Salud, Ciencia y Tecnología. 2024; 4:.1358
doi: 10.56294/saludcyt2024.1358

ORIGINAL

Model based on Machine Learning for the classification of banking transactions carried out through PSE

Modelo basado en Machine Learning para la clasificación de transacciones bancarias realizadas a través de PSE

Fabio Alberto Vargas Agudelo[1], Dario Enrique Soto Duran[1], Mauricio Urrego Álvarez[1], Edison Javier Yepes Sanchez[1], Iván Andrés Delgado González[2]

[1]Tecnológico de Antioquia – Institución Universitaria, Antioquia. Medellín, Colombia.
[2]Fundación Universitaria Juan de Castellanos, Boyacá. Tunja, Colombia.

Citar como: Vargas Agudelo FA, Soto Duran DE, Urrego Álvarez M, Yepes Sanchez EJ, Delgado González IA. Model based on Machine Learning for the classification of banking transactions carried out through PSE. Salud, Ciencia y Tecnología. 2024; 4:.1358. https://doi.org/10.56294/saludcyt2024.1358

Enviado: 08-03-2024 **Revisado:** 01-06-2024 **Aceptado:** 26-08-2024 **Publicado:** 26-08-2024

Editor: Dr. William Castillo-González

ABSTRACT

The financial sector, and specifically banking entities, have experienced changes in recent years thanks to technology, such as the digitization of transactions and the creation of applications such as digital wallets and PFM (Personal Finance Manager), generating gigabytes of information. Managing knowledge becomes essential to face new competitors, provide better services, understand the financial behavior of clients and face great challenges when processing and analyzing the volume of information available, which in most cases requires a complex preprocessing process and data quality. This is the case of banking transactions, which include free text information in their observation fields, making analysis and classification difficult, preventing the bank and its clients from analyzing financial behavior over a period of time. To solve this problem, the use of Machine Learning techniques was proposed to automate the transaction classification process based on text written in natural language, and provide the information that allows an analysis of the financial behavior and personal expenses of each user. Once the training, evaluation and comparison of different models was completed, using the CRISP-DM methodology as a development framework, an optimized solution was reached that solves the classification problem using the KNN algorithm, with an accuracy close to 96 %. The results showed a high level of confidence when classifying a transaction, based on a description, into a category.

Keywords: Machine Learning; Natural Language; Bank Transactions; Natural Language Processing.

RESUMEN

El sector financiero, y en concreto las entidades bancarias, experimentan cambios en los últimos años gracias a la tecnología, como la digitalización de las transacciones y la creación de aplicaciones como billeteras digitales y PFM (*Personal Finance Manager*), generando gigabytes de información. Gestionar el conocimiento se vuelve fundamental para enfrentar nuevos competidores, brindar mejores servicios, comprender el comportamiento financiero de los clientes y afrontar grandes retos a la hora de procesar y analizar el volumen de información disponible, que en la mayoría de las ocasiones requiere de un complejo proceso de preprocesamiento y calidad de datos. Es el caso de las transacciones bancarias, que incluyen información en texto libre en sus campos de observación, dificultando el análisis y su clasificación, impidiendo al banco y a sus clientes analizar el comportamiento financiero durante un periodo de tiempo. Para resolver esta problemática, se propuso el uso de técnicas de *Machine Learning* para automatizar el proceso de clasificación de transacciones a partir del texto escrito en lenguaje natural, y disponer la información que permita

realizar un análisis del comportamiento financiero y de los gastos personales de cada usuario. Finalizado el entrenamiento, evaluación y comparación de diferentes modelos, utilizando la metodología CRISP-DM como marco de desarrollo, se llegó a una solución optimizada que resuelve el problema de clasificación mediante el algoritmo KNN, con una precisión cercana al 96 %. Los resultados mostraron un nivel de confianza alto al momento de clasificar una transacción, a partir de una descripción, en una categoría.

Palabras clave: Machine Learning; Transacciones Bancarias; Lenguaje Natural; Procesamiento de Lenguaje Natural.

INTRODUCCIÓN

Las entidades financieras producen y almacenan grandes volúmenes de datos, gran parte de ellos por la interacción que los usuarios tienen con las pasarelas de pagos electrónicos, como lo es la Pasarela de Pagos Seguros (PSE). Con base a esta interacción, se realizan una serie de análisis de los gastos o egresos de los productos financieros de los clientes, por ejemplo, soluciones que permitan conocer en detalle el comportamiento financiero de una persona, y el análisis de la distribución de los gastos mensuales en las categorías como servicios públicos, educación, comida, etc. También realizar análisis más complejos desde el punto de vista de predicción, para proyectar un hecho en función del comportamiento financiero histórico.

La banca forma parte del sistema financiero y juega un rol de mediador entre los recursos y su asignación, buscando mejorar la distribución de los ingresos y disminuir la pobreza en un país.[1] En otras palabras, el concepto de consultoría ha cambiado y todos los días los consumidores necesitan conocer cuál es su comportamiento financiero para decidir: cuánto gastar, ahorrar, qué productos comprar, cuánto y dónde invertir. En tiempos de bancarización, una asesoría financiera es necesaria en todo momento, lo cual supone un reto para el sector bancario como es el de ayudar a sus clientes a tener un mayor conocimiento de sus finanzas personales. De hecho, algunas empresas emergentes del sector de la banca comercializan sus servicios haciendo uso intensivo de las tecnologías de la información, de una manera más rápida y oportuna sobre los productos financieros de sus clientes, con el objetivo de ayudarles en el control de sus finanzas personales.

Se realizan millones de transacciones comerciales a través de plataformas virtuales de pago, que no son categorizadas, clasificadas, ni etiquetadas, por lo cual su posterior análisis o interpretación resulta más compleja. En el 2018, el 50 % de las operaciones bancarias se realizaron a través de internet y generaron millones de gigabytes de datos, que, con el tratamiento adecuado, permitirían conocer en detalle el comportamiento financiero de las personas.[2]

Este trabajo desarrolla un modelo basado en *machine learning* para clasificar la información y los datos de las transacciones bancarias con el fin de disponer de la información y realizar un análisis del comportamiento financiero y los gastos personales de los usuarios en diferentes categorías.

Trabajos relacionados

En esta sección, se presentan trabajos sobre modelos de clasificación aplicados para resolver problemas similares, como, por ejemplo, en los estudios.[3,4,5,6] Un artículo[5]propone técnicas de *Machine Learning* como clasificación, regresión, *clustering* y reglas de asociación para clasificar y seleccionar las señales emitidas por un Sistema de Navegación por Sonido (SONAR), categorizando según el mejor conjunto de características basado en la incertidumbre simétrica.

Con el objetivo de lograr una mayor precisión del modelo de clasificación, el artículo[4] propone mediante la reducción de datos, utilizar algoritmos genéticos, buscando mejorar la predicción de los movimientos del mercado de valores. En la clasificación de las transacciones bancarías generadas a través de PSE, se aplican métodos de limpieza de cadenas de texto como el manejo de los conectores o *stopwords*.

Por su parte,[6] establece una comparación de los algoritmos de clasificación en función de la cantidad de datos que se requiere en el proceso de entrenamiento, en este caso, de un modelo que pretende predecir la intención de pago del cliente. Se busca unir en una misma categoría a los clientes que olvidan intencionalmente realizar un pago y los identificados como fraudulentos. Para esta actividad, se tuvo en cuenta los envíos repetidos de notificaciones, datos tomados en el momento del pedido y la información del pago.

Una investigación[3] plantea una problemática similar a la definida en este artículo, como es la de clasificar transacciones a partir de texto en lenguaje natural, esperando etiquetarlas en categorías de débito y crédito. Proponen el uso de un kit de herramientas de lenguaje natural, conocido como *Natural Lenguaje Tool Kit* (NLTK). El proceso consiste en tomar el texto de la transacción y realizar una segmentación inicial con el objetivo de dividir la fecha del texto que describe la transacción, continúa un proceso de tokenización donde se divide el texto en palabras formando una matriz, posteriormente se aplican las librerías de NLTK a dicha matriz para etiquetar las palabras de manera que se identifiquen los verbos, valores y sustantivos, entre otros; luego

son cruzadas con una base de datos de sinónimos clasificados previamente, y de esta manera asociar palabras a una categoría. Este método requiere de datos de entrada muy bien estructurados de manera que no genere error en la asignación de la etiqueta.

Fuentes de datos y herramientas de procesamiento

En este artículo, la clasificación de transacciones bancarias se basa en el procesamiento del texto escrito en leguaje natural, y consiste en identificar la cercanía o similitud semántica de las palabras para realizar asociación entre los conjuntos del corpus y las categorías de gastos definidas.

Un set de datos es creado con transacciones bancarias realizadas a través de PSE. Consta de 329 251 registros y 6 columnas que proporcionan información sobre la transacción (Fecha y hora, identificador de la persona, descripción de la transacción, sector económico de la transacción, detalle del sector y valor). Dicho set de datos es utilizado para la extracción de las características, aplicando técnicas estadísticas (Descriptiva, Agrupamiento, Análisis de variaciones) que permitan identificar la relevancia para el modelo.

Para cada transacción se extraen 3 características categóricas descritas a continuación:

- Descripción: hace referencia a la razón de la transferencia. Es un campo de texto libre que permite escritura en lenguaje natural.
- Sector: corresponde al sector económico al que va dirigida la transacción. Este dato permite complementar la descripción.
- Detalle sector: subsector económico al que va dirigida la transacción.

Para el proceso de validación, se construyó un segundo set de datos únicamente con la descripción de la transacción. Dicho set de datos contiene información diferente a la utilizada en el proceso de entrenamiento, con el objetivo de identificar el comportamiento del modelo ante datos desconocidos.

MÉTODO

El presente estudio de tipo descriptivo-exploratorio se ubica en una investigación de corte experimental. Para llevar a cabo el desarrollo de esta investigación se trabajó en las siguientes tareas agrupadas por fases.

Fase de exploración

Se realizó una búsqueda de estudios o trabajos similares en bases de datos especializadas, con el objetivo de determinar el estado del arte de la aplicación de *Machine Learning* en el dominio de la banca y otros sectores con problemas de clasificación, donde se logró inferir las diferentes técnicas usadas para la interpretación del lenguaje natural y la clasificación de etiquetas. Por otro lado, se analizaron y compararon distintas metodologías vigentes para proyectos de minería de datos, evaluando las ventajas y desventajas de las mismas como objetivo de utilizarla en el desarrollo de un modelo de *Machine Learning*.

Fase de preparación

Se realizan las transformaciones y cálculos necesarios que permitieron limpiar y enriquecer los datos, disminuyendo el error en la clasificación. En esta fase, se seleccionaron también los atributos importantes para el entrenamiento del modelo, y se descartó la información irrelevante para el mismo.

Fase de desarrollo y validación

Se entrenan los diferentes modelos aplicando los algoritmos de clasificación KNN (K Vecinos más Cercanos), SVM (Maquinas de Soporte Vectorial), Regresión Logística y Árboles de decisión. Posteriormente, se evaluó el rendimiento de cada uno de los modelos haciendo uso de un set de datos de validación, de esta manera se logran establecer las bases y fundamentos necesarios para desarrollar la estructura del modelo de *Machine Learning* propuesto para la clasificación de las transacciones PSE y se redactan las conclusiones y recomendaciones sobre el mismo.

RESULTADOS

A continuación, se plasman los resultados obtenidos a partir de cada una de las fases del modelo.

Fase de exploración

Como se evidencia en el apartado “Trabajos Relacionados”, es amplia la aplicabilidad del *Machine Learning* en los diferentes sectores económicos, especialmente cuando se busca dar solución a problemas de clasificación de datos. Un factor común entre los procedimientos utilizados por los autores es establecer una actividad de limpieza de datos que busque disminuir el error en el proceso de clasificación. Se realizó un comparativo de metodologías representativas para proyectos de minería de datos, teniendo en cuenta aspectos como uso y aceptación, punto de partida, nivel de detalle, grado de garantía, documentación y aplicabilidad en proyectos de *Machine Learning* (tabla 1).

Tras una revisión detallada de las etapas, actividades, documentación y entregables de cada metodología, se puede notar que estas son muy similares entre sí, a excepción de SEMMA que omite la etapa encargada del entendimiento del negocio, siendo esta una fase importante en todo proyecto de *Machine Learning*, dado que una solución de este tipo siempre busca ajustarse a un entorno específico, y entender cómo se generan los datos y su valor que representan en el negocio.

Por su parte, ASUM-DM se puede considerar la evolución de CRISP-DM, pero en esta transformación incorpora actividades como infraestructura, operación y retroalimentación para llevar a cabo la etapa de implementación del proyecto.[7] Basados en el aspecto "Nivel de detalle", y teniendo en cuenta que el alcance de este proyecto no contempla la etapa de implementación, seleccionar esta metodología como base en el desarrollo del proyecto lo convertiría en algo más robusto innecesariamente, evaluando, justificando o documentando actividades que finalmente no serían realizadas.

Respecto al "Grado de garantía", encontramos que todas las metodologías cumplen con una etapa de "Evaluación", lo cual nos brinda más confianza respecto al cumplimiento de los objetivos establecidos en el proyecto.

KDD[8] es una metodología muy conocida y ampliamente documentada, a diferencia de CATALYST[9] que se dificulta encontrar documentación, lo cual le resta importancia en la elección para el desarrollo del proyecto. Sin embargo, está orientada a desarrollar proyectos de minería de datos tradicionales o inteligencia de negocios, es decir, encontrar conocimiento en bases de datos, teniendo como objetivo el llevar datos de una fuente a estructuras de datos centralizadas. Ya que este proyecto, más que un proceso de extracción, transformación y carga busca el desarrollo de un modelo de aprendizaje automático, KDD no es la metodología más recomendada para ser utilizada en este caso.

Las ventajas notables de CRISP-DM[10] frente a las demás metodologías corresponde a su gran aceptación en los proyectos de *Machine Learning*, sus etapas inician con el entendimiento del negocio y los objetivos del proyecto, detalla las actividades que se deben realizar en cada etapa, facilitando su uso. Aunque en su fase de despliegue no especifica una tarea de traducción o representación de patrones encontrados a lenguaje de negocio, sí define la necesidad de comunicar y hacer visible los resultados.

Esta articulo incorpora dentro de su alcance las fases de CRISP-DM hasta la etapa de evaluación.

Tabla 1. Comparativo de metodologías

Aspectos	KDD	CRISP-DM	SEMMA	CATALYST	ASUM-DM
Uso y aceptación	Medianamente usado	Altamente usado	Medianamente usado	No registra	No registra
Punto de partida	Comprende el entendimiento del negocio	Comprende el entendimiento del negocio	Comprende el entendimiento del negocio	Comprende el entendimiento del negocio	Comprende el entendimiento del negocio
Nivel de detalle	7 etapas/promedio	6 etapas/promedio	5 etapas/bajo	6 etapas/promedio	12 etapas/muy alto
Grado de garantía	Cumple con la etapa de evaluación	Cumple con la etapa de evaluación	Cumple con la etapa de evaluación	Cumple con la etapa de evaluación	Cumple con la etapa de evaluación
Documentación	Se encuentra fácilmente	Se encuentra fácilmente	Poca documentación encontrada	Poca documentación encontrada	Poca documentación encontrada
Aplicación en proyectos de *Machine Learning*	No aplica	Sí aplica	Sí aplica	Sí aplica	Sí aplica

Preparación

Una transacción bancaria es la relación financiera entre dos actores, donde el pagador es el actor que realiza la transferencia desde una cuenta pagadora, y el beneficiario es el actor que recibe el dinero en una cuenta recaudadora.

PSE[11] (Pasarela de Pagos Seguros en Línea) brinda a los usuarios la posibilidad de realizar sus pagos y/o compras a través de internet, debitando los recursos en línea de la entidad financiera donde se encuentra registrado el pagador, y depositando este dinero en la entidad financiera del beneficiario. Ahora bien, al momento de realizar una transferencia a través de PSE, el usuario cuenta siempre con la opción de ingresar una nota de texto que haga referencia al asunto de la transferencia. Este campo libre recibe texto en lenguaje natural, caracteres especiales, mala ortografía, que elevan la complejidad para identificar a que hace referencia la transacción y de esta manera clasificarla en una categoría de gastos.

En primera instancia se buscó identificar las columnas, tipos de datos, valores nulos y estadísticas básicas, aplicando técnicas de exploración de datos como agrupaciones, descripciones, conteos, medias, frecuencias,

mínimos y máximos, entre otras. Este proceso permitió descartar las características que no aportan al entrenamiento y la precisión del modelo. Ya que el objetivo es clasificar una transacción en una categoría a partir de su descripción en lenguaje natural, campos como la fecha de la transacción, el pagador y el valor de la transacción, no aportan información relevante y no permiten desarrollar un modelo generalizado.

Como se estableció anteriormente, la calidad de los datos es un punto crítico para el proyecto, y más tratándose de una clasificación a partir de texto libre. Se aplican técnicas para eliminar valores atípicos como son *Stopwords*, mediante conteos y nube de palabras, y *Outliers* haciendo uso de técnicas de estadística, dejando el mayor porcentaje de datos y descartando valores exageradamente grandes dentro del *dataset*. El método aplicado es el método estadístico clásico, que consiste en obtener la media +- 3 * desviación estándar. Esto permite mantener el 98,8 % de los datos originales.

Una vez los datos se encuentran limpios, se procede a aplicar un proceso lingüístico llamado Lematización, que permite, dada una forma flexionada, hallar la raíz de esta,[12] tal cual lo muestra la tabla 2. El objetivo es estandarizar el lenguaje escrito de manera que se pueda transformar las palabras de diferente escritura, pero igual significado lingüístico, a una palabra que las integre.

Tabla 2. Proceso de lematización

Entrada	**Proceso**	**Salida**
Buscar Búsqueda Buscando Buscaremos Buscarás	Lematización	Busca

Hasta este punto, los datos se encuentran lo más limpios posibles y listos para ser usados en el entrenamiento, sin embargo, debido al número de términos diferentes que tiene el idioma y el lenguaje natural en general, la dimensionalidad del vector resultante sería muy elevada y haría que las palabras no guardaran relación entre sí. Para solucionar esto, se hace uso de técnicas de optimización como TF-IDF (*Term Frecuency Times Inverse Document Frecuency*).[13] Esto permite tener una reducción en el número de variables usadas en el modelo de aprendizaje, permitiendo aumentar así el rendimiento del procesamiento y eficiencia en los resultados. Los modelos predictivos no permiten hacer uso de las variables categóricas de manera directa, y es necesario transformarlas a una representación numérica previamente. Se puede ver un ejemplo del proceso paso a paso, donde se muestran los documentos y las descripciones de cada uno. La tabla 3 muestra como en el primer documento aparecen cuatro términos de los cuales “pago” aparece dos veces.

Tabla 3. Descripción de transacciones

Documento	**Texto**
doc1	pago servicios públicos pago
doc2	servicios públicos servicios
doc3	servicios hogar

Por lo tanto, para la tabla anterior, podemos decir que la palabra “pago” representa 2/4 partes del documento. La tabla 4 muestra la matriz completa con la participación de cada una de las palabras de los documentos mencionados.

Tabla 4. Matriz de participación

TF	**Pago**	**Servicios**	**Públicos**	**Hogar**
d1	2\4	1\4	1\4	0/4
d2	0/4	2\3	1\3	0/4
d3	0/4	1\2	0/4	1\2

Fuente: elaboración propia.

Los documentos se representan en un vector con la multiplicación de las partes, para cada término y cada documento, obteniendo como resultado un espacio multidimensional por cada término; donde se seleccionan los N términos con valores más altos. Lo cual indica que los términos más cercanos a 0 (cero), son poco discriminados. Entonces, después de todos los cálculos, en la tabla 5 se tiene una matriz de palabras y la relevancia de esta en la totalidad de los documentos.

Esta información sirve como entrada al proceso de entrenamiento de los modelos, enriqueciendo los datos y permitiendo una mayor precisión en la clasificación.

Tabla 5. Matriz de términos y relevancia

TF-IDF	Pago	Servicios	Públicos	Hogar
IDF	Log(3/1)	Log(3/3)	Log(3/2)	Log(3/1)
d1	0,24	0	0,04	0
d2	0	0	0,06	0
d3	0	0	0	0,24

Entrenamiento y evaluación

La tarea de clasificación de transacciones se lleva a cabo mediante la construcción y comparación de diversos clasificadores: KNN (*K-Nearest Neighbors*),[14] que asigna una categoría a un nuevo ejemplo basado en la observación de la categoría del vecino más cercano; SVM (*Support Vector Machine*),[14] que propone encontrar un hiperplano que separe mejor los datos, tal que maximice el margen; Regresión Logística, que se basa en un modelo lineal que minimiza el costo de la función "acertar o fallar" en lugar de minimizar la suma de las raíces residuales de la función como en la regresión convencional [15]y Árboles de decisión, que genera iterativamente un árbol y selecciona un atributo en cada nodo, lo que puede maximizar la cantidad de información sobre el conjunto de entrenamiento.

Para el entrenamiento de los modelos se utilizaron diferentes algoritmos de clasificación, para los cuales su parametrización final se obtiene tras realizar diferentes pruebas ajustando sus variables, configuraciones e híper-parámetros. Estos algoritmos fueron alimentados con los siguientes sets de datos:

Set de datos de entrenamiento, que corresponde al 80 % de los datos; y set de datos de pruebas que corresponde al 20 % restante. Estos datos previamente preparados con todo el proceso de limpieza e ingeniería de datos. Este Split se realiza de manera estratificada, buscando garantizar que existan registros de todas las categorías en ambos conjuntos de datos.

Una vez preparados los datos y separados en Train y Test, se procede a realizar el entrenamiento de los modelos, generando los siguientes resultados figura 1:

	Jaccard	F1-score	LogLoss
KNN	0.965026	0.963863	NaN
Decision Tree	0.897755	0.857698	NaN
SVM	0.949350	0.942665	NaN
Logistic Regression	0.866483	0.812009	0.740939

Figura 1. Resultados de precisión obtenidos para cada uno de los algoritmos de clasificación

Se puede decir que la exactitud promedio para KNN es el promedio del score F1 para todas las etiquetas, cuyo valor es de 96,50 %, definiéndolo como el algoritmo que presenta los mejores resultados en la clasificación de transacciones bancarias a partir de lenguaje natural en función de los datos adquiridos. Por su parte, los algoritmos *Decision Tree y Logistic Regression* arrojan los resultados más bajos en el cálculo de los indicadores de medición, haciéndolos los algoritmos menos convenientes para ser usados en el proceso de clasificación. El algoritmo SVM presenta resultados similares, y aunque se puede considerar aceptable, no superan los resultados de KNN.

CONCLUSIONES

En la actualidad, la banca está en camino a la transformación digital, haciendo uso de las diferentes herramientas y tecnologías para mejorar la experiencia de los clientes, desarrollando nuevos productos y servicios, haciendo uso eficiente de estas herramientas para generar, almacenar y analizar datos.

Esta investigación permitió sugerir un modelo para clasificar las transacciones realizadas a través de PSE, aplicando técnicas de *Machine Learning* a partir del uso de la interpretación de texto con la capacidad lógica de un humano que permite el NLP (Procesamiento de Lenguaje Natural) en el campo "Referencia" de la transacción, el cual recibe texto libre, surgiendo así un modelo que permita a la máquina clasificar las transacciones bancarias en categorías previamente definidas y conocer el comportamiento financiero de sus clientes, disponiendo de insumos que faciliten tener una mayor inteligencia financiera.

El desarrollo del Modelo propuesto se llevó a cabo mediante la metodología de CRISP-DM, que resultó de la evaluación realizada como la metodología más adecuada para el proyecto. En el proceso de evaluación de los modelos analizados, se logra concluir que las mejores estadísticas las proporciona: KNN (K Nearest Neighbor):

Jaccard: 96,5 %; F1-score: 96,3 %.

REFERENCIAS BIBLIOGRÁFICAS

1. Ratnawati K. The impact of financial inclusion on economic growth, poverty, income inequality, and financial stability in Asia. The Journal of Asian Finance, Economics and Business. 2020; 10(7): 73-85.

2. Davenport TH, Mittal N. All-in on AI: How smart companies win big with artificial intelligence. 1 a ed. Estados Unidos: Harvard Business Review Press. 2023.

3. Agarwal S, Mukherjee P, Chakraborty B, Nandi D. A Novel Automated Financial Transaction System Using Natural Language Processing. En Hassanien A, Azar A, Gaber T, Bhatnagar RF, Tolba M. The International Conference on Advanced Machine Learning Technologies and Applications. Advances in Intelligent Systems and Computing. Springer International Publishing. 2020. 535-545.

4. Kim Y, Enke D. Instance Selection Using Genetic Algorithms for an Intelligent Ensemble Trading System. Procedia Computer Science. 2017; 114: 465–472.

5. Potharaju SP, Sreedevi M. A Novel Subset Feature Selection Framework for Increasing the Classification Performance of sonar Targets. Procedia Computer Science. 2018; 125: 902–909.

6. Takahashi M, Azuma H, Tsuda K. A Study on Validity Detection for Shipping Decision in the Mail-order Industry. Procedia Computer Science. 2017; 112: 1318–1325.

7. Díaz Y, Hidalgo MÁ, Lagunes V, Pichardo O, Martínez B. A Hybrid Methodology Based on CRISP-DM and TDSP for the Execution of Preprocessing Tasks in Mexican Environmental Laws. En Pichardo O, Martínez J, Martínez B. Advances in Computational Intelligence. Lecture Notes in Computer Science. Springer International Publishing. 2022. 68-82.

8. Dogan A, Birant D. Machine learning and data mining in manufacturing. Expert Systems with Applications. 2021; 166: 114060.

9. Ma S, Liu ZP. Machine learning for atomic simulation and activity prediction in heterogeneous catalysis: current status and future. ACS Catalysis. 2020; 10(22): 13213-13226.

10. Cazacu M, Titan E. Adapting CRISP-DM for social sciences. Broad Research in Artificial Intelligence and Neuroscience. 2021; 11(2): 99-106.

11. Salinas E, Barrientos, AF, Quiroz JF. Pasarelas de pago en Colombia, un mercado cambiante y altamente competitivo. Colombia. Institución Universitaria de Envigado. 2021.

12. Obando J, Pulido J, Gómez J. (2020). Procesamiento del lenguaje natural para reconocer mensajes de textos extorsivos a través del análisis sintáctico y lematización. 2020; 16(1): 33-42.

13. Akuma S, Lubem T, Adom IT. Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. International Journal of Information Technology. 2022; 14(7): 3629-3635.

14. Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal. 2022; 3: 100071.

15. Shah K, Patel H, Sanghvi D, Shah M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research. 2020; 5(1): 12.

FINANCIACIÓN

Ninguna

CONFLICTO DE INTERESES

Ninguno

CONTRIBUCIÓN DE AUTORÍA

Conceptualización: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez.

Curación de datos: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez, Ivan Andres Delgado.

Análisis formal: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez, Ivan Andres Delgado.

Investigación: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez.

Metodología: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez.

Administración del proyecto: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez.

Recursos: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez.

Software: Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez.

Supervisión: Fabio Alberto Vargas, Dario Enrique Soto.

Validación: Ivan Andres Delgado.

Visualización: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez.

Redacción – borrador original: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez.

Redacción – revisión y edición: Fabio Alberto Vargas, Dario Enrique Soto, Mauricio Urrego Álvarez, Edison Javier Yepes Sanchez, Ivan Andres Delgado.

https://doi.org/10.56294/saludcyt2024.1358