Procedimiento de Revisión Sistemática

El proceso metodológico del estudio siguió cuatro fases principales:

1) **Identificación**: búsqueda exhaustiva de literatura en bases de datos como Scopus, Web of Science, IEEE Xplore y Google Scholar, utilizando combinaciones de palabras clave como *"public procurement fraud"*, *"corruption detection"*, y *"machine learning"*.
2) **Filtrado**: aplicación de criterios de inclusión y exclusión, priorizando estudios con aplicación empírica en el contexto de contrataciones públicas.
3) **Clasificación**: codificación de los artículos según tipo de fraude, técnica utilizada y dominio de aplicación.
4) **Síntesis y análisis**: integración de hallazgos cuantitativos y cualitativos para identificar tendencias y vacíos de investigación.

Clasificación de Estudios y Técnicas Analíticas

Los artículos seleccionados fueron clasificados según tres dimensiones principales:

- **Tipo de técnica analítica**:
  - Métodos estadísticos: **52.7%** (regresiones lineales, logísticas y análisis basados en la Ley de Benford).
  - Aprendizaje automático: **28%** (modelos de **árboles de decisión**, **Random Forest**, **SVM**, y **Gradient Boosting**).
  - Ciencia de redes: **17%** (detección de colusión mediante análisis de grafos y métricas de centralidad).
- **Tipo de fraude analizado**:
  - **Colusión**: acuerdos entre oferentes o manipulación de licitaciones.
  - **Favoritismo**: asignación sesgada de contratos a proveedores específicos.
  - **Anomalías generales**: irregularidades administrativas o transaccionales.
- **Fuentes de datos y alcance geográfico**: predominio de estudios basados en **datos públicos europeos** y **latinoamericanos**, con creciente interés en la creación de **datasets sintéticos** para entrenamiento de modelos.

Method group	Anomalies	Collusion	Favoritism	Others
Clusters	–	–	89.0	–
Ensemble Methods	85.95	95.30	–	55.67
Linear Models	26.60	95.98	28.0	–
Naïve Bayes	–	96.70	7.0	–
Nearest Neighbors	–	92.95	–	–
Neural Networks	–	91.07	84.55	–
Others	–	91.0	60.70	43.67

TABLE IV: Desempeño promedio de los grupos de métodos según el tipo de fraude analizado.

Resultados

El análisis de los estudios revisados permitió identificar varias tendencias y hallazgos clave:

- El uso de **métodos estadísticos tradicionales** sigue siendo dominante, aunque los enfoques de **aprendizaje automático** están ganando relevancia por su mayor capacidad predictiva.
- Los modelos más efectivos en la detección de fraude son aquellos que incorporan **relaciones entre actores** (compradores y proveedores), en lugar de centrarse únicamente en variables contractuales individuales.
- Las técnicas de **Network Science** han demostrado ser especialmente útiles para identificar **patrones de colusión** y **comunidades de proveedores interconectados**.
- Se evidencia una carencia de **datasets públicos y reproducibles**, lo que limita la comparación y validación de resultados entre estudios.

Sector	Fraud group	Model	Papers (ID and ref)
Construction	Collusion	ML	2 [128], 3 [127], 9 [100], 13 [50], 22 [58]
Construction	Collusion	NS	81 [139], 86 [129]
Construction	Collusion	Statistics	19 [112], 26 [105], 41 [61], 42 [108], 66 [107], 67 [109], 92 [131], 93 [130]
Construction	Other Frauds	Statistics	63 [21]
General	Collusion	ML	4 [49], 10 [52], 27 [70], 85 [9]
General	Collusion	NS	5 [124], 43 [16], 51 [74], 60 [88], 90 [137]
General	Collusion	Statistics	87 [77]
General	Favoritism	ML	6 [103], 16 [119], 17 [118], 23 [42], 30 [81], 31 [71], 57 [41], 68 [25], 72 [79]
General	Favoritism	NS	20 [96], 28 [46], 54 [34], 61 [126]
General	Favoritism	Statistics	18 [19], 35 [20], 36 [5], 38 [59], 46 [131], 47 [29], 55 [37], 59 [28], 64 [132], 69 [38], 70 [102], 78 [27], 82 [18], 83 [73], 91 [135]
General	Other Frauds	ML	11 [115], 77 [89], 84 [78]
General	Other Frauds	NS	48 [87]
General	Other Frauds	Statistics	40 [110], 45 [72], 53 [112], 56 [97], 76 [53]
General	Anomalies	ML	12 [30], 14 [35], 50 [17]
General	Anomalies	NS	21 [64], 62 [91]
General	Anomalies	Statistics	15 [98], 49 [1], 52 [8], 58 [125], 71 [39], 73 [92], 88 [101]
Others	Collusion	NS	33 [32]
Others	Collusion	Statistics	24 [10], 25 [22], 29 [36], 37 [62], 80 [23]
Others	Anomalies	NS	34 [15]
Others	Anomalies	Statistics	1 [90], 44 [82]
Transport	Collusion	ML	7 [111]
Transport	Collusion	Statistics	39 [4], 65 [106], 74 [26], 89 [67]
Transport	Favoritism	Statistics	79 [6]
Transport	Anomalies	ML	8 [24]

TABLE V: Principales tendencias identificadas en el sector económico sobre detección de fraude en contrataciones públicas.

Conclusiones

Los autores concluyen que la investigación sobre detección de fraude en contrataciones públicas se encuentra en una fase de madurez intermedia, caracterizada por la consolidación de métodos estadísticos y el crecimiento sostenido de enfoques basados en aprendizaje automático y análisis de redes. Sin embargo, resaltan la necesidad de:

- Fomentar la creación de **bases de datos abiertas y estandarizadas**.
- Desarrollar **modelos híbridos** que integren información tabular, textual e incluso visual (por ejemplo, con modelos como Donut).

- Promover la **implementación práctica** de sistemas de alerta temprana en organismos gubernamentales.

El estudio aporta una visión integral y sistematizada del estado actual del uso de la ciencia de datos en la detección de corrupción pública, sirviendo como una guía fundamental para investigaciones futuras en **auditoría digital**, **transparencia gubernamental** y **ética pública**.

OCR-free Document Understanding Transformer (Donut)

El trabajo titulado *"OCR-free Document Understanding Transformer (Donut)"* [2], desarrollado por Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han y Seunghyun Park, fue publicado en el año **2022** por el laboratorio de inteligencia artificial de NAVER (NAVER AI Lab). Surge de la motivación por superar las limitaciones de los modelos tradicionales de comprensión de documentos (*Visual Document Understanding, VDU*), los cuales dependen de motores de reconocimiento óptico de caracteres (*OCR*) que suelen ser costosos, inflexibles y propensos a errores.

Los autores proponen un modelo denominado **Donut (Document Understanding Transformer)**, basado completamente en una arquitectura **Transformer end-to-end**, capaz de procesar imágenes de documentos (como facturas, recibos o contratos) **sin necesidad de OCR**. La hipótesis central sostiene que es posible entrenar un modelo que aprenda directamente la correspondencia entre la imagen del documento y su representación estructurada en formato **JSON**, lo que permite eliminar etapas intermedias y reducir errores acumulativos.

El modelo Donut está compuesto por un **codificador visual** basado en *Swin Transformer*, encargado de extraer características de la imagen, y un **decodificador textual** basado en *BART*, responsable de generar secuencias de texto estructuradas. Su entrenamiento se desarrolla en dos fases: una etapa de **pre-entrenamiento**, utilizando datos sintéticos multilingües generados mediante el sistema **SynthDoG**, y una fase de **fine-tuning** orientada a tareas específicas como clasificación de documentos, extracción de información y respuestas visuales a preguntas (*DocVQA*).

Este trabajo destaca por introducir un nuevo paradigma en la comprensión de documentos digitales, permitiendo el procesamiento directo de imágenes con alta eficiencia y precisión. Los autores concluyen que Donut representa un avance significativo hacia la **automatización de la lectura y análisis de documentos**, con aplicaciones potenciales en la **detección de fraude documental**, la **auditoría automatizada** y la **gestión inteligente de información gubernamental**.

Base de Datos

Para el desarrollo y evaluación del modelo, los autores emplearon tanto **datasets reales como sintéticos**. Durante la fase de preentrenamiento, se utilizó el conjunto **SynthDoG (Synthetic Document Generator)**, un corpus multilingüe creado específicamente para el entrenamiento de modelos de comprensión de documentos sin OCR. Este conjunto simula facturas, recibos, contratos y formularios en diferentes idiomas y estructuras.

Posteriormente, en la fase de evaluación, se emplearon tres bases de datos ampliamente utilizadas en la literatura:

- **RVL-CDIP**: conjunto de 400,000 imágenes de documentos escaneados, utilizados para la tarea de *clasificación de tipo de documento*.
- **CORD**: base de datos de recibos y facturas con anotaciones estructuradas en formato JSON, utilizada para tareas de *extracción de información*.
- **DocVQA**: conjunto de imágenes con preguntas y respuestas asociadas, empleado para la evaluación en *Visual Question Answering*.

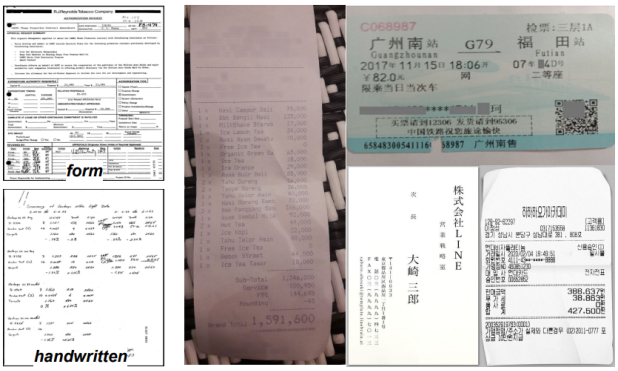

Fig. 10: Documentos de los datasets utilizados.

Metodología

Preparación de Datos

Durante la etapa de entrenamiento, todas las imágenes fueron redimensionadas a una resolución fija de 1280×960 píxeles y normalizadas entre 0 y 1. Los textos asociados se convirtieron a representaciones estructuradas en formato **JSON**, donde cada campo (por ejemplo, nombre, fecha, total, etc.) se mapeó a un token específico del vocabulario de salida.

3) Preprocesamiento y Estructura del Corpus:

- **Estandarización visual**: normalización del tamaño, brillo y contraste de las imágenes de documentos.
- **Tokenización estructural**: codificación jerárquica de los elementos textuales en secuencias JSON para aprendizaje supervisado.
- **Aumento de datos**: rotaciones, recortes y variaciones tipográficas simuladas para mejorar la robustez del modelo.

Arquitectura del Modelo Donut

El modelo **Donut (Document Understanding Transformer)** se basa completamente en una arquitectura **Transformer end-to-end** que prescinde del uso de OCR. Su estructura general consta de dos módulos principales:

- **Codificador visual**: construido sobre *Swin Transformer*, responsable de extraer representaciones espaciales jerárquicas del documento de entrada.
- **Decodificador textual**: basado en *BART (Bidirectional and Auto-Regressive Transformer)*, encargado de generar

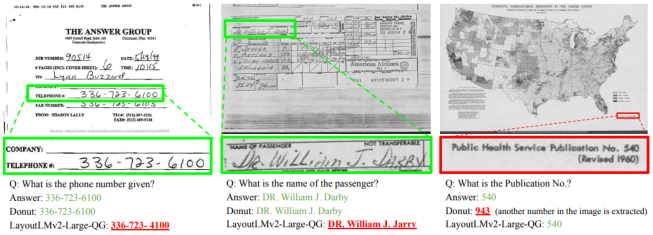

Fig. 11: Ejemplos de como son las salidas de Donut y LayoutLMv2 en DocVQA.

secuencias JSON interpretables directamente desde las características visuales.

El modelo se entrena mediante un esquema de dos fases:

1) **Pre-entrenamiento con SynthDoG**: aprendizaje general del formato y la estructura de documentos multilingües sintéticos.
2) **Fine-tuning en tareas específicas**: ajuste del modelo en tareas concretas como clasificación, extracción de información y respuesta a preguntas.

Conjunto de ajuste fino	OCR	#Params	Tiempo	ANLS test set
BERT	train set	110M + α	1517	63.5
LayoutLMv3	train set	113M + α	1519	69.8
LayoutLMv2	train set	200M + α	1610	78.1
Donut	train set	176M	782	67.5
LayoutLMv2-Large-QG	train+dev+QG	390M + α	1698	86.7

TABLE VI: Puntajes de Similitud de Levenshtein Normalizada Promedio (ANLS) en DocVQA.

Entrenamiento y Evaluación

Durante el entrenamiento, se utilizaron lotes de 32 documentos por iteración, con una tasa de aprendizaje de 3×10^{-5} y optimizador **AdamW**. El entrenamiento completo requirió aproximadamente 20 horas en una GPU A100.

La evaluación se realizó sobre las tres tareas principales:

1) **Clasificación de documentos (RVL-CDIP)**: precisión del **95.3%**.
2) **Extracción de información (CORD)**: **F1 = 91.6%**, **accuracy = 93.5%**.
3) **Visual Question Answering (DocVQA)**: **Exactitud del 86.7%**.

Además, el modelo logró un tiempo de inferencia **dos veces más rápido** que los sistemas dependientes de OCR, con una reducción significativa en el consumo de memoria y una mejora de precisión en idiomas con escritura compleja.

Resultados

Los resultados demuestran que el modelo Donut no solo supera a los sistemas tradicionales basados en OCR, sino que además ofrece una solución más eficiente y generalizable para la comprensión de documentos escaneados. Su capacidad de aprender directamente desde la imagen permite evitar errores de segmentación y transcripción típicos del OCR, mejorando la precisión global del sistema.

En particular, el modelo muestra un desempeño robusto en contextos multilingües, lo que lo convierte en una herramienta potencial para la **detección automatizada de irregularidades en documentos públicos, facturas o contratos gubernamentales**. Al eliminar la dependencia del OCR, Donut representa un avance clave hacia la **automatización completa del análisis documental**, con aplicaciones directas en **auditoría digital, transparencia y control anticorrupción**.

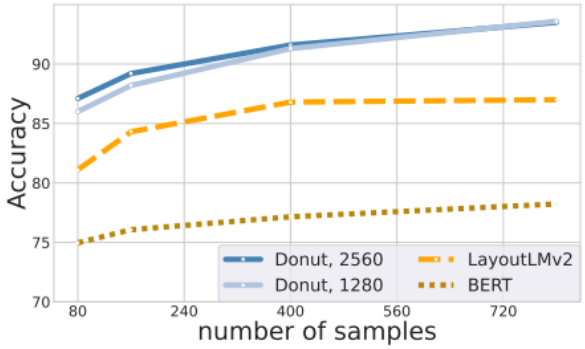

Fig. 12: Comparación de desempeño entre Donut y modelos OCR-dependientes en distintas tareas.

III. Bases Teóricas

A. *Contratos públicos y riesgos de anomalías*

Los procesos de contratación pública constituyen uno de los pilares para el funcionamiento adecuado del Estado. Según el Banco Mundial, "las compras públicas representan en promedio entre el 15% y el 20% del PIB en los países en desarrollo, lo que las convierte en una actividad especialmente vulnerable a prácticas irregulares" [20]. En el contexto peruano, el Organismo Supervisor de las Contrataciones del Estado (OSCE) establece lineamientos para garantizar transparencia; sin embargo, la gran cantidad de transacciones digitales y documentos asociados dificulta la supervisión manual.

Las **anomalías en contratos públicos** pueden manifestarse como valores atípicos en montos adjudicados, proveedores recurrentes, sobrecostos injustificados, duraciones atípicas o documentos con información inconsistente. La literatura especializada señala que "la detección temprana de irregularidades permite reducir pérdidas económicas y mejorar la integridad institucional" [21]. En consecuencia, la automatización mediante técnicas de aprendizaje automático se vuelve una herramienta estratégica para identificar patrones inusuales en grandes volúmenes de datos administrativos.

Como se observa en la Figure 13, el proceso de contratación pública incluye etapas que van desde la identificación de la necesidad hasta la ejecución contractual, y en cada una de ellas pueden surgir irregularidades que requieren ser analizadas mediante métodos automatizados.

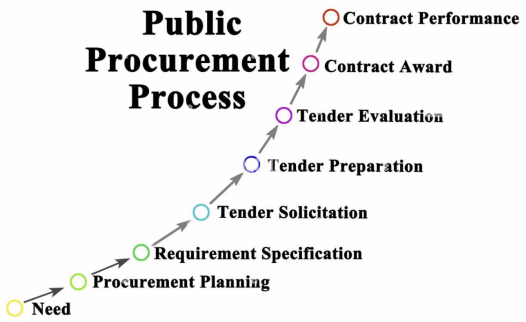

Fig. 13: Etapas del Proceso de Contratación Pública: necesidad, planificación, preparación, evaluación y ejecución [**?**].

B. Preprocesamiento de datos en análisis de contratos públicos

El preprocesamiento de datos es una etapa esencial para garantizar la calidad de los modelos de aprendizaje automático aplicados a contratos públicos. Según Han, Kamber y Pei, "la calidad de un modelo depende directamente de la calidad de los datos que recibe; los algoritmos no pueden corregir de manera automática errores, inconsistencias o valores faltantes" [22]. En el contexto de la contratación estatal, las bases de datos suelen contener información heterogénea, valores nulos, codificaciones diferentes y textos no estructurados provenientes de documentos administrativos.

El proceso inicia con la **limpieza de datos**, que incluye la detección y corrección de errores tipográficos, el tratamiento de valores faltantes y la identificación de outliers numéricos que puedan sesgar el entrenamiento. Posteriormente, se aplica la **normalización y estandarización** de variables numéricas para mejorar la estabilidad de modelos sensibles a la escala, como KNN o regresión logística. Para variables categóricas, se emplean técnicas como *One-Hot Encoding*, lo que permite representar información institucional (proveedor, tipo de proceso, modalidad de compra) en un formato numérico compatible con algoritmos supervisados.

Además, el preprocesamiento incluye el manejo del **texto contractual**, que requiere tokenización, eliminación de stopwords y lematización antes de aplicar representaciones vectoriales como TF-IDF. Como señalan Feldman y Sanger (2007), "el procesamiento sistemático de documentos textuales es esencial para detectar patrones semánticos relevantes y apoyar la toma de decisiones" [23].

El flujo general de este proceso, mostrado en la Figure 14, permite garantizar que los datos finales sean coherentes, estructurados y representativos, favoreciendo que los modelos identifiquen anomalías con mayor precisión.

C. Embeddings semánticos mediante modelos preentrenados para el análisis de contratos públicos

Los *embeddings* de texto constituyen una de las técnicas más avanzadas dentro del procesamiento de lenguaje natural (NLP), ya que permiten representar documentos mediante vectores densos que capturan relaciones semánticas profundas.

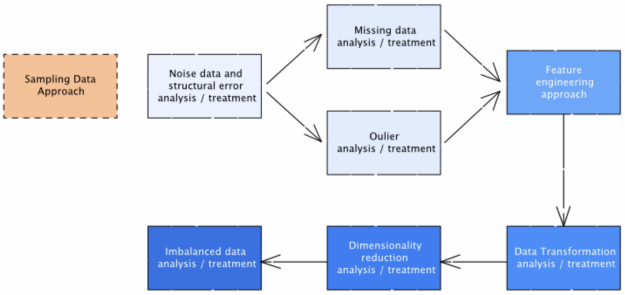

Fig. 14: Pipeline de preprocesamiento de datos utilizado en técnicas de aprendizaje automático.

Según Mikolov et al., "los embeddings permiten mapear palabras y textos en un espacio continuo donde la distancia geométrica refleja su similitud contextual" [29]. Esta capacidad resulta esencial en entornos complejos como la contratación pública, donde los documentos contienen lenguaje técnico, jurídico y administrativo cuya interpretación excede las capacidades de técnicas tradicionales basadas únicamente en frecuencias de palabras.

En este proyecto se utiliza el modelo preentrenado *paraphrase-multilingual-MiniLM-L12-v2*, perteneciente a la familia Sentence-BERT (SBERT). De acuerdo con Reimers y Gurevych, SBERT "modifica la arquitectura de BERT para generar representaciones semánticas altamente efectivas en tareas de comparación y agrupamiento de textos" [30]. Este modelo aplica una arquitectura basada en *Transformers* capaz de procesar dependencias contextuales dentro de los documentos, produciendo vectores que reflejan tanto el contenido como el significado implícito del texto contractual.

El modelo utilizado es *multilingüe*, lo que permite procesar contratos redactados en distintos idiomas y estilos institucionales, manteniendo la coherencia semántica de las representaciones. Cada documento convertido a embedding genera un vector de dimensión reducida pero altamente informativa, lo que facilita su uso en algoritmos de aprendizaje automático. En el presente estudio, estos embeddings se emplean como entrada para el algoritmo de *clustering* K-Means, con el objetivo de identificar grupos temáticos dentro del contenido de los contratos.

El resultado de este proceso es la variable *cluster_texto*, una categorización construida a partir de la similitud semántica de los documentos. Esta variable resume la temática predominante de cada contrato y aporta información estructural relevante para la fase de modelado supervisado, fortaleciendo la detección de patrones asociados a riesgos o anomalías dentro de los procesos de contratación pública.

D. Detección de anomalías supervisada en contratos públicos

La detección de anomalías supervisada consiste en entrenar modelos de clasificación capaces de distinguir entre transacciones regulares y aquellas consideradas sospechosas según un conjunto de etiquetas previamente definidas. Chandola, Banerjee y Kumar describen esta técnica como "el uso de algoritmos que aprenden un modelo discriminativo a partir

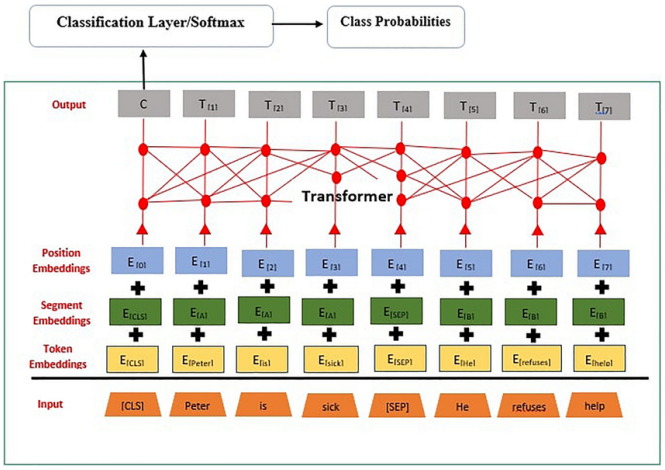

Fig. 15: Esquema general del proceso de generación de *embeddings* mediante modelos basados en Transformers.

de ejemplos etiquetados para identificar patrones inusuales en datos futuros" [24]. En el ámbito de la contratación pública, este enfoque permite aprovechar bases de datos históricas donde ya se han detectado casos de irregularidad, transformándolas en insumos para el aprendizaje automático.

Este tipo de detección ofrece ventajas clave frente a métodos no supervisados: permite identificar anomalías específicas asociadas a sobrecostos, proveedores recurrentes o procedimientos inusuales, al mismo tiempo que genera predicciones más precisas gracias a la información contextual proporcionada por las etiquetas. Además, estudios recientes indican que "los modelos supervisados tienden a superar a los métodos tradicionales al analizar conjuntos de datos complejos en compras públicas" [25]. Por ello, se integran modelos como regresión logística, árboles de decisión, KNN y AdaBoost, los cuales permiten capturar distintas formas de irregularidad tanto en datos numéricos como textuales.

En este proyecto, la detección supervisada es esencial debido a la disponibilidad de registros con categorizaciones históricas de riesgo. Estos datos permiten entrenar modelos capaces de predecir automáticamente la probabilidad de anomalía en contratos futuros, fortaleciendo la transparencia estatal y optimizando los mecanismos de fiscalización automatizada.

El flujo general del proceso de detección supervisada, ilustrado en la Figure 16, permite comprender cómo los datos son preprocesados, clasificados y evaluados para estimar el riesgo de anomalías en nuevas transacciones. Este esquema constituye un componente fundamental del proyecto, pues permite automatizar la predicción de contratos públicos sospechosos basándose en patrones previamente aprendidos.

E. Modelos supervisados utilizados en la detección de anomalías

Los modelos supervisados permiten aprender patrones a partir de datos etiquetados y son ampliamente utilizados en la detección de anomalías en contratos públicos. Según Bishop, "el aprendizaje supervisado busca construir una función que

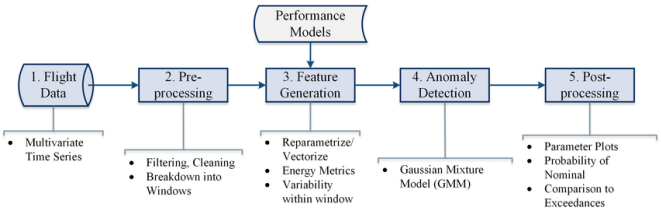

Fig. 16: Arquitectura del proceso de detección de anomalías basado en modelos de rendimiento.

asigne correctamente etiquetas a nuevas observaciones a partir de ejemplos previos" [26]. En este proyecto se emplean cuatro modelos: *Decision Tree*, *Logistic Regression*, *KNN* y *AdaBoost*, cada uno con propiedades que favorecen la identificación de riesgos contractuales.

El **árbol de decisión** es un modelo interpretable que divide los datos en regiones mediante reglas lógicas. Su ventaja clave es que "proporciona modelos fácilmente explicables, especialmente útiles en contextos gubernamentales" [27]. La **regresión logística** es un modelo lineal que estima probabilidades de pertenecer a una clase mediante la función sigmoide, siendo efectiva para problemas binarios en los que se desea una medida clara de riesgo.

Por otro lado, el **K-Nearest Neighbors (KNN)** clasifica basándose en la proximidad de una observación respecto a sus vecinos más cercanos, permitiendo detectar anomalías mediante patrones locales. Finalmente, **AdaBoost** es un método de ensamble que combina múltiples modelos débiles para mejorar la precisión; como indican Freund y Schapire, "AdaBoost ajusta de manera adaptativa el peso de los ejemplos difíciles de clasificar, incrementando la capacidad predictiva" [28].

La combinación de estos modelos permite capturar distintos tipos de irregularidad estructural: patrones lineales, relaciones no lineales, vecindarios anómalos y comportamientos complejos detectados mediante ensambles, fortaleciendo la predicción de riesgo en licitaciones públicas.

El conjunto de estos algoritmos, ilustrado en la Figure 17, permite capturar relaciones lineales, no lineales, vecindarios anómalos y patrones complejos, fortaleciendo significativamente la capacidad de predicción de riesgo en licitaciones públicas.

IV. Propuesta Metodológica

La metodología desarrollada para la detección automatizada de anomalías en contratos públicos del Estado peruano se estructura en distintas fases que abarcan desde la recolección y consolidación de datos hasta la preparación del conjunto de entrenamiento para los modelos de aprendizaje automático. A continuación, se describen las etapas iniciales del proceso metodológico.

A. Recolección de Datos

La información utilizada en este estudio fue obtenida a partir de fuentes oficiales del Gobierno del Perú, específicamente de los portales de datos abiertos y transparencia pública. El conjunto principal de contratos se descargó del portal *CONOSCE*

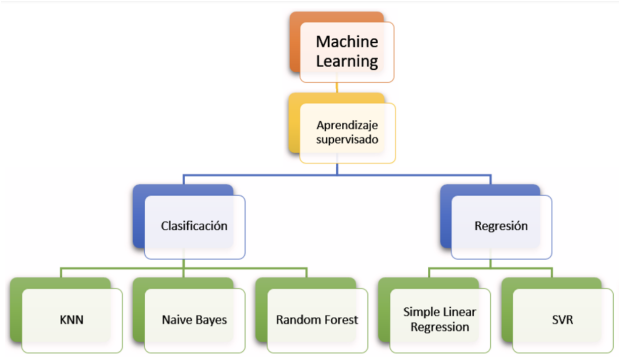

Fig. 17: Diagrama de los principales algoritmos del aprendizaje supervisado dentro del campo de *Machine Learning*.

– *Datos Abiertos de Contratos Públicos* del SEACE, el cual contiene información detallada sobre los contratos registrados por las entidades públicas, incluyendo fechas de suscripción, inicio, fin, prórrogas, adicionales y reducciones.

Complementariamente, se consultaron las plataformas:

- **Mapa de Contratos Públicos** [11] donde se verifican los contratos vigentes y adjudicados por región, entidad y tipo de proceso.

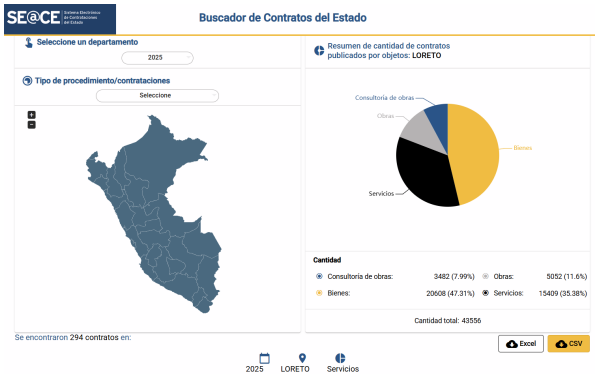

Fig. 18: Visualización del Mapa de Contratos Públicos (SEACE).

- **Portal de Transparencia Económica del MEF** [12] que permite consultar por número de RUC el historial de contrataciones, montos adjudicados, sanciones e infracciones de las empresas proveedoras.

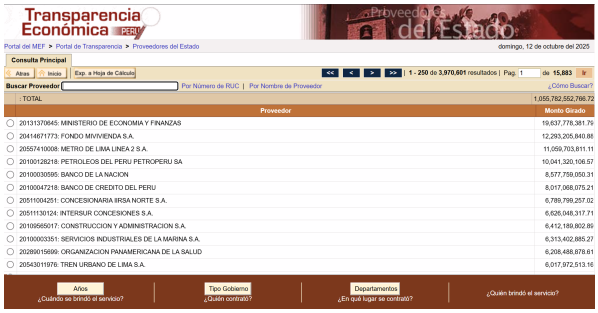

Fig. 19: Interfaz del Portal de Transparencia Económica del MEF.

Para este proyecto, se consolidaron los registros de contratos desde el año 2018 hasta julio de 2025, considerando todas las adjudicaciones registradas en las plataformas mencionadas. Los archivos disponibles fueron descargados de manera anual y posteriormente unificados en un único conjunto de datos con un total de **365,341 registros**.

B. Construcción del Dataset

El dataset consolidado incluye tanto información contractual como atributos relacionados con las entidades y contratistas. Cada registro representa un contrato individual con sus respectivos montos, fechas y metadatos administrativos. En la Tabla VII se presenta un resumen de los principales atributos empleados en la investigación.

Además del dataset principal de contratos públicos, se integró un segundo conjunto de datos fundamental para la construcción de variables derivadas y para definir la variable objetivo: el Dataset de Penalidades. Este archivo contiene información sobre sanciones administrativas, penalidades económicas e incumplimientos registrados contra contratistas por diversas entidades del Estado peruano.

En la Tabla VIII se presenta el diccionario de datos del conjunto utilizado para esta integración.

TABLE VIII: Diccionario de datos del Dataset de Penalidades

Atributo	**Descripción**	**Tipo**
ID_CONTRATO	Identificador del contrato asociado a la penalidad	Número
RUC_CONTRATISTA	RUC del contratista sancionado	Texto
TIPO_PENALIDAD	Categoría o tipo de penalidad aplicada	Texto
OBJETO_CONTRATO	Objeto contractual vinculado a la penalidad	Texto
ENTIDAD_CONTRATANTE	Entidad estatal que emitió la penalidad	Texto
FECHA_PENALIDAD	Fecha oficial del documento sancionador	Fecha
DESCRIPCION_MOTIVO	Motivo o descripción resumida de la penalidad	Texto
MONTO	Monto económico de la sanción	Número

La incorporación de este dataset permitió **vincular cada contrato con el historial sancionatorio de su proveedor**. Para ello, se realizó un proceso de emparejamiento basado en el campo `ID_CONTRATO`, garantizando la correspondencia entre los registros contractuales y los eventos asociados en el dataset de penalidades.

A partir de esta integración, se generaron dos variables derivadas clave:

- **Frecuencia_Penalidades:** número total de sanciones asociadas a un contrato o contratista. Esta variable se obtuvo mediante el cruce del campo `ID_CONTRATO` entre los archivos `contratosunidos.xlsx` y `penalidades.csv`.
- **Anomalía:** variable binaria que indica la presencia de comportamientos irregulares en la adjudicación o ejecución del contrato. Un registro se clasifica como **1 (Anómalo)** cuando el contratista presenta uno o más

TABLE VII: Diccionario de datos del dataset de contratos públicos

ATRIBUTO	**DESCRIPCIÒN**	**TIPO DE ATRIBUTO**
CODIGOCONVOCATORIA	Código de la convocatoria	Número
N_COD_CONTRATO	Identificador del contrato	Número
DESCRIPCION_PROCESO	Descripción del proceso de selección	Texto
FECHA_PUBLICACION_CONTRATO	Fecha de publicación del contrato. Es campo registrado por la Entidad.	Fecha
FECHA_SUSCRIPCION_CONTRATO	Fecha de suscripción del contrato. Es campo registrado por la Entidad.	Fecha
FECHA_VIGENCIA_INICIAL	Fecha de inicio de la vigencia del contrato. Es campo registrado por la Entidad.	Fecha
FECHA_VIGENCIA_FINAL	Fecha de final de la vigencia del contrato. Es campo registrado por la Entidad.	Fecha
FECHA_VIGENCIA_FIN_ACTUALIZADA	Fecha de final de la vigencia actualizada del contrato. Es campo registrado por la Entidad.	Fecha
CODIGO_CONTRATO	Código del contrato	Número
NUM_CONTRATO	Número de contrato	Número
MONTO_CONTRATADO_TOTAL	Monto total contratado en moneda original. Dato registrado por la Entidad.	Número
MONTO_CONTRATADO_ITEM	Monto del ítem contratado en moneda original. Datos registrados por la Entidad.	Número
MONTO_ADICIONAL	Monto del adicional en moneda original, en caso el contrato tuviera un adicional.	Número
MONTO_REDUCCION	Monto de la reducción en moneda original, en caso el contrato tuviera una reducción.	Número
MONTO_PRORROGA	Monto de la prórroga en moneda original, en caso el contrato tuviera una prórroga.	Número
MONTO_COMPLEMENTARIO	Monto complementario en moneda original.	Número
URLCONTRATO	URL del archivo del contrato	Texto
CODIGOENTIDAD	Código de la entidad contratante	Texto
NUM_ITEM	Número del ítem	Número
MONEDA	Moneda del contrato	Texto
RUC_CONTRATISTA	RUC del contratista	Texto
RUC_DESTINATARIO_PAGO	RUC del destinatario de pago	Texto
TIENERESOLUCION	Indica si el contrato tiene resolución y/o nulidad	Número
FRECUENCIA_PENALIDADES	Cantidad de sanciones o penalidades asociadas a un contrato o licitación	Número
ANOMALIA	Indica si hubo irregularidades en la licitación o si el contratista fue sancionado	Número

eventos sancionatorios previos; y como **0 (No Anómalo)** en caso contrario.

La construcción de la variable *Anomalía* se basa en un supuesto ampliamente respaldado en la literatura de auditoría pública: *los proveedores con historial de sanciones tienen mayor probabilidad de incurrir en incumplimientos o prácticas irregulares*. Por ello, esta etiqueta se convierte en un insumo esencial para el entrenamiento de los modelos supervisados.

La integración del historial sancionatorio permite enriquecer el dataset con información relevante sobre la conducta pasada de los contratistas, facilitando la identificación temprana de patrones anómalos en los procesos de contratación pública.

C. Preprocesamiento y Selección de Atributos

El preprocesamiento de datos tiene como finalidad garantizar la consistencia y calidad de los registros antes del modelado, y las principales acciones a considerar fueron:

- **Validación de identificadores (RUC):** Se realizó un proceso de verificación entre las columnas *ruc_contratista* y *ruc_destinatario_pago* con el fin de detectar inconsistencias. Se identificaron 434 registros distintos entre ambos campos, los cuales fueron almacenados en un archivo separado para documentar las discrepancias, y de los 365,338 registros coincidentes, algunos valores en *ruc_contratista* no cumplían con la longitud estándar de 11 dígitos, por lo que se consideraron no válidos.

TABLE IX: Diccionario de datos del dataset final utilizado en el modelo

ATRIBUTO	DESCRIPCIÓN	TIPO DE ATRIBUTO
DESCRIPCION_PROCESO	Descripción del proceso de contratación o selección. Utilizada para análisis semántico mediante embeddings.	Texto
MONTO_CONTRATADO_TOTAL	Monto total contratado en moneda original. Representa el valor principal del contrato.	Número
MONTO_ADICIONAL	Monto adicional en moneda original, en caso de ampliaciones o adendas contractuales.	Número
MONTO_REDUCCION	Monto de la reducción en moneda original, en caso de modificaciones contractuales que disminuyan el valor.	Número
MONTO_PRORROGA	Monto correspondiente a prórrogas contractuales.	Número
MONTO_COMPLEMENTARIO	Monto complementario en moneda original asociado al contrato.	Número
RUC_DESTINATARIO_PAGO	RUC del destinatario de pago o proveedor responsable del contrato.	Texto
TIENERESOLUCION	Indica si el contrato presenta resolución o nulidad registrada.	Número
URLCONTRATO	Enlace al archivo PDF del contrato, utilizado para extracción de texto mediante OCR.	Texto
FRECUENCIA_PENALIDADES	Cantidad de sanciones o penalidades históricas asociadas al contratista o proceso.	Número
ANOMALIA	Variable binaria que indica si el contrato presenta irregularidades o antecedentes de sanciones.	Número
DURACION_CONTRATO_DIAS	Duración total del contrato expresada en días, calculada a partir de las fechas de vigencia.	Número

Por lo tanto, se decidió trabajar con la columna *ruc_destinatario_pago* como referencia principal, al presentar una estructura más consistente. Asimismo, se conservaron los registros anómalos en un archivo independiente, manteniendo la trazabilidad del proceso.
Además, se excluyeron los registros correspondientes a personas naturales, limitando el análisis a los RUC de empresas jurídicas, con el propósito de garantizar un cruce confiable con fuentes externas y de enfocarse en las entidades de mayor relevancia en la contratación pública.

- **Homogeneización de valores monetarios:** Se unificaron las diferentes denominaciones de moneda, convirtiendo todos los montos al equivalente en *soles peruanos (PEN)*, esta normalización permitió realizar comparaciones homogéneas y evitar distorsiones en las variables financieras.
- **Consolidación y derivación de fechas:** Durante el análisis se verificó que las columnas *fecha_vigencia_final* y *fecha_vigencia_fin_actualizada* contenían valores equivalentes, por lo que se mantuvo una sola variable representativa. A partir de *fecha_vigencia_inicial* y la fecha de finalización seleccionada, se generó una nueva variable denominada *duracion_contrato_dias*, que expresa la cantidad total de días de vigencia del contrato.
- **Eliminación de variables no relevantes:** Se suprimieron aquellas columnas que no aportaban información significativa al análisis o que presentaban alta redundancia, reduciendo así la dimensionalidad del dataset y optimizando el desempeño de los modelos posteriores.
- **Creación de variables derivadas:** Además de la variable de duración del contrato, se generaron nuevas características relevantes, tales como la *frecuencia_penalidades* (número total de penalidades o sanciones asociadas al contratista) y la variable binaria anomalia.
- **Herramientas utilizadas:** Todo el proceso de limpieza, integración y transformación de datos se desarrolló en *Python*, utilizando las librerías *pandas*, *numpy* y *datetime* para la manipulación de estructuras tabulares, el tratamiento de fechas y la validación de registros. Estas herramientas permitieron realizar de manera eficiente la detección de valores inválidos, el filtrado de inconsistencias y la creación del dataset final que serviría como insumo para los modelos de aprendizaje automático.

En resumen, el proceso de preprocesamiento y selección de atributos permitió obtener un conjunto de datos depurado, coherente y enriquecido, que preserva la integridad de la información original y garantiza la validez de los análisis posteriores. La elección de trabajar exclusivamente con RUC de empresas y de establecer una estructura de datos uniforme asegura una base sólida para la aplicación de técnicas de detección de anomalías.

Como resultado de todas las etapas de limpieza, integración, transformación y enriquecimiento de datos descritas previamente, se obtuvo el conjunto final de variables que fueron empleadas como insumo en los modelos de detección de anomalías.

Este dataset incorpora tanto atributos originales de los

contratos como nuevas variables derivadas, representando de manera estructurada la información relevante que describe cada proceso de contratación.

En la **Tabla IX** se detalla el *diccionario de datos del dataset final*, que resume los campos seleccionados y creados tras el análisis exploratorio, asegurando que cada atributo aporte valor explicativo al modelo.

D. Flujo Metodológico General del Proyecto

El proceso metodológico completo propuesto se compone de dos fases principales, las cuales integran el tratamiento de datos estructurados y no estructurados, así como el entrenamiento y la aplicación de modelos de aprendizaje automático para la detección de anomalías en contratos públicos.

En la Figura 22 se presenta la primera fase, orientada a la recolección, depuración y procesamiento de datos; mientras que la Figura 23 describe la segunda fase, correspondiente al modelado predictivo y la clasificación de nuevos contratos.

Fase 1: Recolección, Preprocesamiento e Ingeniería de Características

Esta fase tiene como objetivo construir un dataset robusto, integrando múltiples fuentes de información y transformando los datos en un formato adecuado para el modelado. El flujo se desarrolla mediante dos rutas paralelas: el procesamiento de datos estructurados y el procesamiento de datos no estructurados.

Ruta de Datos Estructurados: El proceso inicia con la consolidación de registros provenientes de bases institucionales y fuentes oficiales del SEACE. Posteriormente, se realiza una validación exhaustiva de los datos, que incluye la estandarización de valores monetarios, la conversión de tipos de datos, la homologación de formatos y la verificación de la validez de identificadores como el RUC. El objetivo es garantizar la coherencia y confiabilidad de los datos tabulares.

Luego se lleva a cabo la ingeniería de características, en la cual se generan nuevas variables relevantes para el análisis. Entre estas destacan la variable *duracion_contrato_dias*, derivada de las fechas contractuales, y la variable *Frecuencia_Penalidades*, obtenida mediante el cruce directo entre el número de contrato y el registro histórico de sanciones. Finalmente, se construye la variable objetivo *anomalia*, que clasifica como observaciones anómalas aquellas asociadas a proveedores que presentan antecedentes de sanciones, inhabilitaciones o incumplimientos administrativos documentados por entidades supervisoras.

Ruta de Datos No Estructurados: En paralelo, se procesa el contenido textual de los contratos, compuesto principalmente por descripciones cortas (*descripcion_proceso*) y documentos completos en formato PDF. Para la columna *descripcion_proceso*, se aplica un proceso de vectorización mediante *embeddings* utilizando el modelo *paraphrase-multilingual-MiniLM-L12-v2*, perteneciente a la familia Sentence-BERT (SBERT). Dicho modelo, basado en la arquitectura Transformer, genera representaciones densas capaces de capturar relaciones semánticas profundas, incluso en textos de múltiples idiomas como el español.

El procesamiento de los contratos completos se realiza mediante la descarga automática de los archivos PDF utilizando técnicas de automatización con Selenium WebDriver. Una vez descargados, se aplica un pipeline de reconocimiento óptico de caracteres (OCR) empleando *pdf2image* y *Pytesseract*. Este enfoque permite extraer texto incluso de documentos escaneados o sin contenido digital embebido. El texto resultante se almacena en la variable *texto_contrato*.

Posteriormente, tanto el contenido de *descripcion_proceso* como el texto extraído mediante OCR son transformados en vectores semánticos utilizando nuevamente el modelo *paraphrase-multilingual-MiniLM-L12-v2*. Estos embeddings representan cada contrato como un vector denso de alta dimensionalidad (384 características), preservando su estructura semántica.

Finalmente, los embeddings del texto completo se utilizan como entrada para el algoritmo de *clustering* K-Means. Este algoritmo permite agrupar los contratos según su similitud semántica, generando así la variable *cluster_texto*, la cual resume la temática predominante de cada documento y aporta una dimensión adicional útil para la posterior fase de modelado. Para ello se buscó el mejor K, para el cluster.

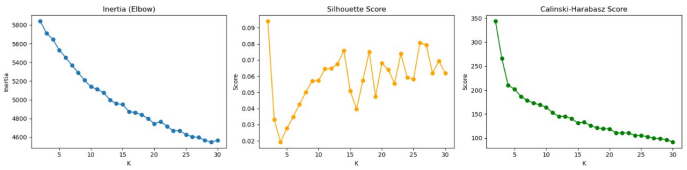

Fig. 20: Evaluación de distintos valores de K mediante los criterios de Inercia (Elbow), Silhouette y Calinski-Harabasz.

Fase 2: Modelado Predictivo y Clasificación

La segunda fase abarca el entrenamiento, evaluación y aplicación del modelo encargado de detectar anomalías en nuevos contratos.

Consolidación y División del Dataset: El dataset final, que integra tanto las variables estructuradas como las generadas mediante procesamiento de texto (*cluster_texto* y los *embeddings*), se divide en dos subconjuntos: 80% para entrenamiento y 20% para prueba.

Balanceo del Dataset de Entrenamiento: El análisis exploratorio de la variable objetivo *anomalia* evidenció una distribución moderadamente desbalanceada entre las dos clases, donde el 74.33% de los registros correspondieron a contratos clasificados como *no anómalos* (clase 0), mientras que el 25.67% representaron casos *anómalos* (clase 1), como se muestra en la figura 20. Si bien la clase minoritaria posee un tamaño considerable, su menor proporción podría afectar el desempeño de los modelos supervisados, los cuales tienden a favorecer la clase mayoritaria durante el proceso de aprendizaje.

Para mitigar este sesgo y mejorar la capacidad del modelo para identificar adecuadamente las anomalías, se aplicó un proceso de balanceo exclusivamente sobre el conjunto de entrenamiento. Este balanceo se realizó mediante la técnica

de *undersampling* controlado, reduciendo parcialmente la clase mayoritaria sin comprometer su representatividad estadística. De esta manera, el modelo recibe una señal más equilibrada durante el entrenamiento, fortaleciendo su habilidad para reconocer patrones asociados a comportamientos irregulares.

Cabe resaltar que el conjunto de prueba se mantuvo sin alteraciones con el fin de garantizar una evaluación objetiva del rendimiento del modelo en un escenario realista. En este contexto, la métrica *recall* adquiere especial relevancia, dado que permite cuantificar la capacidad del sistema para detectar correctamente los casos de anomalías, los cuales constituyen el principal foco de interés para la supervisión y el control de riesgos en la contratación pública.

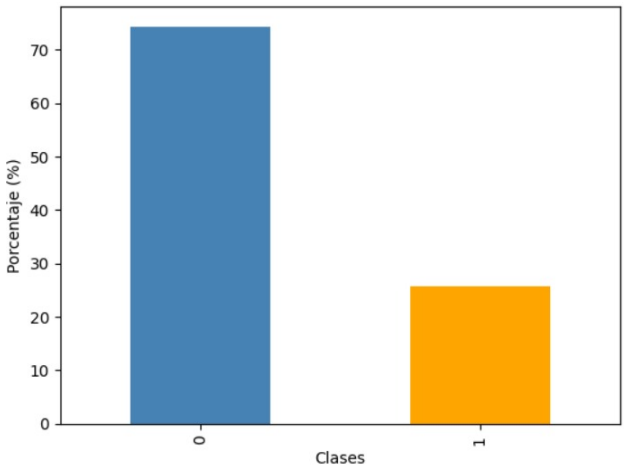

Fig. 21: Distribución de clases de la variable objetivo *anomalia*, donde la clase 0 representa contratos sin anomalías y la clase 1 corresponde a casos con antecedentes sancionatorios.

Entrenamiento de Modelos Supervisados: Se evalúa un conjunto de algoritmos de clasificación supervisada, incluyendo K-Nearest Neighbors, Regresión Logística, AdaBoost, Árboles de Decisión, Perceptrón Multicapa (MLP) y un modelo de red neuronal profunda implementado en TensorFlow. Cada algoritmo recibe la misma matriz de características, lo que permite comparar su desempeño de manera justa.

Evaluación y Selección del Mejor Modelo: Los modelos se comparan mediante métricas como precisión, *recall* y F1-score, siendo esta última la métrica principal para la selección del mejor clasificador debido al desbalance natural de la clase anómala. El modelo con el mayor F1-score es guardado en formato *pickle* para su uso en producción.

Predicción de Nuevos Contratos: Para la clasificación de contratos futuros, el registro pasa por el mismo pipeline de preprocesamiento definido en la Fase 1: extracción de texto, generación de embeddings, asignación de cluster semántico y construcción del vector de características. Finalmente, este vector se ingresa al modelo entrenado, obteniendo como resultado una predicción binaria que indica si el contrato puede considerarse anómalo o no.

V. Resultados

En esta sección se presentan los resultados obtenidos tras el entrenamiento y evaluación de los modelos supervisados utilizados para la detección de anomalías en contratos públicos. El conjunto de datos final fue dividido en un 80% para entrenamiento y un 20% para prueba, manteniendo el conjunto de prueba sin balancear para asegurar una evaluación realista.

Se evaluaron cuatro modelos principales: *Decision Tree*, *Logistic Regression*, *K-Nearest Neighbors (KNN)* y *AdaBoost*. Para cada modelo se calcularon las métricas *Accuracy*, *Precision*, *Recall* y *F1-score*, siendo esta última la métrica de referencia debido al desbalance natural de la variable objetivo.

La Tabla X resume los resultados obtenidos:

TABLE X: Rendimiento de los modelos supervisados en el conjunto de prueba

Modelo	Accuracy	Precision	Recall	F1-score
Decision Tree	0.7371	0.7350	0.7371	0.7360
Logistic Regression	0.7301	0.7607	0.7301	**0.7408**
K-Nearest Neighbors	0.7092	0.7704	0.7092	0.7256
AdaBoost	0.7330	0.7265	0.7330	0.7295

A. Análisis de los resultados

Los modelos presentan desempeños similares en términos generales, con valores de *accuracy* entre 0.70 y 0.74. Sin embargo, la métrica más relevante para el presente estudio es el **F1-score**, dado que la clase positiva (*anomalia*) representa la minoría y es la de mayor importancia para la supervisión contractual.

El modelo **Logistic Regression** obtuvo el mayor *F1-score* (0.7408), superando ligeramente al *Decision Tree* (0.7360) y a *AdaBoost* (0.7295). Asimismo, alcanzó la mayor precisión entre todos los modelos (0.7607), lo que indica una menor tasa de falsos positivos.

Aunque el modelo **KNN** presenta la mayor precisión (0.7704), su *recall* es el más bajo del conjunto (0.7092), lo cual puede resultar desfavorable en un contexto donde es crítico identificar la mayor cantidad de contratos irregulares.

B. Modelo seleccionado

Considerando la importancia de equilibrar precisión y sensibilidad en un entorno de riesgo institucional, se seleccionó **Logistic Regression** como el mejor modelo del estudio. Su rendimiento estable, su F1-score más alto y su bajo costo computacional lo convierten en una alternativa óptima para una eventual implementación en sistemas reales de auditoría digital.

En consecuencia, el modelo final fue guardado en formato *pickle* para su integración dentro del pipeline de predicción de anomalías en nuevos contratos públicos.

VI. Discusión

Los resultados obtenidos evidencian el potencial del aprendizaje automático para la detección temprana de anomalías en procesos de contratación pública. En la evaluación comparativa, los cuatro modelos supervisados presentaron desempeños

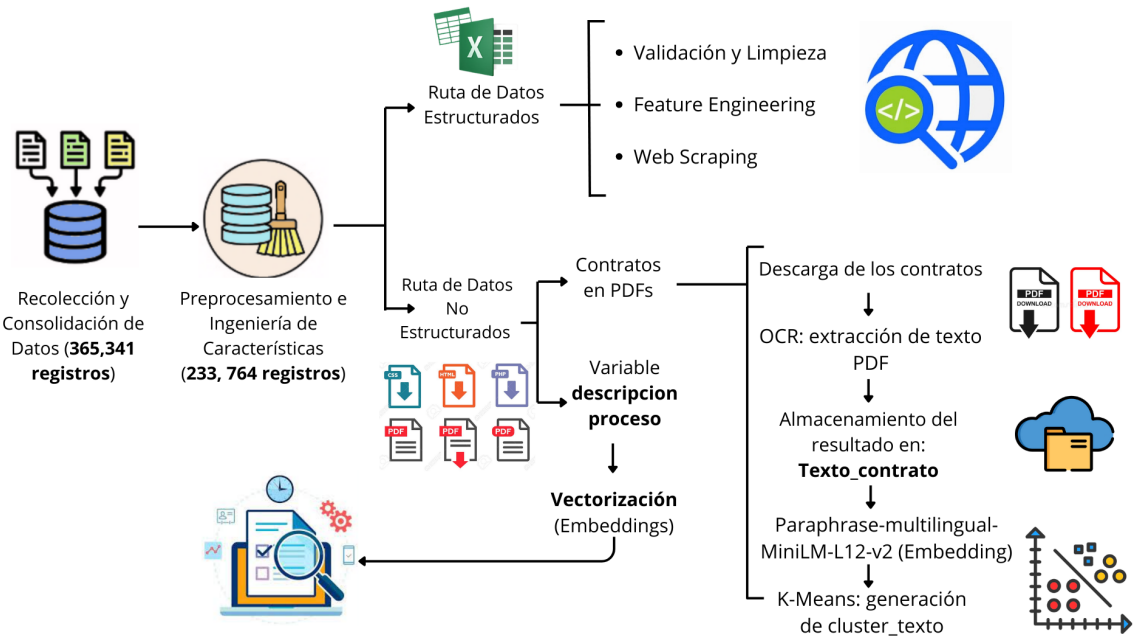

Fig. 22: Fase 1: Recolección, depuración y procesamiento de datos estructurados y no estructurados.

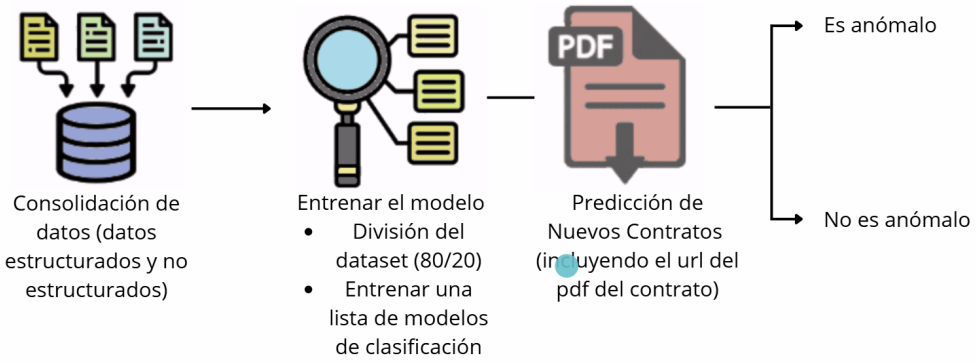

Fig. 23: Fase 2: Entrenamiento del modelo y detección de anomalías en nuevos contratos.

relativamente similares; sin embargo, Logistic Regression obtuvo el mayor *F1-score*, lo cual es particularmente relevante dada la naturaleza desbalanceada de la variable objetivo. Este comportamiento confirma que, en escenarios donde la clase positiva (anomalía) es minoritaria, las métricas tradicionales como *accuracy* no son suficientes para evaluar adecuadamente la capacidad discriminativa del modelo.

Un aspecto clave identificado en los experimentos es la influencia del procesamiento semántico del texto contractual. Los *embeddings* generados con *paraphrase-multilingual-MiniLM-L12-v2* permitieron capturar patrones lingüísticos asociados a posibles irregularidades, mejorando la representación de los documentos frente a técnicas tradicionales como TF-IDF. Asimismo, la incorporación del *clustering* temático mediante K-Means aportó una dimensión adicional que contribuyó a diferenciar contratos según su naturaleza semántica, enriqueciendo el conjunto final de características.

En términos de efectividad, se observó que los modelos basados en árboles y métodos de ensamble (Decision Tree y AdaBoost) mostraron una alta estabilidad en precisión y *recall*, aunque sin superar el desempeño global de Logistic Regression. Esto sugiere que las relaciones entre las variables

contractuales y la ocurrencia de anomalías presentan patrones parcialmente lineales que este último modelo logra capturar de manera más eficiente.

Por otro lado, la tasa moderada de *recall* en algunos algoritmos indica que todavía existen casos anómalos difíciles de identificar, probablemente asociados a comportamientos irregulares no reflejados explícitamente en las variables estructuradas o textuales. Esto abre la puerta a futuros trabajos que integren fuentes adicionales de información, como redes de proveedores, alertas tempranas basadas en grafos o análisis de documentos completos mediante modelos OCR-free como Donut.

En conjunto, los hallazgos señalan que el enfoque propuesto no solo es viable, sino que también constituye una base sólida para la automatización de auditorías públicas. La combinación de datos estructurados, embeddings semánticos y modelos supervisados genera un sistema capaz de identificar patrones de riesgo con mayor precisión que los métodos tradicionales, contribuyendo al fortalecimiento de la transparencia y la toma de decisiones basada en evidencia dentro del sistema de contratación estatal.

VII. Conclusiones

El presente estudio demuestra que es posible desarrollar un sistema automatizado y eficaz para la detección de anomalías en los contratos públicos del Estado peruano mediante el uso combinado de aprendizaje automático, procesamiento de lenguaje natural y análisis semántico. La integración de datos estructurados provenientes del SEACE con información adicional sobre sanciones administrativas permitió construir un dataset robusto de 365,341 contratos, enriquecido con variables derivadas como *frecuencia_penalidades* y la etiqueta binaria *anomalia*, fundamentales para el modelado supervisado.

Los resultados obtenidos evidencian que los modelos supervisados evaluados alcanzan niveles consistentes de desempeño, con valores de *accuracy* entre 0.70 y 0.74. Entre ellos, **Logistic Regression** destacó al obtener el mayor *F1-score* (0.7408), convirtiéndose en la alternativa más adecuada para contextos donde la identificación de casos irregulares es prioritaria. Este hallazgo confirma que, en situaciones de desbalance de clases, las métricas centradas en la capacidad de detección, como el F1-score, son más representativas que las métricas globales.

Asimismo, el uso de *embeddings* semánticos mediante el modelo *paraphrase-multilingual-MiniLM-L12-v2* permitió capturar patrones lingüísticos presentes en los textos contractuales y mejorar la representación de la información no estructurada. La incorporación del *clustering* temático mediante K-Means aportó una dimensión adicional que enriqueció el conjunto final de características, contribuyendo al desempeño general del sistema.

Las técnicas aplicadas de limpieza, validación de identificadores, estandarización monetaria y generación de nuevas características resultaron cruciales para asegurar la coherencia y calidad del dataset final. De igual forma, la automatización del proceso de extracción OCR y la transformación semántica de los documentos fortalecieron el pipeline metodológico propuesto.

El sistema desarrollado constituye un aporte significativo para la auditoría digital en el sector público, ya que permite identificar contratos con mayor probabilidad de presentar riesgos o irregularidades, facilitando la asignación de recursos de fiscalización. Los resultados obtenidos evidencian el potencial del aprendizaje automático como herramienta de apoyo para mejorar la transparencia, la supervisión y la lucha contra la corrupción en la administración pública.

Finalmente, se recomiendan líneas futuras de investigación orientadas a la incorporación de modelos OCR-free más avanzados, como Donut, el uso de grafos para detectar redes de colusión entre proveedores, y la integración de variables externas que fortalezcan la capacidad predictiva del sistema. Estas mejoras permitirían avanzar hacia soluciones más precisas, escalables y aplicables a diferentes dominios de la gestión estatal.

Agradecimientos

Los autores, Beatriz Bravo y Jefferson Garay, expresan su sincero agradecimiento al profesor del curso de Machine Learning por su acompañamiento constante, su disciplina académica y su exigencia metodológica a lo largo del desarrollo de este proyecto. Su paciencia, claridad y compromiso con la formación técnica fueron fundamentales, especialmente considerando que esta es nuestra primera experiencia aplicando técnicas avanzadas de aprendizaje automático.

Asimismo, agradecemos a la Universidad ESAN, Perú, por brindar el entorno académico y los recursos necesarios que hicieron posible esta investigación. El curso ha representado una oportunidad valiosa para comprender el potencial de la inteligencia artificial en el análisis de datos públicos y en la mejora de la transparencia institucional.

Finalmente, agradecemos también a nuestras familias y compañeros por su apoyo y motivación durante la elaboración de este trabajo.

References

[1] A. Aldana, A. Falcón-Cortés, and H. Larralde, "A machine learning model to identify corruption in Mexico's public procurement contracts" *arXiv preprint arXiv:2104.05611*, 2021. [Online]. Available: https://arxiv.org/abs/2104.05611

[2] G. Kim, T. Hong, M. Yim, J. Nam, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "OCR-free Document Understanding Transformer (Donut)," in *Proceedings of the European Conference on Computer Vision (ECCV 2022)*, 2022. [Online]. Available: https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136880493.pdf

[3] Secretaría de la Función Pública de México, "Unidad de Política de Contrataciones Públicas (Compranet)," [Online]. Available: https://upcp-compranet.buengobierno.gob.mx/.[Accessed: Oct. 11, 2025].

[4] Servicio de Administración Tributaria (SAT), "Portal del SAT," [Online]. Available: https://www.sat.gob.mx/portal/public/home. [Accessed: Oct. 11, 2025].

[5] Gobierno de México, "Portal de Datos Abiertos del Gobierno de México," [Online]. Available: https://datos.gob.mx/. [Accessed: Oct. 11, 2025].

[6] A. Falcón-Cortés, A. Aldana, and H. Larralde, "Practices of public procurement and the risk of corrupt behavior before and after the government transition in Mexico," *EPJ Data Science*, vol. 11, no. 1, p. 19, 2022.

[7] R. Sánchez Fernández, "Predicción de la corrupción vía red neuronal," *Trabajo Fin de Grado en Ingeniería Informática*, Universidad Autónoma de Madrid, Escuela Politécnica Superior, Feb. 2017. [Online]. Available: https://repositorio.uam.es/handle/10486/ [Accessed: Oct. 12, 2025].

[8] A. Groll, A. Khanna, and L. Zeldin, "A Machine Learning-based Anomaly Detection Framework in Life Insurance Contracts," *arXiv preprint arXiv:2411.17495*, Nov. 2024. [Online]. Available: https://arxiv.org/abs/2411.17495 [Accessed: Oct. 12, 2025].

[9] E. Schneider dos Santos, M. M. dos Santos, M. Castro, and J. T. Carvalho, "Detection of fraud in public procurement using data-driven methods: a systematic mapping study," *EPJ Data Science*, vol. 14, no. 52, 2025. [Online]. Available: https://doi.org/10.1140/epjds/s13688-025-00569-3 [Accessed: Oct. 12, 2025].

[10] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[11] Mapa de Contratos Públicos. Disponible en: https://prod4.seace.gob.pe/contratos/publico/#/mapa. Última consulta: 2025.

[12] Portal de Transparencia Económica del MEF. Disponible en: https://apps5.mineco.gob.pe/proveedor/. Última consulta: 2025.

[13] R. E. Martínez Huamán, *La corrupción en el Perú: situación, respuestas y resultados*, Revista Oficial del Poder Judicial, vol. 15, no. 19, pp. 163–183, 2023. doi: 10.35292/ropj.015i19.719.

[14] J. F. Valdivia Sánchez and G. R. Bautista Quispe, *Corrupción en las contrataciones estatales y su impacto en la vulneración del derecho a la educación en Perú*, Revista INVE.COM, vol. 5, no. 2, 2025.

[15] Contraloría General de la República, *Nota de prensa: Corrupción e inconducta funcional habrían ocasionado pérdidas por S/ 24,268 millones en el 2023*, 2024. [Online]. Available: https://www.gob.pe/institucion/contraloria/noticias/912182. Accessed: Oct. 13, 2025.

[16] Contraloría General de la República, *Conferencia Anual Internacional por la Integridad (CAII) 2023*, Nota de Prensa N° 1404-2023-CG/GCOC, 2023.

[17] A. Aldana, A. Falcon-Corte, and H. Larralde, *A machine learning model to identify corruption in Mexico's public procurement contracts*, Journal of Data Science and Analytics, 2024.

[18] A. Groll, A. Khanna, and L. Zeldin, *A Machine Learning-based Anomaly Detection Framework in Life Insurance Contracts*, Insurance Analytics Journal, 2024.

[19] E. S. dos Santos, M. M. dos Santos, M. Castro, and J. T. Carvalho, "Detection of fraud in public procurement using data-driven methods: a systematic mapping study," *Government Information Quarterly*, 2025.

[20] World Bank. *Benchmarking Public Procurement 2018: Assessing Public Procurement Systems*. World Bank Publications, 2018. Available at: https://openknowledge.worldbank.org/entities/publication/c1c5b8ed-205c-5cc0-9427-743c1c970905

[21] Organisation for Economic Co-operation and Development (OECD). *Preventing Corruption in Public Procurement*. OECD Publishing, 2016. Available at: https://baselgovernance.org/sites/default/files/2020-03/oecd_preventing_corruption_in_public_procurement_2016.pdf

[22] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2012. Disponible en: https://www.elsevier.com/books/data-mining-concepts-and-techniques/han/978-0-12-381479-1

[23] Feldman, R. & Sanger, J., *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007. Disponible en: https://www.cambridge.org/core/books/text-mining-handbook/0634B1DF14259CB43FCCF28972AE4382

[24] Chandola, V., Banerjee, A., & Kumar, V. "Anomaly Detection: A Survey." *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009. https://dl.acm.org/doi/abs/10.1145/1541880.1541882

[25] A. Groll, A. Khanna, and L. Zeldin, *A Machine Learning-based Anomaly Detection Framework in Life Insurance Contracts*, Insurance Analytics Journal, 2024. https://www.jstor.org/stable/48842644

[26] Bishop, C. *Pattern Recognition and Machine Learning*. Springer, 2006.

[27] Breiman, L., Friedman, J., Stone, C., & Olshen, R. *Classification and Regression Trees*. Wadsworth, 1984.

[28] Freund, Y., & Schapire, R. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *Journal of Computer and System Sciences*, 1997.

[29] Mikolov, T., Chen, K., Corrado, G., & Dean, J. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*, 2013. https://arxiv.org/abs/1301.3781

[30] Reimers, N., & Gurevych, I. "Sentence-BERT: Sentence embeddings using Siamese BERT-networks." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. https://arxiv.org/abs/1908.10084