Revista Cubana de Ciencias Informáticas
Vol. 18, No. 3, Julio-Septiembre, 2024
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 16-32

Tipo de artículo: Artículo original
Temática: Inteligencia artificial

Detección de facturas sospechosas utilizando modelos no supervisados

Detection of suspicious invoices using unsupervised models

Liany Mendoza Cruz [1*] https://orcid.org/0009-0008-5391-6270
Vitali Herrera-Semenets [2] https://orcid.org/0000-0001-7094-2835

[1] Countigo S.R.L., calle 27 #508 Bajos entre calle E y calle F, Vedado, Plaza de la Revolución, La Habana, Cuba.
[2] Advanced Technologies Application Center (CENATAV). 7a #21406, Playa, C.P. 12200, Havana, Cuba Cuba.

* Autor para la correspondencia. (lianymendoza90@gmail.com)

Revista Cubana de Ciencias Informáticas
Vol. 18, No. 3, Julio-Septiembre, 2024
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 16-32

RESUMEN

El fraude y la evasión fiscal son delitos globales que afectan negativamente a la sociedad. El empleo de las nuevas tecnologías junto a la experticia de especialistas sobre dichos delitos puede producir resultados satisfactorios en su detección y enfrentamiento. Una forma de abordar este problema es mediante el uso de técnicas de Machine Learning (ML) con modelos no supervisados, que permitan a los auditores identificar posibles casos de fraude sin conocimiento previo del dominio. Por tanto, se decide investigar los posibles modelos no supervisados que puedan detectar posibles casos de fraude, por lo que se analizaron ejercicios implementados utilizando diferentes algoritmos que den una respuesta eficiente y precisa. La investigación no solo se centrará en aplicar la detección de valores atípicos, también busca lograr que los auditores puedan confiar en los resultados utilizando mecanismos de explicación del sector financiero. El objetivo principal es encontrar eventos anómalos que puedan ser fraudulentos o no, para un posterior estudio y análisis, y así complementar la base de conocimiento de los auditores.

Palabras clave: Detección de fraude; evasión fiscal; Inteligencia Artificial; Machine Learning; algoritmos.

ABSTRACT

Tax fraud and evasion are global crimes that negatively affect society. The use of new technologies together with the expertise of specialists on these crimes can produce satisfactory results in their detection and confrontation. One way to address this problem is through the use of Machine Learning (ML) techniques with unsupervised models, which allow auditors to identify possible cases of fraud without prior knowledge of the domain. Therefore, it was decided to investigate possible unsupervised models that can detect possible cases of fraud, which is why exercises implemented using different algorithms that provide an efficient and accurate response were analyzed. The research will not only focus on applying outlier detection, but also seeks to ensure that auditors can trust the results using explanation mechanisms from the financial sector. The main objective is to find anomalous events that may or may not be fraudulent, for subsequent study and analysis, and thus complement the auditors' knowledge base.

Revista Cubana de Ciencias Informáticas
Vol. 18, No. 3, Julio-Septiembre, 2024
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 16-32

Keywords: Fraud detection; tax evasion; Artificial Intelligence; Machine Learning; algorithms.

Recibido: 09/06/2024

Aceptado: 01/10/2024

Introducción

Los impuestos son el punto de referencia y el punto de inflexión del desarrollo general de un país, las prácticas de evasión fiscal son más graves en los países en desarrollo que cuando se comparan con los países desarrollados. La evasión fiscal es como una pandemia para los países ya que no son capaces de controlarla, el objetivo principal de cobrar los impuestos a los contribuyentes es mejorar el nivel de vida de los ciudadanos y asignar presupuesto para el gasto público, según las estadísticas, se estima que hay una pérdida del 20% de los ingresos por impuestos sobre la renta (Kassa, 2021). Según un informe de la Red de Justicia Fiscal, la Alianza Global por la Justicia Fiscal y la federación sindical mundial Internacional de Servicios Públicos, los gobiernos pierden cerca de 500 mil millones de dólares en ingresos fiscales al año debido al abuso fiscal global, principalmente debido a las corporaciones (McGoey, 2021).

El tema de fraude fiscal es una problemática global, y Cuba no queda exenta. La recaudación tributaria como porcentaje del PIB de Cuba disminuyó en 4.6 puntos porcentuales del 42.1% en 2019 al 37.5% en 2020 (OECD, 2022). Lo anterior pone de manifiesto una tendencia a la reducción de la recaudación tributaria, lo cual va asociado al aumento de eventos fraudulentos. En este contexto, la aplicación de técnicas de Machine Learning con un enfoque no supervisado plantea una alternativa prometedora. Sin embargo, la complejidad de la utilización de estos modelos está relacionada con la identificación precisa de facturas sospechosas en un conjunto de datos que puede contener una variedad de patrones de fraude y transacciones legítimas. La problemática radica en cómo desarrollar un modelo que pueda aprender automáticamente patrones sutiles y complejos asociados con las facturas sospechosas, sin requerir etiquetas previas de clasificación. Además, se

debe abordar cómo evaluar y validar la efectividad de dicho modelo en comparación con los enfoques convencionales (Gavoille et al., 2022).

Métodos o Metodología Computacional

Para el desarrollo de la investigación se utilizaron varios métodos científicos, como el Análisis-Síntesis para el estudio de las fuentes bibliográficas relacionados con los temas de fraude financiero y modelos de Machine Learning, la Observación para estudiar el comportamiento de los modelos utilizados con los datos proporcionados, la Medición para obtener las métricas para la eficacia y la interpretabilidad de los resultados dados por los diferentes modelos y el Experimental para realizar la comparación de los diferentes algoritmos (Wei R et al., 2019).

Técnicas de Aprendizaje Automático

La revisión de la literatura provee diferentes tipos de técnicas de aprendizaje automático no supervisado para la detección de fraudes en datos financieros, por lo que se realizará un resumen de alguna de las técnicas más mencionadas por la comunidad científica.

- **Árboles de Decisión** (*Decisions Tree*): Es Robusto y simple para entender e interpretar. Tiene un manejo eficiente de datos numéricos y categóricos, y brinda buenos resultados con gran volumen de datos. (Ali et al., 2022). Presenta sobreajuste especialmente cuando el número de características predictivas es alto y pequeños cambios en los datos de entrada pueden suponer un árbol de decisión completamente diferente. Tiene una probabilidad alta de sesgo hacia aquellas características con mayores valores numéricos. (Tang, 2020)
- **Agrupamiento** *(Clustering)*: Agrupa instancias similares y resuelve problemas complejos. Permite lograr resultados a largo plazo mediante el aprendizaje por refuerzo y se adaptan a diferentes medidas de similitud y niveles de compactación y separación entre los grupos. Presenta mayor riesgo de resultados inexactos y requieren una complejidad computacional elevada debido al gran volumen de

datos de entrenamiento. Como desventaja son sensibles: a los valores iniciales, la función del kernel, el umbral de distancia y la dimensión de los datos, también tienen dificultades para adaptarse a los cambios en los datos (Vanhoeyveld et al., 2020).

- **Reducción de Dimensionalidad** *(Dimensionality Reduction):* Son muy útiles para trabajar con una gran cantidad de datos y disminuye la redundancia requiriendo menos recursos para procesar los datos. Muy útil para la visualización de los datos y requiere menos recursos computacionales por lo que aumenta el rendimiento general de los algoritmos. Elimina la correlación entre variables lo que aumenta el rendimiento del algoritmo (Vanhoeyveld et al., 2020). Debido a la reducción de la dimensionalidad se pueden perder algunos datos y provocar una pérdida de interpretabilidad de la información. Posibilita el aumento del riesgo de sobreajuste si se reduce demasiado la dimensionalidad.

Árboles de Decisión

Isolation Forest: Algoritmo de detección de anomalías no supervisado que se utiliza para identificar observaciones inusuales en un conjunto de datos. Su funcionamiento se basa en la construcción de múltiples árboles de aislamiento, donde cada árbol divide aleatoriamente los datos en ramas. Calcula un puntaje de anomalía para cada observación, midiendo cuántos pasos son necesarios para aislarla en los árboles. Las ventajas de Isolation Forest incluyen su eficiencia computacional, robustez ante valores atípicos y capacidad para manejar distribuciones multimodales. Sin embargo, tiene como desventaja detectar anomalías locales y reducción de la eficacia con la alta dimensionalidad (Mensi et al.,2021).

Agrupamiento

K-means: Se enfoca en un agrupamiento sencillo y rápido basado en la distancia. El algoritmo reúne principalmente k muestras en K grupos según la similitud. Como ventajas podemos acotar que es un algoritmo simple, eficiente y fácil de implementar, pero la necesidad de especificar el número de clústeres (k) de antemano, puede ser un problema en algunos casos, y es válido destacar que no funciona bien cuando los clústeres tienen formas no esféricas o tamaños muy diferentes. También se puede mencionar como desventaja

que es un algoritmo sensible a los valores atípicos y que no garantiza convergencia global, lo que significa que la solución encontrada puede no ser la mejor posible. (Tang et al., 2020)

Spectral Clustering: Algoritmo de agrupamiento espectral conocido por su capacidad para representar grupos de diversas formas y densidades, lo que lo hace aplicable a conjuntos de datos de alta dimensión y capaz de manejar variables categóricas, es muy útil para el análisis de conglomerados, particularmente en el contexto de conjuntos de datos dispersos. Su algoritmización incluye la creación de una matriz de similitud de objetos, el cálculo de una matriz laplaciana, el cálculo de vectores propios y valores propios y, finalmente, la agrupación en el espacio de vectores propios seleccionados. Como desventaja se puede decir que la interpretación de los resultados puede ser muy compleja y su rendimiento puede ser bajo cuando tiene una alta dimensionalidad (Starosta et al., 2023).

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Algoritmo común de agrupamiento no supervisado que puede lograr clústeres al encontrar áreas de alta densidad separadas por áreas de baja densidad. A diferencia de otros métodos de agrupamiento, DBSCAN puede funcionar bien para clústeres de cualquier forma y puede identificar datos excepcionales de manera efectiva. Este algoritmo requiere la preconfiguración de dos parámetros, épsilon (EPS) y MinPts, para evaluar la distribución de densidad de los puntos en el espacio. Una de las principales desventajas de DBSCAN es la necesidad de configurar precisamente estos parámetros para cada conjunto de datos específico con el fin de lograr resultados óptimos de agrupamiento. Este proceso de ajuste de parámetros puede limitar la aplicabilidad de DBSCAN en diferentes escenarios (Yang et al., 2022).

OPTICS (Ordening Points To Identify the Clustering Structure): Algoritmo de agrupamiento que se basa en la idea de ordenar los puntos de datos según su densidad alcanzable. A diferencia de DBSCAN, OPTICS no requiere la configuración de parámetros fijos como EPS y MinPts, lo que lo hace más flexible y adaptable a conjuntos de datos con diferentes densidades y formas de clusterización. OPTICS puede identificar clústeres de diferentes formas y tamaños, así como puntos de ruido, lo que lo convierte en una opción robusta para el análisis de datos (Yang et al., 2022). Este modelo presenta algunas limitaciones como la dependencia de los

Revista Cubana de Ciencias Informáticas
Vol. 18, No. 3, Julio-Septiembre, 2024
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 16-32

parámetros min_samples y xi, el costo computacional es alto por lo que puede requerir más recursos computacionales, la interpretación de los resultados puede ser muy compleja y puede tener dificultades en conjuntos de datos de alta dimensionalidad o con una gran cantidad de muestras (Hajihosseinlou et al., 2024).

Reducción de Dimensionalidad

El Análisis de Componentes Principales (PCA) es una técnica que identifica la base más significativa de los datos y la transforma en otro conjunto de datos con menor dimensión. Esta nueva base revele la estructura oculta en el conjunto de datos y filtra el ruido que esta puede contener. Sus principales aplicaciones son: la reducción de dimensionalidad, compresión de datos, extracción de características y visualización de datos. Como desventaja podemos mencionar que puede causar pérdida de información y dificultad para la interpretación de los datos. Es un modelo dependiente de la escala de las variables, por lo que es importante normalizar o estandarizar los datos antes de aplicar esta técnica. También es válido mencionar que es un modelo que se centra en la varianza de los datos y no tiene en cuenta la información contextual o el significado de las variables (Kurita, 2020).

Métricas para la evaluación de modelos de aprendizaje automático no supervisado

La selección de las métricas adecuadas para la validación de los modelos es fundamental, ya que algunas pueden dar una falsa expectativa sobre el rendimiento obtenido. Por consiguiente, es importante utilizar diferentes métricas para validar los modelos. A continuación, se describen brevemente las métricas a utilizar en la presente investigación:

1. Silhouette Score: Esta métrica evalúa la cohesión y la separación entre los grupos. Un valor cercano a 1 indica que los puntos dentro de los grupos están bien agrupados y separados de otros grupos, lo que podría ser un indicativo de un buen agrupamiento. (Vanhoeyveld et al., 2020)

$$\text{silhouette\_score} = \frac{1}{N}\sum_{I=1}^{N}\frac{\text{b(i)}-\text{a(i)}}{\max\{\text{a(i)},\text{b(i)}\}} \quad (1)$$

2. Homogeneidad: Se centra en la similitud de los elementos dentro de un mismo clúster en relación con

Revista Cubana de Ciencias Informáticas
Vol. 18, No. 3, Julio-Septiembre, 2024
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 16-32

las etiquetas verdaderas. Una puntuación de 1 indica una homogeneidad perfecta, lo que significa que cada clúster contiene solo muestras de una sola clase. Por otro lado, una puntuación más baja indica una mezcla de clases dentro de los clústeres. (Vanhoeyveld et al., 2020)

$$H(y \mid c) = 1 - \frac{H(y|c)}{H(y)} \quad (2)$$

3. Visualización de componentes principales: Es utilizado comúnmente para visualizar datos de alta dimensionalidad en un espacio de dimensiones reducidas. Se puede evaluar la calidad de la visualización observando si los grupos de datos se separan de manera significativa en el espacio de los componentes principales.

Resultados y discusión

En este apartado se mostrará los resultados obtenidos mediante la experimentación de diferentes algoritmos. Para el entrenamiento de los modelos se utilizaron los siguientes conjuntos de datos obtenidos de la plataforma Kaggle:

Tabla 1 - Información sobre los conjuntos de datos.

Entradas	**Características**	**Etiquetas**
https://www.kaggle.com/datasets/mathchi/online-retail-ii-data-set-from-ml-repository/data?select=Year+2009-2010.csv		
2928	8	-
https://www.kaggle.com/datasets/gopalmahadevan/fraud-detection-example		
101613	11	Legítimo: 101497 Fraude: 116

En primera instancia se entrenó el algoritmo Isolation Forest para el conjunto de datos con 2928 entradas. Antes del entrenamiento del modelo se aplicó la ingeniería de características y se utilizó basamentos financieros como la Ley de Benford, la cual establece que, en muchos conjuntos de datos del mundo real, la frecuencia de aparición del primer dígito (es decir, el dígito más a la izquierda) sigue un patrón específico y predecible. Según dicha ley, el dígito "1" aparecerá con mayor frecuencia que cualquier otro dígito (Iosifidou and E. M, 2023). A continuación, se presentan las columnas originales del conjunto de datos:

- Invoice: Identificador de la factura.
- Customer ID: Identificador del cliente.
- Country: País emisor de la factura.
- StockCode: Código del producto facturado.
- Description: Descripción del producto facturado.
- Quantity: Cantidad del producto facturado.
- InvoiceDate: Fecha de emisión.
- Price: Precio del producto.

Después de un análisis de las características existentes, la primera acción sobre el conjunto de datos fue la limpieza de datos nulos, duplicados e inválidos. Posterior a esto, se creó una nueva columna con el precio total obtenido de la multiplicación del precio y la cantidad del producto facturado, por lo que se eliminaron dichas columnas del conjunto de datos. También se requirió transformar la característica País, codificándola en base a la frecuencia de ocurrencia en el conjunto de datos, ya que el Isolation Forest, funciona mejor con características numéricas en lugar de características categóricas o nominales.

Teniendo en cuenta los enfoques tradicionales donde se observan los comportamientos de los clientes, por ejemplo: la inactividad de una determinada empresa, la fecha y hora de la facturación, la ausencia de datos importantes en la documentación, entre otras, se procedió a construir otras características que pudieran ser relevantes para la detección de anomalías (Hamelers et al., 2021). Se utilizó PCA para la reducción de la

dimensionalidad (Ikeda et al., 2021) y este demostró que las nuevas características mantienen una varianza acumulada y poca correlación entre las variables, ver figura 1.

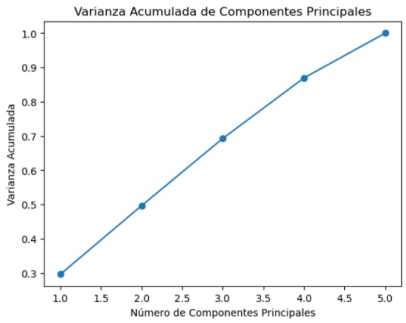

Fig. 1 - Aplicación de PCA.

Finalmente, solo permanecieron las siguientes características:

- Country_Freq: Frecuencia de ocurrencia del país emisor de la factura.
- Amount: Importe total de las facturas emitidas al mismo cliente, con la misma fecha, el mismo identificador y el mismo país.
- QtyTransacctions: Cantidad de transacciones por el mismo cliente, donde se representa que si la cantidad es mayor que 100 el valor es 1 sino es 0.
- Inactive: Tiempo de inactividad del cliente expresado en días, donde se representa que si la cantidad de día de inactividad es mayor que 300 el valor es 1 sino es 0.
- BenfordAmount: Característica binaria que representa que si el importe cumple la Ley de Benford el valor es 1 sino es 0.

Después del preprocesamiento de los datos y el análisis exploratorio de los mismos, se procedió a entrenar el modelo con un nivel de contaminación 0.01, teniendo en cuenta que los conjuntos de datos de facturación son altamente desbalanceados y en el contexto de la detección de fraudes, el nivel de contaminación puede ser interpretado como el porcentaje de transacciones fraudulentas en el conjunto de datos. El modelo reportó como resultado final 191 entradas anómalas.

Revista Cubana de Ciencias Informáticas
Vol. 18, No. 3, Julio-Septiembre, 2024
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 16-32

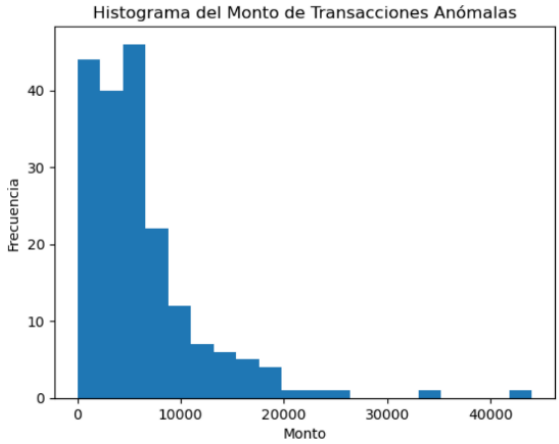

Fig. 2 - Monto de Transacciones Anómalas.

La figura 2 muestra la relación del monto facturado con la frecuencia de entradas existente en el conjunto de datos, lo que evidencia que la mayor cantidad de anomalías se encuentra en el rango de 0 a 10 mil pesos en más de 40 transacciones. A continuación, en la figura 3 también se puede observar que el Reino Unido (United Kingdom) es el país que más incidentes de posibles fraudes presenta.

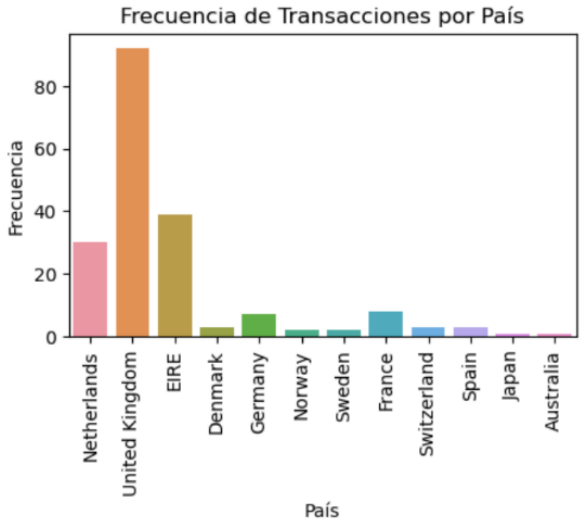

Fig. 3 - Transacciones anómalas por país.

Revista Cubana de Ciencias Informáticas
Vol. 18, No. 3, Julio-Septiembre, 2024
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 16-32

Se realizó otro caso de estudio con un conjunto de datos de 101613 entradas y 8 características utilizando el modelo DBSCAN y PCA para la visualización de los resultados. Antes del entrenamiento del modelo se dividió el conjunto en un 70% para entrenamiento y un 30% para prueba. Después se aplicó el algoritmo DBSCAN probando diferentes hiperparámetros para descubrir cuál es el valor óptimo de épsilon y seleccionar el valor mínimo de ejemplares. La tabla 2 muestra los resultados obtenidos:

Tabla 2 - Resultados obtenidos con diferentes hiperparámetros.

eps	min_samples	Coeficiente de la Silueta	Clústeres	Puntos atípicos	Homogeniedad
0.2	3	0.773	7	11	0.155
0.2	5	0.818	6	25	0.150
0.3	**3**	**0.830**	**5**	**8**	**0.157**
0.3	**5**	**0.830**	**5**	**8**	**0.157**
0.5	3	0.822	5	1	0.146
0.5	5	0.82	5	2	0.144

Posteriormente se utilizó la técnica de reducción de dimensionalidad (PCA) para la visualización y análisis de los clústeres del conjunto de prueba, la figura 4 visualiza el agrupamiento y donde se encuentran ubicados las transacciones fraudulentas.

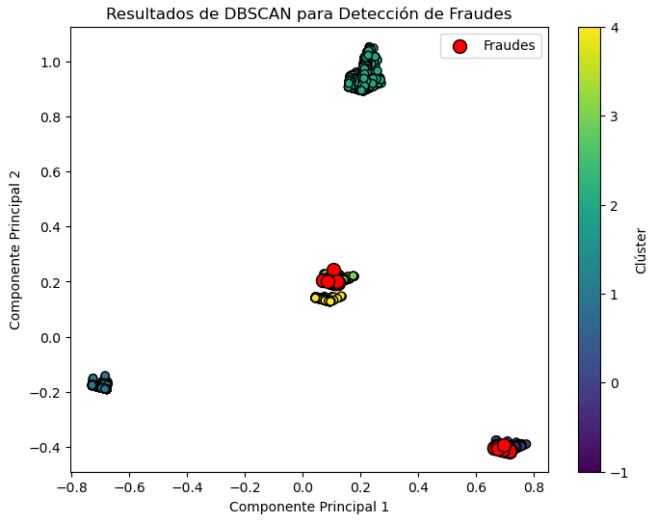

Fig. 4 - Visualización de Clústeres

Analizando los resultados obtenidos planteados en la tabla 2 y la visualización de los clústeres en la figura 4, podemos concluir que el valor de épsilon óptimo es 0.3 y el mejor valor mínimo de ejemplares es 3 o 5. Como se observa en la figura 4 las transacciones fraudulentas se encuentran agrupadas, aun estando con transacciones genuinas este resultado puede prevenir o detectar fraudes, ya que el objetivo principal es la detección, aunque el agrupamiento incluya las transacciones no fraudulentas. Por otra parte, se analizaron otros algoritmos como KMeans, Spectral Clustering y OPTICS, donde los resultados en conjunto no fueron tan eficaces (ver tablas 3 y 4), según la métrica del Coeficiente de la Silueta y la Homogeneidad:

Tabla 3 - Resultados obtenidos de KMeans y Spectral Clustering.

Algoritmo	Clústeres	Coeficiente de la Silueta	Homogeniedad
KMeans	5	0.832	0.135
	8	0.755	0.135
	10	0.686	0.157
	12	0.63	0.176
Spectral Clustering	5	-0.099	0.085
	8	-0.234	0.067
	10	0.243	0.079
	12	0.158	0.073

Tabla 4 - Resultados obtenidos de OPTICS.

min_samples	Coeficiente de la Silueta	Clústeres	Puntos atípicos	homogeneidad
5	-0.352	3051	36355	0.272
10	-0.549	936	45025	0.280
20	-0.686	280	50893	0.245
50	-0.320	49	35909	0.152
80	0.010	23	28183	0.131
100	0.016	19	34017	0.141

Los resultados obtenidos al aplicar los distintos algoritmos de agrupamiento sobre los conjuntos de datos de facturas arrojaron conclusiones valiosas en cuanto a la detección de anomalías y patrones sospechosos. El

modelo Isolation Forest demuestra ser un algoritmo eficaz para el análisis exploratorio de los datos, y combinándolo con otro modelo puede brindar excelentes resultados. La evaluación basada en métricas como el coeficiente de silueta en conjunto con la homogeneidad reveló que DBSCAN superó a otros algoritmos en términos de su desempeño, aunque la homogeneidad del modelo OPTICS fue superior, los valores del coeficiente y los puntos atípicos demuestra que no hay una correspondencia entre las métricas, ya que un mayor número de clústeres coadyuva a una mayor homogeneidad. DBSCAN al emplear un enfoque basado en densidades, mostró una capacidad notable para identificar clústeres de formas y tamaños variables, lo que resultó en una mejor cohesión en la distribución de los datos. Analizando los valores de las métricas se puede notar que no son del todo satisfactoria, pero cabe destacar que el agrupamiento coincide con el etiquetado de los datos, a pesar de que en el mismo grupo se encuentran transacciones genuinas. Con este resultado el auditor podrá observar el comportamiento de los clústeres potenciales que agrupan las transacciones fraudulentas, y así su revisión será hacia una menor cantidad de facturas.

Conclusiones

En este artículo, se abordó la detección de facturas sospechosas utilizando modelos no supervisados, se presentaron los resultados de la experimentación con el algoritmo Isolation Forest en un conjunto de datos de facturación, donde la Ley de Benford y los comportamientos comunes de los clientes fraudulentos fueron cruciales para la ingeniería de características. Además, se destacó la necesidad de entrenar modelos capaces como DBSCAN para el agrupamiento de los datos, con el objetivo de proporcionar una herramienta valiosa a los auditores forenses para que identifique los posibles casos fraudulentos. El uso de conjuntos públicos permitió entrenar los diferentes modelos que se utilizaron con el fin de validarlos con datos reales. Se emplearon diversos métodos de investigación, como el análisis-síntesis, observación, medición y experimentación, identificado así los valores óptimos para los parámetros claves en la detección de clústeres.

Referencias

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature

Hajihosseinlou, M., Maghsoudi, A., & Ghezelbash, R. (2024). A comprehensive evaluation of OPTICS, GMM and K-means clustering methodologies for geochemical anomaly detection connected with sample catchment basins. *Geochemistry*, 126094. https://doi.org/10.1016/J.CHEMER.2024.126094

Gavoille, N., Riga, S., & Zasova, A. (2022). FREE POLICY NETWORK BRIEF SERIES #AcademicsStandWithUkraine Detecting Labor Tax Evasion Using Administrative Data and Machine-Learning Techniques.

Hamelers, L., Rijk Marije den Ouden Vincent Gorka, A., Poel, M., ir van Capelleveen, E. G., & ProfDr van Hillegersberg, B. J. (2021). *Detecting and Explaining Potential Financial Fraud Cases in Invoice Data with Machine Learning*.

Ikeda, C., Ouazzane, K., Yu, Q., & Hubenova, S. (2021). New Feature Engineering Framework for Deep Learning in Financial Fraud Detection. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 12, Issue 12). www.ijacsa.thesai.org

Iosifidou, E. M. (2023). Financial Fraud Detection. http://dspace.lib.uom.gr/hadle/2159/28901

Kassa, E. T. (2021). Factors influencing taxpayers to engage in tax evasion: evidence from Woldia City administration micro, small, and large enterprise taxpayers. Journal of Innovation and Entrepreneurship, 10(1). https://doi.org/10.1186/s13731-020-00142-4

Kurita, T. (2020). Principal Component Analysis (PCA). *Computer Vision*, 1–4. https://doi.org/10.1007/978-3-030-03243-2_649-1

McGoey, S. (2021, November 19). Nearly $500 billion lost yearly to global tax abuse due mostly to corporations, new analysis says - ICIJ. https://www.icij.org/inside-icij/2021/11/nearly-500-billion-lost-yearly-to-global-tax-abuse-due-mostly-to-corporations-new-analysis-says/

Mensi, A., Franzoni, A., Tax, D. M. J., & Bicego, M. (2021). An Alternative Exploitation of Isolation Forests for Outlier Detection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial*

Intelligence and Lecture Notes in Bioinformatics), *12644 LNCS*, 34–44. https://doi.org/10.1007/978-3-030-73973-7_4/COVER

OECD. (2022, June 4). Estadísticas tributarias ALC: principales resultados para Cuba. https://www.oecd.org/tax/tax-policy/estadisticas-tributarias-america-latina-caribe-cuba.pdf

Starosta, B., Kłopotek, M. A., & Wierzchoń, S. T. (2023). *Explainable Graph Spectral Clustering of Text Documents*. http://arxiv.org/abs/2308.00504

Tang, P., Qiu, W., Huang, Z., Chen, S., Yan, M., Lian, H., & Li, Z. (2020). Anomaly detection in electronic invoice systems based on machine learning. *Information Sciences*, *535*, 172–186. https://doi.org/10.1016/j.ins.2020.03.089

Vanhoeyveld, J., Martens, D., & Peeters, B. (2020). Value-added tax fraud detection with scalable anomaly detection techniques. *Applied Soft Computing Journal*, *86*. https://doi.org/10.1016/j.asoc.2019.105895

Wei R, Dong B, Zheng Q, Zhu X, Ruan J, He H (2019) Unsupervised conditional adversarial networks for tax evasion detection. In: 2019 IEEE International Conference on Big Data (Big Data), pp 1675-1680, DOI 10.1109/BigData47090.2019.9005656

Yang, Y., Qian, C., Li, H., Gao, Y., Wu, J., Liu, C. J., & Zhao, S. (2022). An efficient DBSCAN optimized by arithmetic optimization algorithm with opposition-based learning. *Journal of Supercomputing*, *78*(18), 19566–19604. https://doi.org/10.1007/s11227-022-04634-w

Conflicto de interés

El autor autoriza la distribución y uso de su artículo.

Contribuciones de los autores

Conceptualización: Liany Mendoza Cruz, Vitali Herrera-Semenets

Análisis formal: Liany Mendoza Cruz

Investigación: Liany Mendoza Cruz

Metodología: Liany Mendoza Cruz

Software: Liany Mendoza Cruz

Visualización: Liany Mendoza Cruz

Redacción – borrador original: Liany Mendoza Cruz

Redacción – revisión y edición: Vitali Herrera-Semenets