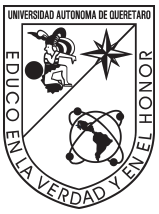


Universidad Autónoma de Querétaro
Facultad de Informática
Doctorado en Ciencias de la Computación

Arquitectura para Análisis Drill-Down mediante el uso de Multilayer Perceptron Neural Networks

Tesis

Que como parte de los requisitos para obtener el Grado de

Doctor en Ciencias de la Computación

Presenta

Victor Hugo Silva Blancas

Dirigido por:

Dr. José Manuel Álvarez Alvarado

Co-dirigido por:

Dr. Hugo Jiménez Hernández

Sinodal Presidente	**Dr. José Manuel Álvarez Alvarado**
Sinodal Secretario	**Dr. Hugo Jiménez Hernández**
Sinodal Vocal	**Dra. Ana Marcela Herrera Navarro**
Sinodal Suplente	**Dra. Diana Margarita Córdova Esparza**
Sinodal Suplente	**Dr. Juvenal Rodríguez Reséndiz**

Centro Universitario, Querétaro, Qro.

Junio 2025

México

Índice

Índice de tablas

Índice de figuras

Resumen

Actualmente se están usando tecnologías emergentes para capitalizar los datos, crear valor de negocios y competir en un mundo controlado digitalmente. La implementación de análisis de grandes bases de datos es generalmente complicada y consumidora de recursos y por tal razón es necesario desarrollar nuevas herramientas. El análisis *drill-down*, DD, que es un entorno efectivo para codificar múltiples consultas en una representación compacta y eficiente, potencia significativamente los métodos actuales de recuperación de información y puede ser aplicado al sistema financiero para la creación de indicadores, normalizaciones y predicciones, como aquellas basadas en el precio. Esta investigación a diseñado una nueva metodología de análisis de datos integrada con herramientas de *machine learning*, ML, como *K-means* y *MLP*, con el objeto de sesgar los valores difusos provocados por los problemas de análisis sobre-saturados y permitir una predicción y una proyección más certeras aprovechando la información de los mercados de valores micro y macro económico. Como resultado, durante el análisis de datos se pudo observar que el problema de la sobre-saturación se redujo al momento de realizar operaciones de clasificación a través de algoritmos aplicados de K-means previos al uso de MLP. Se dedujeron teoremas y corolarios, y también hipótesis de carácter matemático, que fundamentaron teóricamente los cálculos realizados. También, con respecto al precio y sus determinantes, se produjo un conjunto de indicadores que demostraron que las particularidades de cada tendencia informativa son capaces de ofrecer diferentes perspectivas en el aprovechamiento del conjunto de datos. La aplicación de K-means y MLP redujo la sobre-saturación durante el *exploratory data analysis* del conjunto de datos. Por el lado del mercado financiero, se pudo observar que la sofisticación del esquema de datos (dada por la cantidad de tipos, perfiles, profundidad y volumen) permitió una normalización y racionalización durante la experimentación capaz de ofrecer diferentes enfoques y perfiles. El precio como factor fundamental se vio potenciado por el tipo de transacción, su manipulación, su temporalidad y su disponibilidad.

Abstract

At this moment, emerging technologies are being used to capitalize data, achieve business value and compete in a digital controlled world. Análisis implementation for big data is generally complicated and resource demanding, and for that reason it is necessary to develop new tools. Analysis drill-down, DD, which is an effective environment for codifying multiple queries in an compact and efficient representation, empowers significantly current information retrieving methods and can be applied to the financial system on the creation of indicators, normalizations and predictions, like those based on price. This research has designed a new data analysis methodology integrated with machine learning, ML, tools like K-means classification algorithm and multilayer perceptron, MLP, predictor algorithm, with the aim of bias diffuse problem values caused by overfitting procedures, and aloud both a prediction and a projection plus accurate, exploiting the stock market micro and macro economics information. As a result, pending data analysis it could be observed that overfitting was reduced to the moment of the application of K-means algorithms, before MLP usage. Theorems and corollaries were deduced, and also mathematics hypotheses which theoretically founded all implemented calculations. At the same time, with respect to price and its determinants, a set of indicators were produced which demonstrated that each particularity for each informative tendency was able to offer different perspectives during the datasets leverage. K-means plus MLP application reduced overfitting during the dataset's exploratory data analysis, EDA. On the financial markets side, it was observed that data sophistication schema (caused by multiple types, profiles, depth and volume) allowed both normalization and rationalization, through experimentation, capable of offering a variety of focus and profiles. Price, as a fundamental factor, it was powered for transaction type, manipulation, temporality and availability.

1. Introducción

La historia de la humanidad organizada es la historia de los datos. Desde que el hombre dejó de ser recolector y cazador para convertirse en agricultor ahí en las estepas de la ahora Turquía hace 6 mil años (Fairbairn, 2005) se produjo la necesidad de control de la producción, de saber cuánto se iba a producir, a cuánta gente se iba a alimentar y cuánta semilla debía guardarse para poder sembrar al siguiente año. Lo mismo ocurrió con la ganadería, aquella de los celtas irlandenses de hace 5200 BP (notación por carbono 14 que significa *before present* y que es a partir de 1950), quienes tuvieron que aprender cómo organizar en las hembras los ciclos de la gestación, separar la leche para el alimento y la carne para el matadero aumentando al mismo tiempo el crecimiento del inventario de animales en aquellos antiguos corrales (Connell y Molloy, 2023). Después llegó la economía tribal y más tarde la nacional cuando los individuos de una región tuvieron que organizarse para sobrevivir no sólo biológica sino culturalmente bajo la égida de un pequeño grupo al que llamaron *gobierno*.

Los datos siempre han sido necesarios para organizar los recursos y controlar los planes, ya sea aquellos para la próxima cosecha, para el próximo destete o para la construcción de un puente sobre el río Tíber (Taylor y Rabun, 2002), que, como proyecto económico nacional, el producto de su peaje convirtió a una tribu en el centro del universo, palabra que en sus albores significaba *mundo*. En conclusión, al crecer la población crecieron las necesidades, aumentaron las producciones y el dinero se midió en caudales, y en el mundo moderno consecuentemente a todo esto junto se le llamó *finanzas* (Real Academia Española, 2023).

Las finanzas son la pieza más importante de la sociedad porque sin finanzas no hay crecimiento y sin crecimiento no hay supervivencia. La administración de las finanzas se soporta en la recolección de los datos y la calidad de los datos determina la eficiencia y efectividad del control económico que entorna la naturaleza de la administración particular de un país, de un sector o de una empresa.

Al amanecer de la humanidad como entidad organizada, el control financiero inició

con la aplicación de los censos, que al ser la primera fuente palpable, y por tal razón medible, de conocimiento, dió origen de forma científica al proceso histórico de los datos.

1.1. Marco histórico

Una de las primeras necesidades cubiertas por la recolección de datos fue aquella que tenía que ver con el fisco y la milicia. Los primeros censos, elaborados en tablas de arcilla en Babilonia alrededor del año 3800 a.C., eran estimaciones de los ingresos para el pago de impuestos. Egipcios, chinos y hebreos también elaboraban censos aunque de forma irregular, siendo posible que los únicos datos considerados hubieran sido un simple nombre e ingreso pues los apellidos tuvieron un surgimiento muy posterior causado por la toponimia y tradicionalmente manifestados primero en lugares como Grecia (Ecured, 2011). En Roma la realización del censo no sólo estaba supeditada a la *lex censui* sino que establecía la fórmula *Censendi* que marcaba el procedimiento para realizarlo (Cañas, 2005). Con respecto a la Edad Media, Carlomagno realizó en 762 un censo geográfico (INE, 2013). En México, durante la colonia, en la Nueva España, el Censo de Revillagigedo, llevado a cabo por el virrey Juan Vicente Güemes, manifestó una gran sofisticación (Castro, 2010). En Norteamérica, George Washington elaboró el primer censo (USCB, 2021) donde los comisionados realizaron divisiones específicas tales como condados, ciudades, etcétera (Wright, 1900).

La Tabla 1 muestra el diseño de datos que los censos de los períodos históricos anteriores habrían tenido si se hubiera aplicado la tecnología actual de nomenclatura y tipos, así como la naturaleza de las bases de datos si acaso fueron lineales o relacionales.

Para 1870, la población de los Estados Unidos superaba los 38 millones de habitantes por lo que se estimaba que el análisis del censo no podría terminarse antes del siguiente (USBC, 2021), fue cuando en 1890 Herman Hollerith utilizó la máquina tabuladora del censo, primero realizado con máquinas, completando el trabajo en tres años (Barral, 2018). Con el paso del tiempo Hollerith fundó la empresa Tabulating Machine que daría lugar a la International Business Machines (IBM) (Euston, 2019). Treinta años

Tabla 1

Historia de las bases de datos

Autor	**Período histórico**	**Base de datos**
Babilonia	3800 a.C.	lineal
Roma	4 d.C. Augusto	Relacional
Carlomagno	762	Lineal
Virrey Güemes	1790	Relacional
Washington	1790	Lineal

después el ingeniero alemán Arthur Schrebius patentó la máquina cifradora Enigma (Rijmenants, 2008) la cual contaba con un libro de claves el cual dependía de la fecha en que se estaba utilizando (Gutiérrez, 2017). Durante la Segunda Guerra Mundial, Alan Turing realizó el trabajo de descifrado de la Enigma (Miret, 2013), llegando a perfeccionar los conceptos de algoritmo, computación e inteligencia artificial (Campos, 2011). La Tabla 2 muestra las estructuras del proceso de datos mecanizadas.

Tabla 2

Estructuras del proceso de datos mecanizados

Autor	**Año**	**Datos**	**Tipo aplicado**	**Característica**
Hollerith	1890	Binarios	Boolean	Relacional
Schrebius	1918	Cifrados	Integer	Lineal
Turing	1939	Cifrados	Integer	Relacional

El proceso de datos mecanizados compartió dos características fundamentales: primera, eran mecánicas o electromecánicas y, segunda, fueron previas al surgimiento de las herramientas modernas de la digitalización como el desarrollo de la electrónica y los semiconductores.

Con el surgimiento de la electrónica en 1946 fue presentada la máquina *Electronical Numerical Integrator And Computer* (ENIAC) (Pandora, 2025). En 1947 Bardeen y asociados elaboraron el primer transistor o semiconductor (Mercado et al., 2016). El código binario fue aplicado a los semiconductores determinando que cuando un estímulo

es positivo el sistema interpreta un 1 y viceversa (Heath, 2004). Los transistores evolucionarion hasta convertirse en microprocesadores como el Hoff que tenía 2300 transistores (Rodríguez, 2008). El procesamiento a través de microprocesadores resolvió tres problemas: velocidad de lectura y escritura (Fortier y Michel, 2003). La primera base de datos (DB) moderna de uso comercial fue el sistema Sabre de IBM y American Airlines (IBM, 2001). IBM también desarrolló el Structured English Query Language (SEQUEL) al que posteriormente llamaron SQL. (Hosch, 2019). La teoría moderna de base de datos se estableció a partir de la teoría de conjuntos (Ivorra, 2015), algunos de sus axiomas son: axioma de elección, principio de numerabilidad, lema de Zorn, y lema de Zorn variante (O Connor y Robertson, 2001). Con esta base, a finales de los años 70 Larry Ellison llevaron a cabo un proyecto de base de datos relacionales llamado Project Oracle, que actualmente Oracle es el líder mundial en bases de datos relacionales (Ferreira, 2015) y que ha establecido sus productos en un alto porcentaje de las empresas lideres mundiales (Enlyft, 2019).

Usando la tecnología disponible, la administración de bases de datos a lo largo de la historia de la humanidad ha ido evolucionando hasta los recursos disponibles en esta época. La Tabla 3 muestra la línea de tiempo entre los diferentes tipos de tecnologías utilizadas para el procesamiento de bases de datos.

Tabla 3

Línea de tiempo de los diferentes tipos de tecnologías

Pre-tecnológicas	**Factores de cambio**	**Mecanizadas**	**Factores de cambio**	**Modernas**
3800-1790		1870-1839		1947-2019
	Electricidad, Mecánica, Código Binario		Semi-conductores, Matemáticas Avanzadas	
Piedra de Rosetta (Egipto)		Tabuladora de Hollerith		Servidor Oracle Exadata

La historia del proceso de datos es el fundamento primario de los controles económicos y financieros. Antiguas tecnologías siempre han sido rebasadas por las necesidades económicas y éstas a su vez provocan tanto oportunidades como problemas financieros que deben ser potenciados tanto en su calidad como en su rapidez. Contextualizar en la historia permite evitar recorrer caminos trillados no por su tecnología sino por su metodología.

1.2. Marco teórico y conceptual

De acuerdo a Keynes (1930) existen tres tipos de dinero: aquel de uso cotidiano como los billetes y las monedas, *commodity money*, aquel que representa el valor intrínseco de las cosas, *fiat money*, y el dinero administrativo o de estado, *manage money*. Actualmente hay que considerar también las criptomonedas, las cuales son sistemas de pago en línea soportados por cadenas de bloques, *blockchains*, conocidas también como sistemas de pago descentralizados, *descentralized payment systems* (DPS) que permiten realizar pagos internacionales a bajo costo y gran trazabilidad, con sofisticados protocolos (Mita et al., 2019). La Figura 1 muestra la distribución de los tipos de dinero.

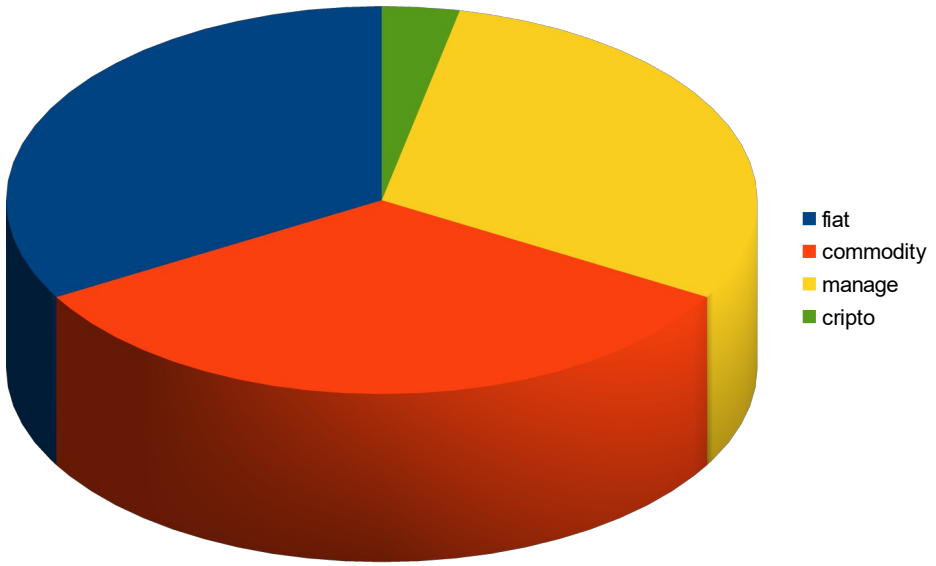
El mercado financiero es básicamente una manipulación del *fiat money* para reproducirse en ósmosis y de vez en cuando obtener *commodity money*, de acuerdo a las fuerzas de la oferta y la demanda del mercado macroeconómico regulado por el *manage money*, el Estado. El factor más importante en los ciclos financieros es el precio del dinero.

La *teoría del precio* trata con la distribución de los recursos para diferentes usos y del precio de cada elemento en su relación con los otros elementos de su entorno. Los precios tienen tres funciones: transmiten información, proveen un incentivo a los usuarios de recursos quienes los utilizan y proveen un incentivo a los dueños de los recursos (Friedman, 1963).

Los mercados *bull* y *bear* son una forma común de describir los ciclos del precio en el valor intrínseco de la macroeconomía, generalmente representado por acciones, y se dan

Figura 1

Tipos de dinero y su distribución.



durante períodos extensos de tiempo en que los precios de estas suben (bull) y bajan (bear) (Pagan y Sossounov, 2002). Las acciones son el *fiat money* que se compra y se vende en tenditas como la de la esquina de Wall y Hanover, en New York, llamadas *bolsas bursátiles*, y para lograr su entendimiento se debe conocer el proceso de generación de sus datos.

1.2.1. Los Mercados *Bull* y *Bear*

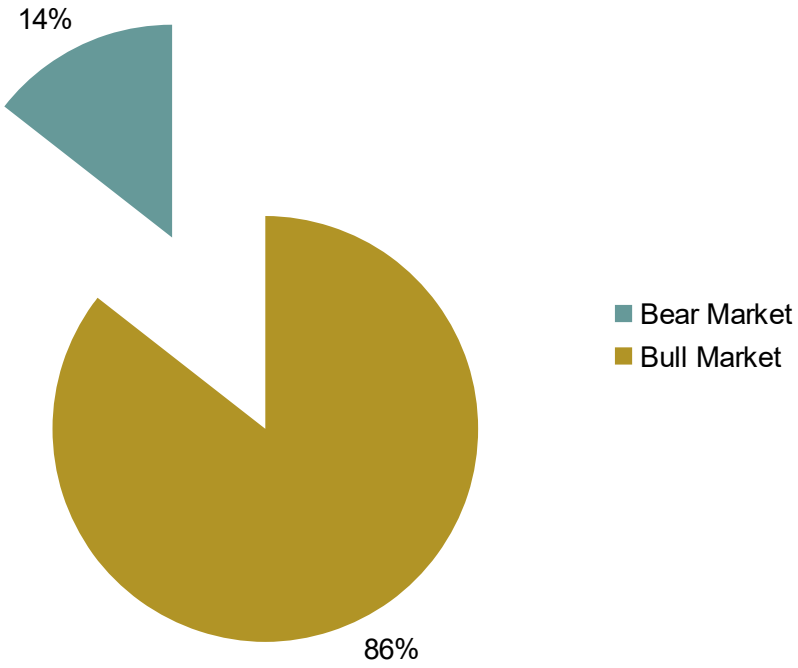
Los mercados *bull* están asociados con el aumento persistente en el precio de las acciones, fuertes intereses de los inversionistas y condiciones favorables mejoradas (González et al., 2005). En este tipo de mercado, los inversionistas atribuyen incorrectamente la generación de ganancias con sus propias habilidades y por lo tanto se vuelven sobreconfidentes de lo que serían en un mercado *bear* y ocasionan un comercio excesivo que no se puede obviar con explicaciones alternativas tales como la disposición a la causa-efecto o la tendencia al juego (Shi y Wang, 2013). En el caso de la *rumorología*, que podría incluirse en la categoría de juego, el porcentaje de ganancias después de la

clarificación de una circunstancia rumorada, como presuntas mergers, adquisiciones o reestructuraciones, ha demostrado ser significativamente positivo en el mercado *bull*, aunque en ambos mercados los inversionistas son incapaces de distinguir la certeza de un rumor o su negación, dando por resultado falta de habilidad al ajustar las estrategias correctivas (Xiaolan y Yongli, 2014). En este sentido, uno de los factores que afectan el comportamiento de ambos mercados es la incertidumbre político-económica la cual generalmente reduce los ingresos (Peng et al., 2018). Los inversionistas reaccionan de forma diferente cuando prestan atención al tono de las noticias las cuales en casos específicos influencían el volumen de compra-venta en el mercado *bull*, lo que es consistente con el *ruido* generado también durante las explosiones especulativas o la *hipótesis de Shiller* acerca de la prensa como propaganda (Hanna et al., 2020). Shiller afirma que los inversionistas racionales fijarán un precio actual en espera de futuros dividendos. Sin embargo, los precios fluctuarán más de lo que pueden explicar por sí mismos a causa de los factores sicológicos haciéndolos actuar irracionalmente, produciendo un mercado ineficiente (Shiller, 2023).

Para estudiar ambos mercados se debe hacer especial énfasis en las anomalías que presentan, estas son: momento, valor, inversión, ganancia, intangibles y fricciones. Aunque la mayoría son particulares a cada mercado otras son propias de ambos. De acuerdo a Nettayanun (Nettayanun, 2023), 27 de 187 anomalías detectadas históricamente son significativas en el mercado bear. La Figura 2 esquematiza esta circunstancia.

Figura 2

Composición de anomalías en el mercado.



1.2.2. Optimización Financiera

Para explorar las complejas relaciones entre los mercados es necesario realizar investigaciones sobre comportamiento asimétrico que puedan aplicar enfoques cruzados de cuartiles (cross-quantilogram) que permitan visualizar la dependencia asimétrica desde una perspectiva gregaria y sectorial, dando por resultado una predicción negativa/positiva en el caso del mercado *bull*/*bear*, respectivamente (Razzaq et al., 2022). Un enfoque similar se ha utilizado para examinar los mercados petroleros de varios países, incluidos los Estados Unidos, México y Canadá, en escenarios de ambos mercados, con la intención de identificar las dependencias más fuertes, la direccionalidad del efecto riesgo y detectar aquellos puntos del mercado con el potencial de causar riesgos sistemáticos globales (Shahzad et al., 2018). Por ejemplo, los efectos del oro en el mercado son asimétricos

en la mayoría de los casos: asimetría negativa es más probable que ocurra sin importar las condiciones del mercado, con poca evidencia de algún resultado significativo en el caso de criptomonedas. En contraposición a la literatura existente, ni el oro ni las criptomonedas son un buen instrumento para cobertura (Thampanya et al., 2020).

Ya en la computadora, las funciones de cópula han demostrado ser flexibles para identificar los puntos de dependencia superiores e inferiores en ambos mercados durante sus movimientos extremos de precio, aunque la correlación gaussiana-copula es incapaz de capturar la dependencia extrema (Mensi et al., 2021). En estadística, copula es una función que une, o hace parejas, a funciones de distribución multivariada en su representación de distribución marginal de una dimensión (Nelsen, 2005). Un conocimiento sólido en matemáticas está asociado con la mayoría de las inversiones productivas o de alto rendimiento, lo que incluye también estilo de ventas (talento), experiencia y diversidad en la formación educativa, sin embargo su aplicación no es automática pues los efectos no son siempre lineales. Un efecto similar positivo en el rendimiento es producto también de la especialización en ciertas áreas, aunque ninguno es capaz de explicar el tema del rendimiento durante el mercado *bear*, de hecho ninguno es sustituto de la experiencia (Talpsepp et al., 2020).

Diversos métodos han sido propuestos para identificar estados actuales y pasados del mercado para predecir cada uno. Estos abarcan tanto *semi-paramétricos basados en reglas*, como *paramétricos de Markov*, los cuales producen pronósticos superiores y llevan a un mejor rendimiento que los primeros (Kole y Van Dijk, 2016). Los modelos de Markov han sido utilizados en aplicaciones para la modelación de procesos estocásticos y secuencias (Fine et al., 1998). Por otra parte, el impacto de los activos globales financieros varía a lo largo de los diferentes horizontes financieros, en la mayoría de los casos, la correlación entre estos activos y los índices de acciones (stock indexes) no es tan significante o es débilmente positivo (Li et al., 2021).

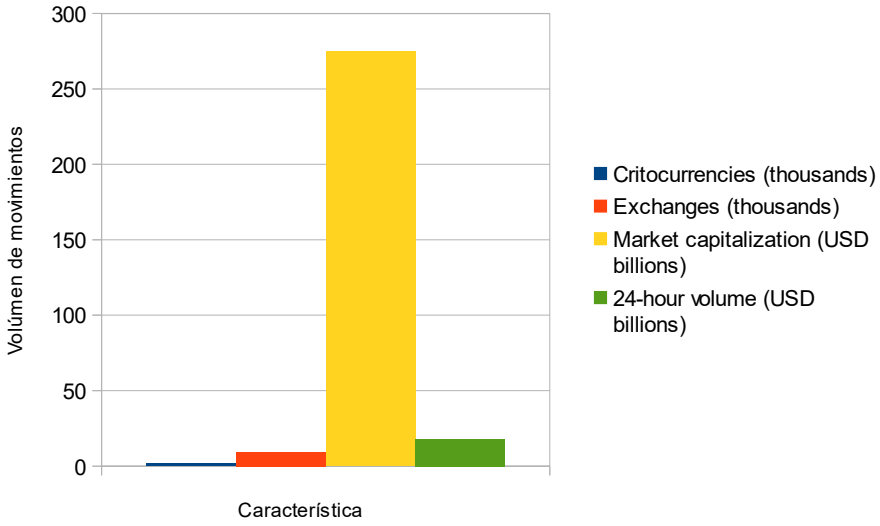
Para la predicción del precio, es necesario incorporar el volumen de operaciones de venta (datos escalares), ganancias diarias (datos funcionales) y manifestaciones emocionales de los inversionistas (datos compulsivos) a través de un entorno de predicción,

lo que ha demostrado su fuerza durante la recuperación diaria de ganancias en ambos mercados (Wang et al., 2019).

En el caso de las criptomonedas, las ganancias producidas durante un mercado *bull* indican una eficiencia mercantil cuando se utiliza un análisis no tendencioso de fluctuaciones (*detrended-fluctuation-analysis*, DFA), sin embargo, cuando las condiciones cambian y se da un mercado *bear* el mercado se vuelve ineficiente, produciendo diferencias de liquidez durante ambos mercados (Zhang et al., 2020). De acuerdo a Jani (Jani y Shailak, 2018) para marzo de 2018 existían poco más de 1564 criptomonedas. La Figura 3 muestra el comportamiento de las criptomonedas y su capitalización.

Figura 3

Criptomonedas y su capitalización.



Desde una perspectiva general, las finanzas actuales tienen como objetivo optimizar los recursos económicos produciendo una ganancia, la ganancia se enfoca en la manipulación del precio y la optimización de este proceso obedece a diferentes técnicas. En la presente investigación, como se verá más adelante, se utilizarán ANN para alcanzar dicho objetivo.

Los avances más recientes en análisis y predicción del mercado de valores caen en cuatro categorías: estadísticas, reconocimiento de patrones, ML y análisis de sentimiento,

las cuales casi siempre llevan a una técnica más amplia de análisis. Sin embargo, dentro de ML se pueden combinar categorías más amplias con enfoques más fundamentales (Shah et al., 2019).

La mayoría de los trabajos que implementan herramientas de AI en la investigación de mercados utilizan modelos de predicción que envuelven *support vector machines* (SVM) and artificial neural networks (ANN) (Mirya et al., 2019). En este sentido, Bukhari et al. (2020) afirma que los métodos tradicionales, tales como *data mining*, estadística y *non-deep neural networks*, no están diseñados para la predicción y proyección del precio de acciones. Por otro lado, el modelo *decentralized finance* (DeFi) que se refiere a la infraestrucura alternativa financiera construida en la cima de la cadena de bloques (*blockchain*) muestra ciertos riesgos pero tiene algunas propiedades interesantes en términos de eficiencia, transparencia, accesibilidad y adaptabilidad (Schär, 2021).

En el caso de técnicas de AI similares, la implementacion de *bayesian optimised recurrent neural network* (RNN) y red *long short term memory* (LSTM) se ha comparado con el modelo *autoregressive integrated moving average* (ARIMA) para series de tiempo, dando por resultado que esos métodos no lineales de deep learning (DL) son superiores a este con una precisión del 52 % (McNally et al., 2018). ARIMA es un modelo de análisis matemático que utiliza información de los datos tanto para percibir el tipo de datos como para predecir tendencias futuras (Dhyani et al., 2020).

1.2.3. Aplicaciones de la Inteligencia Artificial

En la dinámica actual del mercado de capitales, la inteligencia financiera ha demostrado una rápida y precisa potencialidad de ML para manejar datos complejos por lo que ha adquirido gradualmente el potencial de convertirse en *cerebro financiero* (Zheng et al., 2019). Tener un sistema financiero funcional de forma que dirija sus fondos hacia los usos más productivos es un prerrequisito crucial para el desarrollo económico. El sistema financiero consta de la infraestructura, los intermediarios y los mercados, y sus relaciones con respecto al flujo de fondos desde y hacia los hogares, gobiernos, negocios y extranjeros (De Haan et al., 2020). Para robustecer las condiciones de operación de un

financial services market, (FSM), y de las tareas microeconómicas que lo integran, es la formación de un mecanismo efectivo de regulación estatal que incluya elementos tanto del mismo gobierno como autorreguladores dentro del propio mercado. Del mismo modo, la formación de indicadores con la intención de analizar el impacto de dicha regulación resulta relevante (Yazlyuk et al., 2018). En este sentido, los anuncios gubernamentales con respecto a programas de prevención, políticas de pruebas y de cuarentena (en el caso de asuntos de salud) y paquetes de soporte al ingreso, resultan en amplios beneficios para el mercado (Ashraf, 2020).

Los modelos tradicionales de finanzas empíricas - que utilizan: índice de estrés, intercambios y acciones, mercados emergentes y desarrollados, gobierno y corporaciones, y madurez a corto y largo plazo - pueden ser descritos como modelos determinísticos de baja dimensión y funcionan tan bien como modelos estocásticos, ofreciendo una perspectiva adicional en los mecanismos esenciales que manejan el mercado (Orlando et al., 2022). Incluso los flujos de efectivo relativos al gasto financiero, han usado modelos determinísticos basados en la fórmula del flujo declarado en base al monto total, cambios netos en efectivo y la suma de operaciones de inversiones (Mioduchowska, 2022).

Sin embargo, actualmente en el caso del manejo de la información, aunque las normalizaciones estándares de datos basadas en la volatilidad como el nivel del precio o el porcentaje de dispersión no mejoran los resultados durante el entrenamiento en predicciones del mercado financiero, sus estructuras han sido analizadas con DL, que al incluir precio y orden histórico del flujo sobre múltiples observaciones, mejora la precisión del pronóstico, indicando con esto una dependencia sobre el seguimiento en la dinámica del precio (Sirignano y Cont, 2019).

Por parte de las criptomonedas, que entran en la categoría de *innovative financial technology*, (IFinTech), las cuales han invadido el mercado financiero y cambiado el poder de la economía global, se ha demostrado que para motivar la confianza del consumidor es necesario establecer un cuerpo regulatorio y una línea de experiencia para proveer seguridad y aceptación. En algunos casos usuarios con experiencia han demostrado grandes niveles de confianza en aplicaciones basadas en *blockchain* (Albayati

et al., 2020).

Para estudiar ambos mercados se debe hacer espec énfasis en las anomalías que presentan como son: momento, valor, inversión, ganancia, intangibles y fricciones. Aunque la mayoría son particulares a cada mercado, otras son propias de ambos. De acuerdo a Nettayanun (Nettayanun, 2023), 27 de 187 anomalías detectadas históricamente son significativas tanto en uno como en otro.

1.2.4. Retos del Mercado Financiero en cuanto a Data Warehouse

Una de las condiciones más importantes para el desarrollo de un FSM, y de las tareas microeconómicas que lo integran, es la formación de un mecanismo efectivo de regulación estatal que incluya elementos tanto del mismo gobierno como autorreguladores, dentro del propio mercado. Del mismo modo, la formación de indicadores con la intención de analizar el impacto de dicha regulación resulta relevante (Yazlyuk et al., 2018). En este sentido, los anuncios gubernamentales con respecto a programas de prevención, políticas de pruebas y de cuarentena (en el caso de asuntos de salud) y paquetes de soporte al ingreso, resultan en amplios beneficios para el mercado (Ashraf, 2020).

En el caso del manejo de la información, las normalizaciones estándares de datos basadas en la volatilidad como el nivel del precio o el porcentaje de dispersión no mejoran los resultados durante el entrenamiento en predicciones del mercado financiero analizadas durante el *deep learning*, por otro lado la inclusión del precio y el orden histórico del flujo sobre múltiples observaciones mejora la precisión del pronóstico, indicando con esto una dependencia sobre el seguimiento en la dinámica del precio (Sirignano y Cont, 2019).

En el caso de las criptomonedas, que entran en la categoría de IFinTech, las cuales han invadido el mercado financiero y cambiado el poder de la economía global, se ha demostrado que para motivar la confianza del consumidor es necesario establecer un cuerpo regulatorio y una línea de experiencia para proveer seguridad y aceptación. En algunos casos, usuarios con experiencia han demostrado grandes niveles de confianza en aplicaciones basadas en *blockchain* (Albayati et al., 2020).

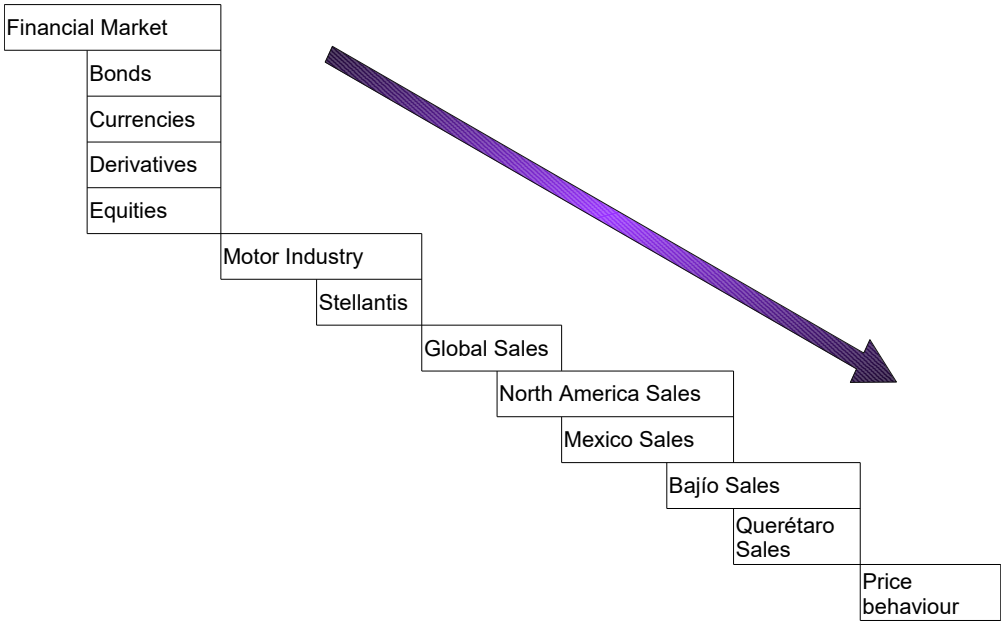
1.3. Drill-Down Analysis

Dentro del DW se encuentra el análisis DD, que es una técnica de investigación deductiva que toma un grupo de información para dividirlo y analizarlo en grupos cada vez más pequeños hasta agotar las posibilidades de división. DD provee una vista granular de los datos asumiendo una relación jerárquica entre los diferentes niveles y permitiendo acceso a perspectivas más detalladas desde un punto de vista más comprensible (Zaric, 2022).

DD significa reemplazar los datos que se están obteniendo con los datos del elemento descendente del dato (child element, en términos de clases). La operación se realiza para obtener mayores detalles. Por ejemplo: (en el caso de una empresa comercializadora de electrodomésticos) la tabla horizontal contiene las ganancias anuales por la venta de lavadoras de platos ordenadas por modelo, el análisis DD mostrará la ganancia obtenida por cada modelo de lavadora (IBM, 2022). La Figura 4 ejemplifica la naturaleza del análisis DD.

Figura 4

Naturaleza del análisis DD.



La necesidad de entender datasets ricos en información, grandes y complejos es una tarea común prácticamente a todos los campos de negocios, ciencia e ingeniería. En el mundo de los negocios, datos personales y corporativos han sido reconocidos como patrimonios estratégicos (Kantardzic, 2011). Dentro de ello, análisis de negocios, *business analytics*, (BA) es la técnica o el arte de utilizar datos cuantitativos en la toma de decisiones e incluye un gran rango de métodos de análisis de datos que van más allá de conteo, chequeo y utilizar aritmética básica. BA es el paso previo a inteligencia de negocios, *business intelligence* que plantean las respuestas a las preguntas *¿qué ha pasado?, ¿qué está pasando?* y *¿qué pasará* (Shmueli et al., 2016). El desarrollo de tecnologías de información ha generado una gran cantidad de bases de datos produciendo un aumento en los enfoques de almacenamiento y manipulación de datos para la toma de decisiones (Tamilselvi y Kalaiselvi, 2013). Para competir de manera efectiva, los directivos deben tomar a tiempo las oportunidades de grandes inversiones, por lo que requieren ser capaces de explotar las grandes cantidades de datos que se generan dentro de sus organizaciones, aunque las dificultades de discernir el valor de la información obtenida, separando lo valioso de los superfluo, evita la completa capitalización de las oportunidades (Sumathi y Sivanyam, 2006). Incluso diferentes metodologías de ciencias sociales tales como sicología, ciencia del conocimiento y comportamiento humano han implementado técnicas de data mining como una alternativa (Liao et al., 2012).

1.3.1. Data management

Para entender mejor la problemática que se busca resolver en la presente investigación, es necesario puntualizar la naturaleza del proceso de datos desde sus diferentes puntos de vista.

El ascenso de la AI, del ML y la automatización robotizada, ha cambiado la perspectiva de las instalaciones experimentales con respecto a la administración de datos, produciendo un cambio de paradigma en la forma en que diferentes sistemas de datos están apuntalados por mejoras en el análisis de datos, operación e inteligencia (Wang et al., 2021).

Al diseñar los mecanismos para guiar el comportamiento de los usuarios con respecto

a la aplicación de datos, no sólo existe el problema de construir la información sino también de invertir la predicción de características previamente analizadas para conducir un análisis de conductas futuras (Tang, 2021).

Actualmente, grandes volúmenes de datos se utilizan para almacenaje, extracción e intercambio a través de aplicaciones basadas tanto en servidores fijos como en Internet, haciendo que las aplicaciones se extiendan verticalmente en forma de enormes bases de datos, repositorios, servicios de nube e incontables instancias del tipo cliente-servidor (Gadicha et al., 2021).

La minería de datos, del inglés Data Mining (DM) es la extracción de patrones y conocimientos originarios de grandes cantidades de datos brutos o simples (aunque también se le define como buscar información oculta en la BD, e incluso analizar datos en grandes volúmenes utilizando aplicaciones) involucra colecciones de datos para almacenarlos y procesarlos de forma eficiente, que por sí misma no es necesariamente productiva (Nivethithaa y Vijayalakshmi, 2021).

Por ejemplo, al ser proveedores de DM, colegios y universidades deben llevar a cabo reformas científicas y procesales que vayan desde ideas, objetivos y métodos para que sus colecciones de BD sean conocidas por el descubrimiento del conocimiento que elaboran (Tang y Lan, 2021).

1.3.2. Data warehouse

El proceso de extraer, transformar y cargar, o *extract-transform-load*, (ETL), es un conjunto de múltiples operaciones que inician con la extracción de los datos requeridos desde los puntos de captura, y su transformación en formatos estándar para cargarlos después en un almacén de datos o *data warehouse* (DW). Sin embargo, las operaciones estándares soportadas por la definición clásica de ETL no están adaptadas para tratar con la evolución masiva de datos, haciendo del proceso de su propio mejoramiento una seria necesidad en orden de lidiar con los dominios del Big Data (BD) (Soussi, 2021).

En la industria médica el proceso general de ETL es provisto por la sociedad de resultados médicos supervisados, *observational medical outcomes partnership*, (OMOP),

que representa el paso más fundamental y central para mapear y transformar los datos en el formato de dato común, *common data model*, (CDM). Dicho proceso es dividido en: mapeo de vocabulario, mapeo de tablas de datos, transformar y cargar los datos locales en el CDM, validación de integridad y, equivalencia con la fuente de datos local (Li y Tsui, 2020).

La tecnología de DW ha sido ampliamente usada en organizaciones que proveen administración, integración de datos y apoyo al proceso de toma de decisiones ya que ofrece un medio efectivo para análisis y estadísticas en el BD (Quitaleg y Ortiz, 2020).

DW asimila y sistematiza datos a lo largo de diferentes departamentos en una organización para obtener un único análisis de toda la información. Así, estos datos son utilizados para tomar decisiones en la organización. En el medio académico, con el incremento del número de instituciones y el volumen de datos que generan, colegios y universidades consideran la integración de dicho sistema de soporte, conducido por datos para tomar mejores decisiones y más organización en los procesos académicos (Yu, 2021).

DW es la estrategia que enlaza la recolección de datos en ambientes físicos, su clasificación y su interpretación, en un repaso general de sus componentes, esta estrategia comprende la definición del entorno de datos los cuales pueden ser primitivos o derivados, divididos en operacionales, atómicos, departamentales e individuales. Por su enfoque, DW es orientado, integrado, no volátil y temporal. Orientado significa que por cada tipo de disciplina o sector económico existe un tipo específico de DW. Integrado significa que los datos son transformados con un objetivo específico. No volátil indica que los datos una vez obtenidos no son cambiados sino interpretados, y temporal que cada registro y cada operación tiene un sello en el tiempo, o time-stamp, que define su inviolabilidad. Otras características importantes son: granularity o granularidad, es la capacidad de presentar información de forma agrupada y por lo tanto más eficaz, y que obedece al axioma: entre más detalle menos granularidad y viceversa. Data modeling o modelado de datos, que es la forma en que las estructuras de datos se estructuran y relacionan. Executive information system o sistema de información ejecutiva, que

corresponde al análisis y explotación de los datos para el monitoreo, rendimiento y resolución de problemas (Inmon, 2005).

1.3.3. Árboles de decisión

Los árboles de decisión son una técnica de clasificación de DM que intenta predecir el comportamiento de la base de datos. Este objetivo es soportado por diversos algoritmos, uno de los cuales es el dicotomizador interactivo, Iterative Dichotomizer 3 (ID3), que muestra las predicciones en una estructura arbolaria. Con la aplicación de los árboles de decisión, los depósitos o las pilas de datos pueden ser procesadas para producir reglas o árboles de decisión como soportes a la toma de decisiones en la solución de problemas (Buaton et al., 2019).

Los árboles de decisión pertenecen a los métodos de clasificación más efectivos y su principal ventaja es la interpretación simple y amigable de los resultados obtenidos. Por otro lado, su principal desventaja es que su algoritmo está construido en un árbol casi óptimo (Mitrofanov y Semenkin, 2021) que implica otro tipo de análisis.

A pesar de sus cualidades, para tratar con BD y problemas de modelos complejos, los árboles de decisión muestran insuficiente precisión y sobrellenado. Para resolver esto, se han introducido redes neurales como un nodo en el mismo árbol, además de un algoritmo mejorado basado en el mismo árbol de decisiones de la red neural (Zhang et al., 2020).

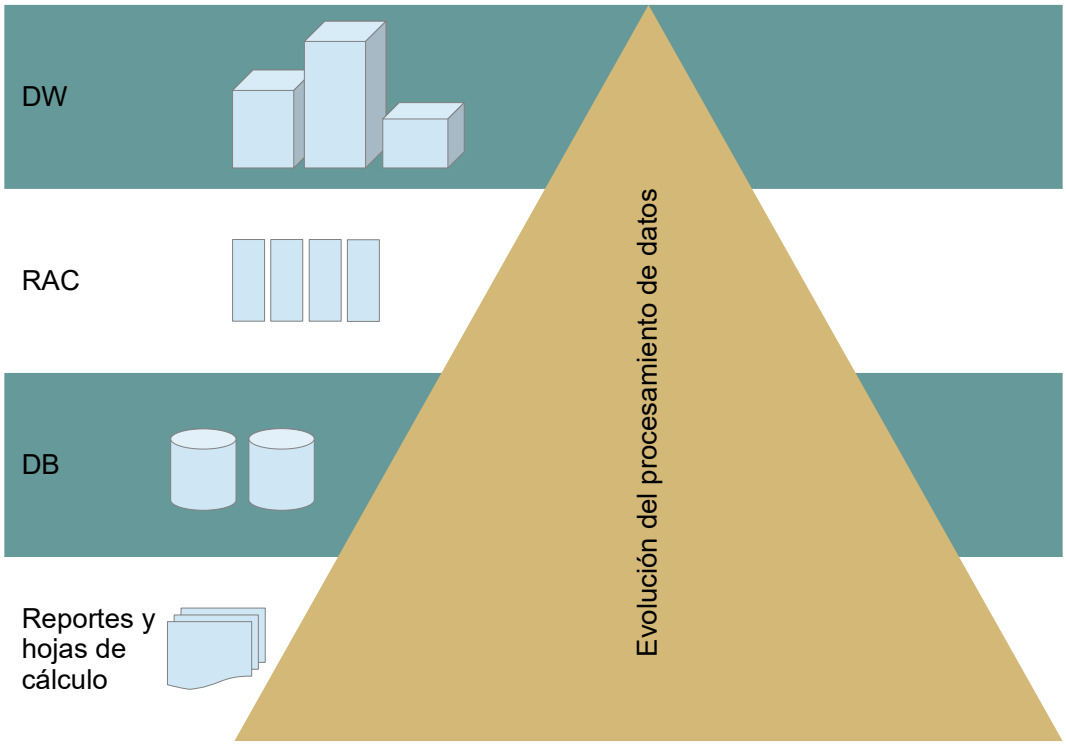
Originalmente los árboles de decisión fueron aplicados en la categorización de datos, pero en la práctica son más comunes en las variables continuas. Al seleccionar variables continuas como variables independientes, esto es, como nodos divisores en los árboles de decisión, la complejidad se incrementó debido a su gran número. Cómo mejorar su rendimiento en variables continuas no tradicionales se ha vuelto el foco de discusión al menos entre académicos (Jiao et al., 2020).

La pirámide evolutiva del procesamiento de datos lleva desde la simple creación de reportes en editores de texto y hojas de cálculo, hasta el uso de procesadores de datos especializados llamados bases de datos, data base (DB), su almacenamiento de

recuperación simultánea en grupos, real application clusters (RAC) y la implementación de diversas herramientas de hardware y software para su mejor explotación, llamadas en conjunto DW. La Figura 5 muestra dicha pirámide.

Figura 5

Pirámide evolutiva del procesamiento de datos.



Existen dos tipos diferentes de DW: aquellos servidores locales y aquellos servidores en la nube, y aunque actualmente los servicios de nube, cloud computing, cloud servers o simple y sencillamente cloud, están relacionados directamente con la actividad humana de redes sociales, herramientas enciclopédicas y recursos de investigación, los servidores locales forman la parte más importante del núcleo operacional del manejo de datos. Por ejemplo, las empresas telefónicas manejan DW locales agrupados en sitios de seguridad llamados búnkers desde donde controlan tanto el flujo de llamadas como los recursos de datos. Empresas comerciales de consumo detallista, o retail, utilizan servidores locales para las operaciones diarias, búnkers para almacenamiento de consolidación y respaldo durante la noche, y servidores de nube para la interacción con los clientes.

Las prácticas para la manipulación correcta de los datos no sólo obedecen a una cuestión estratégica sino también legal. La ley Sarbanes-Oxlay obliga al respaldo y protección de datos tanto con fines fiscales como financieros, haciendo que las empresas tengan que contratar servicios de DW externos e independientes a ellas para salvaguardar los datos y evitar pérdidas por accidentes, eventos meteorológicos y cibercriminales (United States Congress, 2002).

Se han creado diferentes clasificaciones para la gran variedad de herramientas que actualmente ponen a disposición de los usuarios los DW. A las DB especializadas en grandres volúmenes con registros que ocupan amplios espacios en los servidores se les ha dado el título de big data (BD), a la nueva generación de servidores de alta velocidad con procesadores multinúcleos y multiprocesos se les conoce como exadata, a las herramientas que conectan los utensilios de la vida diaria con los recursos de la nube se les llama internet of thins (IoT), y así sucesivamente, conforme van apareciendo innovaciones que potencian los recursos existentes o crean nuevas necesidades que deberán ser cubiertas por nuevas estrategias de consumo.

Ante la implementación de todas estas nuevas tecnologías y al aumento del tráfico intenso tanto en los servidores locales como en la nube, es necesaria una reevaluación de las técnicas utilizadas para la toma de decisiones en base al análisis profundo de datos.

1.3.4. Data Mining

Con la utilización de recursos provistos por la minería de datos, data mining, (DM) algunas estrategias nuevas se han definido pero la aplicación de DM, fuera del contexto de DW, como se verá más adelante, tiene una perspectiva sesgada.

Dentro de las técnicas de DM se encuentra el análisis drill-down (DD) que es una técnica que para ayudar a proveer diferentes imágenes de datos en reportes, esquemas y hojas de cálculo, es sencilla, valiosa y ayuda a revelar el origen de las tendencias expuestas (Morris, 2021). Esta técnica, en los modernos paneles de toma de decisiones, ha sido fácil de entender aunque difícil de implementar ya que su limitación principal ha sido la técnica de explotación de data mining (DM) que ha utilizado en su explotación

desde modelos probabilístico hasta modelos determinísticos generando algunos problemas como la recursividad, llamada overfitting, y dejando de lado la producción de conocimiento base.

El análisis DD ha sido poco desarrollado en los centros de investigación y su implementación ha sido echada a un lado dadas las limitantes en que ha caído a causa de la sencillez de sus algoritmos. Sin embargo, la aplicación de NN, y específicamente de MLP dará un nuevo enfoque y permitirá el reaprovechamiento de esta técnica.

1.3.5. Machine Learning

ML es una vertiente evolucionada de algoritmos matemáticos diseñados para simular la inteligencia humana al aprender del entorno circundante (El Naqa et al., 2015), sus modelos han tenido éxito al aprender patrones sofisticados para usarse en predicciones sobre datos insupervisados y habilidades de interpretación (Murdoch et al., 2023).

Así como los componentes de ML son colaboración en línea, seguimiento del comportamiento y análisis del aprendizaje, una de sus implementaciones es en sistemas de aprendizaje digitales, el sistema de administración del aprendizaje, del inglés Learning Management System (LMS), el cual ayuda a la automatización de ML en procesos usados por profesores para analizar resultados (Herbert et al., 2019).

A su vez, el paradigma de la manufactura está explorando enfoques de producción inteligente para mejorar su competitividad, tal es el caso de: internet de las cosas, del inglés Internet of Things (IoT), gemelos digitales, del inglés Digital Twins (DT) y el mismo ML, entre otras (Kishorre et al., 2021).

Aunado a esto, los modelos híbridos, compuestos por teoría del conocimiento y métodos de aprendizaje de ML, han contribuido a la creación de modelos para la predicción exacta del comportamiento (Saini et al., 2021).

1.3.6. Elementos más significativos de ML

ML es parte de AI, la cual es el centro de los sistemas que pueden aprender por sí mismos sin ser reprogramados muchas veces por humanos. La implementación de ML necesita datos para efectos de entrenamiento antes de producir un resultado. Una aplicación del tipo ML está frecuentemente enmarcada en un tipo específico de dominio, pues no podría interpretarse generalmente para todos los problemas. El resultado del entrenamiento es el mismo conjunto de datos o un conjunto opuesto de los mismos (Allwin et al., 2019).

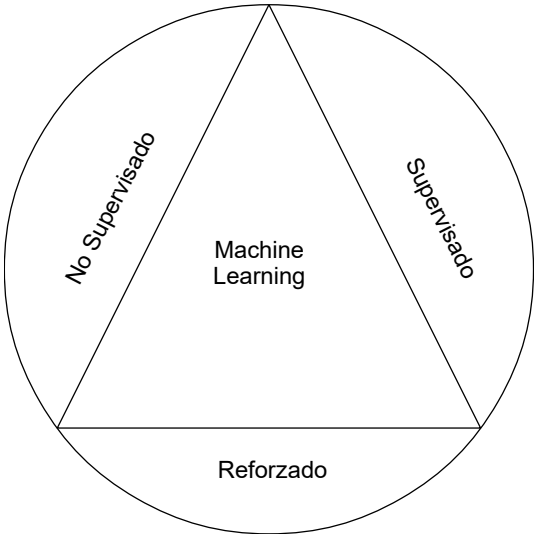
Según (Jin, 2020) la clasificación básica de ML lo divide en:

- **Supervisado** cuando establece objetivos de aprendizaje antes de aprender.
- **No supervisado** cuando la máquina no marca el contenido para cierta dirección durante el proceso completo y
- **Reforzado** cuando el aprendizaje es sistemático para cierto contenido

La Figura 6 ejemplifica la naturaleza del análisis DD.

Figura 6

Clasificación de Machine Learning



Nota. Fuente Jin (2020).

Utiliza a su vez varios tipos de algoritmos entre los que se incluyen:

- Árboles de decisión cuando inicia desde el nodo raíz y busca las posiciones en que los nodos se cruzan
- Bosque aleatorio cuando se crean múltiples conjuntos de árboles de clasificación durante el proceso
- Red neural artificial cuando imita el proceso de la transmisión de información humana
- SVM, cuando durante el proceso de aplicación depende de la máquina de vector
- Impulsar y embolsar para mejorar la precisión del resultado y
- Retro propagación, del inglés Backpropagation (BP), que incluye una capa de entrada, una de salida y una escondida.

La idea detrás de ML es simplificar el desarrollo de modelos analíticos, tales que con la ayuda de los datos disponibles los algoritmos puedan aprender continuamente (Shende, 2021).

1.3.7. Instrumentos teóricos para el análisis de datos

Los métodos clave para el análisis de ML son:

- Clustering para distribuir la multitud de varios objetos en grupos llamados clusters
- Clasificación que es necesaria para descomponer números finitos de objetos en clases predefinidas por uno o más parámetros y
- Regresión que envuelve el estudio del impacto de una o más variables independientes en la variable dependiente. Este último es el más preciso método de análisis de datos.

Ahora bien, según (Hong, 2020), con respecto a las características aplicadas a la tecnología de bases de datos se incluyen:

- Tipos de BD
- Organización
- Selección del modo de almacenamiento
- Protección de archivos encriptados
- Optimización de la tecnología de programación y
- Diseño de la actualización de hardware y software.

Actualmente, en la parte académica, las cátedras de BD de los programas de estudio de ingeniería en informática muestran el ciclo de vida del desarrollo de sistemas de BD en términos de conceptos, utilización, administración de aplicaciones, práctica y entendimiento. El modelo de investigación de diseño de BD inició con la investigación de referencias, análisis de lectura, rigor intensidad e investigación (Supriana, 2020).

Con respecto al hardware, al darse la participación de múltiples dispositivos en las últimas décadas, varias BD multi-máquina han sido construidas para responder cuestionamientos de infraestructura física tanto en lo general como en lo específico, particularmente en lo relativo a la extrapolación de resultados presentes y a la siguiente generación de dispositivos. El desarrollo de técnicas permite determinar si es razonable esperar que la parte física sea la misma en varios equipos o que las entradas de datos no representen inaceptables parcialidades (Murari et al., 2019).

1.3.8. Deep Learning

DL es un subgrupo de ML que tiene la estructura de una NN organizada en múltiples capas lo que le permite realizar tareas complejas (Chassagnon et al., 2019). Son considerados también como un subgrupo de ANN cuando el uso de multicapas (capas ocultas) es preferido ya que pueden manejar más de un problema a la vez y proporcionar una única

respuesta. En su mayoría está basado en las *deep neural networks* (DNN) y *convolutional neural networks* (CNN). Tienen una estructura básica de entrada, llamada *matriz X*, capas ocultas, llamadas *neuronas*, y una capa de salida o *respuesta* (Amigo, 2021).

Los avances en análisis matemático, hardaware y software, y disponibilidad de BData, han hecho posible a estos modelos operar como administradores de inversión, analistas financieros e intermediarios (Culkin et al., 2017).

1.3.9. Definición de MLP-DD

Utilizar MLP para la resolución del overfitting durante el análisis DD requiere de una terminología específica que evite divagar sobre el objetivo de la investigación. Ya que no se usarán algoritmos del tipo k-means, cnn y otros que refieren a la utilización de redes neurales, es necesario proporcionar al investigador una terminología sencilla que pueda identificar durante la reproducción de los experimentos o la creación de nuevos algoritmos.

Por lo tanto, durante la presente investigación se dará al uso de MLP durante el análisis DD el término MLP-DD. Con lo que la expresión se enfocará al uso de dicha técnica y dejará abierto el término para la utilización de otras técnicas en diferentes investigaciones.

1.4. Planteamiento del problema

El planteamiento del problema pasa por dos vertientes: el análisis DD y el estudio del precio. En ambos casos se debe establecer la pregunta de lo que existe y de cómo debería ser. La existencia del análisis DD parte de su uso a través de métodos determinísticos y de cómo debería ser utilizando ANN. De la misma forma, el cálculo del precio, su origen a partir del costo bruto y la sugerencia sobre un estado de predicción micro y macroeconómica también se basan por lo regular en modelos determinísticos y la sugerencia es aplicar también ANN para una proyección más exacta y robusta.

1.4.1. La circunstancia actual del análisis DD

Uno de los problemas en el análisis DD detectados a raíz de la revisión sistemática de la literatura es el overfitting. Overfitting es el uso de modelos o procedimientos que violan el principio de parsimonia, esto es, que incluyen más términos de los necesarios o usan enfoques más complicados que los necesarios. Existen dos tipos de overfitting: el primero cuando se usa un modelo que es más flexible de lo que debe ser, por ejemplo en el caso de una condición cuyos límites no son claros tal como "...la suma de los números mayores de 0...", y el segundo cuando se sobre-representa el rendimiento en un dataset, por ejemplo cuando en una consulta se toma un dato como la fecha y se utiliza como parámetro en todas las consultas. (Hawkins, 2003). Shalev y Ben-David (Shalev y Ben David, 2014), abordando matemáticas teóricas, lo definen como el predictor cuyo rendimiento en los datos de entrenamiento es excelente y sin embargo en el mundo real es muy pobre y se manifiesta cuando los datos de entrenamientos cubren demasiado bien las necesidades de una hipótesis. El teorema 1 muestra un predictor de amplia definición que puede ocasionar overfitting.

$$h_S(x) = \begin{cases} y_i & if\ \exists\, i\ \in\ [m]\ s.t.\ x_i\ =\ x \\ 0 & otherwise \end{cases} \tag{1}$$

Modelos DD con overfitting tienden a memorizar todos los datos, incluyendo el ruido del training set, en lugar de aprender la disciplina oculta detrás de los datos, lo que ocasiona duplicidad en el consumo de recursos y ampliación en los tiempos de proceso con la consiguiente pérdida de la eficacia. Es, en otras palabras, trabajar doble sobre procesos que han producido resultados durante el primer ciclo de operaciones.

La investigación se focalizará en la mayoría de los casos de uso en el aspecto financiero donde el análisis de esquemas microeconómicos se encuentra en constante evolución dadas las nuevas tendencias tecnológicas y los nuevos productos del mercado acordes a las nacientes tendencias tecnológicas como sería, por hacer una comparación, el cambio de la circunstancia del mercado ante el nacimiento de los Chatgpt, Bing Chat,

Jasper y otros, que potencian la predicción de tendencias y evaluación de instrumentos creando nuevos perfiles y que obliga a estudios complementarios dentro análisis de datos.

En el medio financiero, la AI es determinante en los procesos tecnológicos actuales, teniendo un lugar preminente en temas de innovación que influencian a los consumidores, brokers y mercados, obligando a una implementación responsable de todos sus recursos (Ostmann et al., 2023). Ya sea tratándose de una tecnología emergente o como elemento de empoderación, una de las facetas más productivas el análisis DD es en el ámbito financiero.

1.4.2. El estado ideal del análisis DD

Para evitar el problema del overfitting en el análisis DD se han propuesto algunas soluciones:

- Detención temprana (early-stopping), la cual impide que la precisión de los algoritmos deje de mejorar después de cierto punto.
- Reducción de red (network-reduction), se trata de reducir la cantidad de ruido al reducir el tamaño de los clasificadores.
- Expansión del training-data, que es mejorar la cantidad y la calidad del training-dataset, especialmente en el área de aprendizaje supervisado.

El incremento de parámetros demanda una gran cantidad de training-data para sintonizar los hiperparámetos, incluso un entrenamiento perfecto no solo debe ser grande en tamaño si no incluir limitadas porciones de ruido (Ying, 2019).

Las herramientas anteriores implican la utilización de inteligencia artificial y dentro de ésta es factible la aplicación de MLP como el recurso más viable. La integración de AI para la resolución de diversos problemas tecnológicos ha demostrado ser la opción más productiva, los sesgos dejados por técnicas diversas han encontrado solución al aplicarse redes neurales. MLP es una técnica básica y sencilla de aplicar y de comprobar que será adaptable a la solución del overfitting.

1.4.3. El Precio como la ganancia agregada al costo

Medir las primas de riesgos de activos, *asset risk premiums*, es el problema canónico que existe cuando se establece el precio de las acciones y al aplicarse herramientas de machine learning los métodos de mejor rendimiento han sido árboles de decisión y ANN (Shihao et al., 2020). El precio de reserva toma en cuenta la pérdida de oportunidad al vender productos energéticos y su punto fijo puede ser obtenido en base a la capacidad de almacenaje determinado dentro de las mismas locaciones pudiendo ser modificado de acuerdo a los requerimientos del operador del mismo precio (Akbary et al., 2017).

1.4.4. La predicción, la proyección y la suerte del precio con ANN

Dada su gran volatilidad, el precio de las acciones es difícil de predecir pues depende de diversos factores políticos y económicos, cambio de liderazgo, sensibilidad del inversionista y muchos otros. Intentar la predicción basándose en datos históricos o información textual ha mostrado ser insuficiente. Los estudios existentes en análisis de sensibilidad, *sentiment analysis*, han encontrado que hay una gran correlación entre el movimiento del precio y la publicación de artículos científicos, otros han intentado usar algoritmos como *support vector machines*, *naive bayes regression* y *deep learning* (Mohan et al., 2019).

La predicción del mercado ha sido identificada de forma práctica en el campo económico, sin embargo y su sincronización es uno de los mayores desafíos a consecuencia de las características del ruido y volatilidad implicadas las cuales han integrado últimamente modelos de predicción basados en *deep learning* que consideran la tendencia emocional del inversionista (Jin et al., 2019). Algunos estudios han adoptado *long short-term memory* (LSTM) para reconocimiento de habla y texto combinado con *differential evolution* (DE) para la determinación del precio en los casos de commodities como energía al identificar hiperparámetros dentro de este algoritmo (Peng et al., 2018). Diferentes tipos de enfoque han sido tomados para determinar las tendencias como por ejemplo el uso de modelos de RNN, el mismo LSTM, y *bi-directional long short*

term memory (BI-LSTM), notándose una gran precisión al asignar valores propios a los hiperparámetros (Sunny et al., 2020).

Es por lo anterior que esta investigación busca, a partir del diseño de una nueva metodología del análisis DD y su integración con herramientas de ML como K-means y MLP, sesgar los valores difusos provocados por el overfitting y permitir una predicción y una proyección del precio más certera, en el sentido de la precisión, para aprovechar la información del mercado de valores disponible y obtener mayor solidez al predecir factores como el precio, cuyos participantes microeconómicos y macroeconómicos tienen como fortaleza.

1.5. Justificación

Los beneficios de un análisis de DD potenciado por MLP de NN permitirá una reducción del overfitting durante el análisis de datos que se obtiene de aplicar la técnica DD, la cual solamente utiliza modelos probabilísticos y deterministas. También se podrá abatir la falta de creación de herramientas plug-in, para la implementación en librerías de investigación, y la falta de creación de conocimiento base, que fueron los tres problemas principales detectados durante la revisión de la literatura.

La falta de creación de conocimiento base reduce la capacidad de innovación al dejar de incentivar la producción de nuevas ideas y fijar rutas más claras por donde canalizar los últimos avances tecnológicos como sería el desarrollo de nuevos procesadores para la mejorar la velocidad, nuevos sistemas de almacenamiento en discos duros o el diseño de nuevas arquitecturas para servidores exadata.

En el caso de las herramientas plug-ins, el diseño de una nueva arquitectura será la puerta abierta para la participación de tecnología a través de open-source.

Además, al mejorar la estructura de DW con la integración de MLP, los sistemas de bases de datos se estarán potenciando para un mejor aprovechamiento de BDA y su uso en sistemas médicos, comerciales y financieros que requieren de un tipo de inteligencia artificial (AI) que cubra tanto el diagnóstico como la predicción en sus

procesos productivos.

1.6. Motivación

Existen dos vertientes fundamentales que han dado motivado la presente investigación: una que se refiere a la bioética y otra al medio financiero.

1.6.1. Bioética

Para Pérez (2022), la sustentabilidad reconoce tres estados éticos: el pragmático, el obligatorio y el geocéntrico. De acuerdo a este último, el desarrollo de la tecnología de forma directa o indirecta afecta la conducta humana, no obstante esta tecnología sea el desarrollo de bases de datos, pues la operación de servidores, y de toda computadora, implica la producción de minúsculas aunque palpables moléculas de dióxido de carbono. De acuerdo a Statista (2023), existen cerca de 1.5 billones de computadoras operando en el mundo. Si a esta cifra se agrega la producción diaria de moléculas de CO_2, se deberá reconsiderar la verdadera aportación del tiempo que los procesos de análisis de datos cuestan de manera efectiva a la ecología. Una de las principales motivaciones de esta investigación es la eficientización de los motores de análisis y búsqueda que llevará a un menor consumo de recursos de hardware y que por el contrario el overfitting ayuda a incrementar. En el caso de las afectaciones personales por el uso de los datos procesados durante esta investigación, éstos pertenecen a la plataforma del dataverso de Harvard, disponible en (Balogh y Attila, 2023), y no representan información de tipo personal que pudiera poner en riesgo la privacidad, integridad o seguridad de sus generadores, que son empresas cotizantes en las bolsas de valores de los Estados Unidos.

1.6.2. Financiera

Los pronósticos financieros tradicionales se enfocan en valores proyectados y bandas de confianza asimétricas, indicando poco más allá del hecho de que los ingresos futuros sean tan altos como bajos para la media predicha (Trindade et al., 2007). La aplicación

de algoritmos y conceptos desarrollados dentro del MLP-DD representarán una innovación tal que mejorará la certeza de los pronósticos financieros, asegurando la confiabilidad durante las proyecciones y funciones de rentabilidad empresariales.

2. Antecedentes

Con respecto a la importancia del análisis de datos, Mohsen (Mohsen et al., 2018) menciona que los líderes empresariales alrededor del mundo están usando tecnologías emergentes para capitalizar los datos, crear valor de negocios y competir efectivamente en un mundo controlado digitalmente, soportándose en análisis de datos para acelerar tiempos y obtener un mejor entendimiento de las necesidades y deseos de sus clientes.

El análisis de BD permite resultados potenciales ilimitados para el descubrimiento de conocimiento. Sin embargo, la implementación de análisis de BD, o big data analytics (BDA), en cualquier área es generalmente complicada y consumidora de recursos con un gran índice de falla y sin un mapa o estrategia exitosa para guiar a los interesados (Imran et al., 2021). Esta controversia motiva a los investigadores a desarrollar nuevas tecnologías o técnicas en el manejo de grandes volúmenes de datos.

En la era actual de avances tecnológicos sin precedentes, el uso efectivo de BDA se ha vuelto un requerimiento fundamental para las organizaciones y provee oportunidades para cadenas de suplementos sostenibles para incrementar la competitividad y mejorar el rendimiento y la productividad, sin embargo, su implementación plantea riesgos, por lo que es importante desarrollar un aprendizaje profundo de los riesgos en orden de generar estrategias innovativas para superarlos (Kusi et al., 2021).

2.1. Naturaleza del análisis DD

Por su parte, DD es un entorno efectivo para codificar múltiples consultas con una representación compacta y eficiente que extiende significativamente los métodos actuales de recuperación (Tan et al., 2019). Durante la revisión sistemática de la literatura con

respecto al análisis DD, se encontraron diversas investigaciones, su metodología, las problemáticas solucionadas y aquellas que estuvieron fuera del alcance de los investigadores o que manifestaron su falta de cobertura. La Tabla 4 muestra las metodologías utilizadas.

Tabla 4

Metodologías utilizadas

Metodología	**Trabajos**	**%**
Comparative	13	16.88
Descriptive	28	36.36
Experimental	36	46.75
Post-facto	3	3.89

2.1.1. Problemas Resueltos

Los estudios que han resuelto una gran diversidad de problemas clasificados por categoría son descritos enseguida.

Metodología Comparativa

En (Ying et al., 2004), los autores presentan un esquema de rangos que es utilizado para compactar y marcar la correlación en metadatos, lo cual produce adaptabilidad y escalabilidad no probada. En el caso de visualización, esta permite a los investigadores organizar aleatoriamente, mientras que los grupos multi-comparables hacen posible que las comparaciones de algoritmos de clústers y el análisis multidimensional de datos (Lex et al., 2010). La visualización interactiva de una secuencia de filtros y combinaciones lógicas produce más rápidos y eficientes diagramas de flujo (Geymayer et al., 2011). En el camino de analizar datos de procesos de producción contra aquellos obtenidos a través de la simulación, se predice el surgimiento en la precisión de fallas sorprendentes (Nemeth et al., 2021). La implementación de prototipos de expansión proactiva de rápidos DD (McGuffin et al., 2004). Los métodos de accesos métricos hacen posible entender la organización de datos (Vieria et al., 2010). Mientras que al

integrar capas en sistemas SOA se manifestó que el servicio de transporte interno (bus) permitió una definición declarativa de cómo reaccionar frente anomalías y diagnosticar problemas de origen (Psiuk et al., 2012). Los métodos estadísticos gráficos, así como los métodos de data mining producen técnicas de descubrimiento de conocimiento (Nemeth y Michalconok, 2017). Los métodos interactivos máquina-hombre ejecutan data mining para clasificación de datos y análisis de relatividad (Wang et al., 2004).

Al entender el comportamiento a nivel sub-clustering al agregar filtros progresivamente (Jung et al., 2019) se muestra cómo la tendencia a la desviación es atribuida a cambias locales, también llamada *drill-down fallacy*. El cuestionario de usabilidad post-estudios del sistema, o Post-Study System Usability Questionnaire (PSSUQ), una herramienta hecha para pruebas basadas en la satisfacción del usuario, permite a las métricas de SOLAP completar la visualización OLAP en operaciones y datos (Sitanggang et al., 2019). La técnica Kitchenham para selección y agrupación hace posible buscar learning analytics en big data que generalmente intentan aprovechar procesos de aprendizaje (Yunita et al., 2021). La jerarquización en agrupaciones de datos y un modelo híbrido en data warehouse para extracción y análisis resuelve obstáculos en algoritmos de procesos de data mining, incluso en data cube (AlJanabi et al., 2017).

Metodología Descriptiva

Para Angryk and Petry (2005), el conocimiento de data mining multi-nivel hace posible mejorar la metodología para aplicarla científicamente. Identificar métricas relevantes mientras se exploran data cubes ayuda a soportar las funciones de toma de decisiones las cuales están integradas en OLAP comerciales (Cariou et al., 2009). Substrayendo todo el análisis para visualizar la cobertura de enormes datasets permite identificar sus huecos (Adler et al., 2009). Machine learning creado para construir estructura de tres formas que interpretan dependencias en un KPI permite al análisis de negocios procesarlas incluso si dependen de métricas de bajo nivel (Wetzstein et al., 2009). Las técnicas que están basadas en histogramas para reducir las ventanas desplazables han propuesto un rediseño en el árbol multi-estructural (Buccafurri et al., 2010). La identificación cruzada soporta la clasificación de tráfico multilateral y jerárquica (Kim et al., 2012). El entorno

de doble nivel que unifica medidas microscópicas y macroscópicas mide el spam para señalar resultados sospechosos en datasets de clasificación de ratings usados en sitios web de restaurantes (Xie et al., 2015). El uso de técnicas no-supervisadas para descubrir actividades diarias en residentes de casas inteligentes produce una identificación automática de tales actividades (Yin et al., 2015). Cuando el flujo de información es agrupado en células con diseño de matriz para identificar patrones e instancias en grandes redes, este fluye junto con otras cantidades distinguibles (Chen et al., 2017). Para crear un diseño conciso de esquema el proceso de agrupación de datos se debe ajustar para entender la relación entre orígenes de datos para que la estructura de diseño pueda implementar una arquitectura robusta, documentada y actualizable (Jiménez, 2018). El enriquecimiento presentado por diagramas UML y lenguaje PRR hace posible clasificar diagramas y tipología (Prat et al., 2011). Describir cada columna como una regla por el formato $\$f(ab * n)$ optimiza los problemas (Joglekar et al., 2019). Un mapa auto-organizado que trabaja como un algoritmo de aprendizaje no-supervisado para mostrar la visualización de datos multivariables al producir clusters iniciales, y al principio solo mostrando cluster representacionales, hace posible heredar la estructura global (Johansson et al., 2004). La tecnología de aplicación sobre IA hace posible resaltar problemas de producción y fácilmente analizar la información (Chang et al., 2005). El uso de métodos de tipo vector para validar cada operación de DD puede ser aclarante si cada método es realmente eficiente (Zhang et al., 2007). El enfoque hacia el campo de flujo de trabajo al tomar puntos de vista data-céntricos se hace posible sólo si los procesos están conectados pour un registro y el sistema está disponible para conectar procesos con diferentes formatos de datos. (Robinson et al., 2009). Usando arquitectura plug-in, la cual permite el desarrollo modular para la obtención de datos, permite la construcción de sitios web con operaciones sofisticadas de DD (Egenly et al., 2010). Un enfoque para descubrir conocimiento a través de la integración de sumas y presentación de técnicas reduce el tiempo de búsqueda e identificación de información (Fung y Thanadechteemapat, 2010). Una sabia apreciación del sistema y de ejecución de tareas reduce el volumen de código, separa los datos y apuntala el código abierto para diversas

herramientas (Klimentov et al., 2011). El análisis de datos lineal es mucho mejor cuando un mapa de árbol es adaptado con el calendario y se usa el tiempo como el principal atributo jerárquico (De Carvalho et al., 2016). Para conducir métodos de ciencia reproducible, es necesario acumular el rastreo al agrupar los márgenes y nodos con la misma derivación (Xiang et al., 2016). Diseñar y desarrollar el proceso de diseño desarrolla información ejecutiva (Putra et al., 2019). La visualización OLAP propuesta con vistas de análisis tripartito genera expresiones multidimensionales (Zou et al., 2019). Para la implementación de un nuevo data-cube un algoritmo jerárquico es necesario para implementar la indexación espacial y técnicas no relacionales (De Melo et al., 2021). Las vistas DD son ajustadas para percibir ruido durante el análisis de datos (ruido en partes mecánicas vibratorias) permitiendo optimizaciones por diseño y la habilidad de estudiar el ruido durante la simulación de datos más rápidamente (Splechtna et al., 2023). El desarrollo de una herramienta basada en una gráfica de pastel se beneficia del análisis visual de datos categorizados (Guimares et al., 2011). Para explorar el comportamiento de datos sub-agrupados en presentaciones de análisis de aprendizaje (Shabaninejad et al., 2020) propone una perspectiva que recomienda un DD profundo en usuarios de LAD. La arquitectura para resolver consultas NoSQL en almacenes de big data, cuyos resultados precomputados en el sector granular de las colecciones y que son desagrupados, prueba una efectividal del modelo al aplicar consultas DD y drill-up en evaluaciones experimentales extensivas (Franciscus et al., 2018).

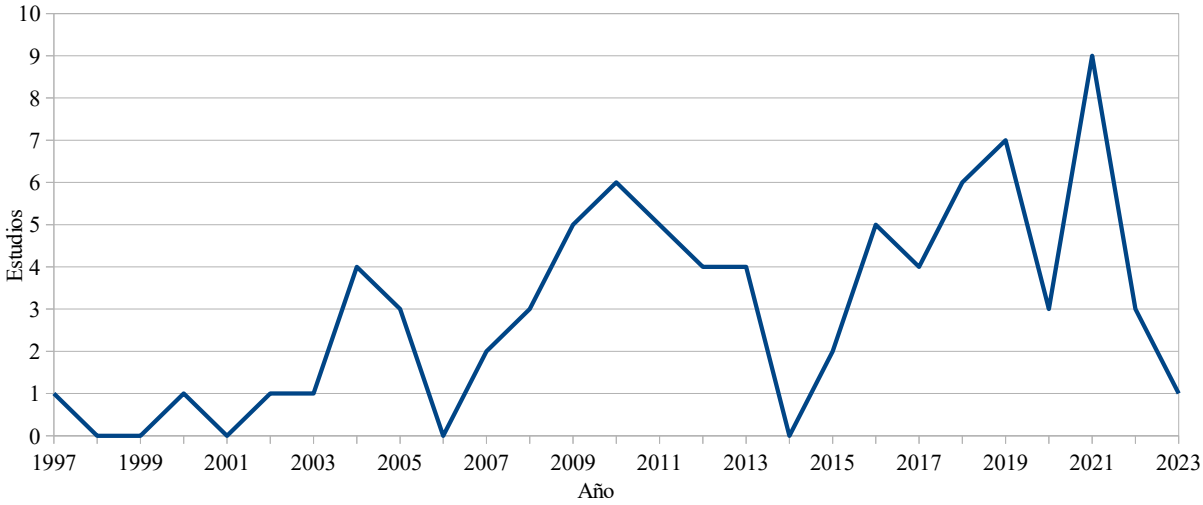
La Figura 7 esquematiza la línea de tiempo de los estudios analizados.

Metodología Experimental

Sen et al. (Sen et al., 2009) muestran que las operaciones OLAP en modelos multidimensionales son posibles después de añadir pequeños cuboides particionados dependiendo de su cardinalidad. La entropía es maximizada cuando sus principios de información son usados para determinar las bases de datos proxy. (Pourabbas y Shoshani, 2010). Técnicas de minería de opinión y herramientas de visualización cuantifican la opinión de los votantes (Soulis et al., 2013). Un solución analítica enfocada en métricas de equipo

Figura 7

Línea de tiempo de los estudios analizados.



se permite para un diseño visual y la navegación (Augustine et al., 2018). La función de regresión y predicción hace posible la detección de tendencias (Ragavi y Geetha, 2021). El uso de consultas dinámicas en línea en capas de datos establece correlaciones, tendencias o resaltar identificaciones (Mathrani, 2021). Se ha usado la herramienta MediSyn para selección, conexión, elaboración, exploración y distribución de conocimientos a través de interacciones (He et al., 2021). La agregación grupo-por-grupo para evaluación del rendimiento crea alternativas posibles para mover grupos de objetos (Baltzer et al., 2013). Para implementar información funcional, la recuperación es de utilidad para combinar técnicas de búsqueda por categoría (Lee et al., 2000). En algunos casos, vistas materializadas de cubos OLAP pueden originarse desde modelos de datos dentro de definiciones multidimensionales y jerárquicas (Palza et al., 2003). El uso de memoria flash en entornos de energía eficiente es posible a causa de sensores de red centrados en el almacenamiento (Tang et al., 2007). El algoritmo completo y mejorado Glide para actualización de vistas elimina las anomalías de datos (Chen et al., 2008). En algoritmos de consultas, decrecen los cubos cerrados en paralelo y el número de bloques de datos se incrementan (You et al., 2008). La colección de datos se hace posible a causa de un

lenguaje restrictivo y de multicapas basado en fuera de línea y análisis DD (aresi y Guinea, 2013). Tiempo de respuesta acortado en evaluaciones de rendimiento se hace posible a través de evaluaciones experimentales en software de open-source (Bianchi et al., 2013). Ejecutar varias instancias en ventanas de tamaño fijo es hecho posible gracias a un algoritmo que soporta el tráfico intenso (Kotamsetty y Govindarasu, 2016). Tendencias y estadísticas para realizar análisis son hechas con la ayuda de entornos soportados por colección y acumulación de eventos (Kritzinger et al., 2017). La inspección de malware en redes de datos permite la activación o desactivación de tiempos de verificación (Lee et al., 2018). La efectividad en los modelos de simulación espacio-temporales provee retroalimentación en datos geo-espaciales (Afzal et al., 2020). La perspectiva basada en rendimiento y procesos muestra errores que se pueden producir junco con árboles de decisión manuales (Khosravi et al., 2021). Nivles de alertas Gaussianas sobre incidentes inalcanzables son posibles gracias a un promedio simple en el nivel de pings de HTTP (Agrawal et al., 2021). Análisis de confiabilidad de rendimiento en grandes datasets es posible al extraer datos de transacciones con un modelo rápido (Franklin, 2021). La comparación y la identificación es posible con un ajuste en la visualización y la jerarquización (Conklin et al., 2002). En la rama de la educación, los datos pueden ser integrados, analizados y procesados con el sistema de aplicación Panda (Ikeda et al., 2012). Herramientas multidimensionales de análisis afectan el diseño de cada función (Zhang et al., 2012). El modelo de citación fue diseñado considerado que los objetivos de usabilidad y experiencia del usuario cumplan con la efectividad, la eficiencia y el aprendizaje (Hartono y Widyantoro, 2016). Las instancias de DW que utilizan sistemas orientados a documentos hacen posible la comparación de modelos y modelos cruzados (Chavalier et al., 2016). La eficiencia en consultas pesadas y frecuentes reside en algoritmos específicos (Ben Basat et al., 2018). Las limitaciones establecidas por listas de datos modeladas pueden ser redefinidas por consultas OLAP (Vassiliadis et al., 2019). El potencial de la narrativa visual de E-learning puede ser demostrada con un enfoque de narrativa (Chen et al., 2019). La etiquetación de nodos en los árboles de jerarquías hace posible seleccionar categorías de tablas al implementarse (Yin et

al., 2015). Ubicar objetivos es más exactos al usar un algoritmo de distorsión desde el diseño de ojo de pescado (*fisheye*) (Shi et al., 2005). La retroalimentación de modelos de avanzada y retroceso de textos soportados por esquemas de peso residen en los fundamentos de contrastes de clusters (Ziegler et al., 2008). El tamaño de elementos de interfaces, incluyendo diseños de emociones, pueden conducirse con el uso de jerarquías de conceptos (Ilyas et al., 2022). El algoritmo para usar estructuras dinámicas de datos identifica una conexión Galois con abstracciones bien definidas y funciones concretizadas (Sen y Chaki, 2011). Redes multi-capas como modelos de datos hacen posible generar diagramas EER, modelos flexibles y sostenibles (Santra et al., 2022).

Metodología Post-Facto

Odoni et al. (2018) presentan Orbis, un entorno extendible para análisis DD con múltiples tareas de anotación y versionamiento que hace posible reconocimiento de entidades, disambigüedades y tipificación de entidades. Por otro lado, la generación de conocimiento genérico necesita un gran conjunto de reglas para deducir las básicas a través de análisis semántico (Grabot, 2020). El uso de un panel que dirige la introspección a nivel de instalaciones hace posible identificar problemas probables y recuperar ocho indicadores de rendimiento visualizados en diversas vistas las cuales permitirán análisis DD en datos específicos (Lechner et al., 2022).

La Tabla 5 muestra los problemas resueltos.

Tabla 5

Problemas resueltos

Metodología	**Problemas resueltos**	**Porcentaje**
Comparative	17	18.28 %
Descriptive	34	36.56 %
Experimental	42	45.16 %
Post-facto	7	7.53 %

2.1.2. Problemas No Resueltos

A pesar de la dificulta que implica el desarrollo de la investigación científica y los errores y las imprecisiones que aparecen frecuentemente, no todos los autores reportan las fallas o circunstancias difíciles durante los procesos de investigación. Aquellas reportadas, listadas por categoría, se presentan en seguida.

Metodología Descriptiva

En 2018, Jiménez (2018) presentó que para crear un esquema de diseño conciso, e proceso de agrupamiento se debe ajustar para entender la relación entre orígenes de datos y tal esquema debe ser actualizado para prevenir futuros problemas. La arquitectura plug-in, que permite desarrollar módulos residentes en el servidor, no permite dicha expansión (Egenly et al., 2010). El desarrollo de una herramienta basada en un diagrama de pastel está basado en un estudio de usabilidad reducido (Guimares et al., 2011). La arquitectura para resolver consultas de big data o depósitos de resultados pre-computados en NoSQL para sectores granulados de colecciones son desagrupadas, notándose que la arquitectura propuesta fue sólo probada en específicos casos de estudio y que se consideró de importancia temporal dada su escasa granularidad (Franciscus et al., 2018).

Metodología Experimental

Mathrani (Mathrani, 2021) practicó con consultas dinámicas en linea sobre capas de datos que mostraron que la distribución no estaba lista para permitir la evaluación de entendimiento durante el rendimiento. La perspectiva sobre el rendimiento y el proceso sugirieron que debieron haberse producido errores a causa de overfitting (Khosravi et al., 2021). Los algoritmos de gran intensidad se han visto ligeramente sobre cargados (Ben Basat et al., 2018). Las redefiniciones de consultas OLAP no permiten ver las listas de problemas a resolver (Vassiliadis et al., 2019). La etiquetación de nodos en los árboles jerárquicos informan de la ausencia de algybas etiquetas, por lo que se deben realizar lecturas al detalle (Wang et al., 1997). Las jerarquías de conceptos que tienden a mostrar interfaces dinámicas carecen de aplicaciones móviles (Ilyas et al.,

2022). Redes multi-capas, usadas como modelos de datos, no reportan la existencia de una verificación total (Santra et al., 2022).

Metodología Post-Facto

Adicionalmente en Orbis, Odoni et al. (Odoni et al., 2018) notaron que las tareas de notación múltiple y versionamiento no integran pruebas de significancia estadística, ni construyen plug-ins para monitoreo, ni desarrollan soporte para evaluaciones extras. La creación de conocimiento base algunas veces es omitido durante las investigaciones, a pesar de que es un paso obligatorio para mejorar la adopción de estas técnicas (Odoni et al., 2018). El uso de páneles lleva a la introspección a nivel de instalaciones, haciendo posible identificar probables problemas, aunque por cuestiones de eficiencia no todos los datos fueron incluidos (Lechner et al., 2022).

La Tabla 6 muestra los problemas no resueltos.

Tabla 6

Problemas no resueltos

Metodología	**Problemas no resueltos**	**Porcentaje**
Descriptive	4	36.36 %
Experimental	7	63.64 %
Post-facto	3	27.27 %

2.1.3. Determinismo

El análisis DD aplicado lleva a la implementación de fórmulas determinísticas, en las experimentaciones que asignaron a cada columna el formato f(a,b,*,n) se resolvieron problemas de optimización (Joglekar et al., 2019). En el mismo rango de ideas, al entender el comportamiento de subgrupos de datos, los cuales se derivan de los árboles de decisión, y añadir filtros de forma progresiva, se obtuvo que la desviación de una tendencia se atribuye a un cambio local o *drill-down fallacy* (DDF) (Jung et al., 2019), lo que en otras palabras es una falla de interpretación de la información obtendia.

Por otro lado, en la investigación de learning analytics dashboards (LAD), la cual exploró el comportamiento de subgrupos de los árboles de decisión y añadió filtros progresivamente, se propuso un enfoque que recomendó análisis DD a profundidad en los usuarios (Shabaninejad et al., 2020). *Orbir*, un entorno de software extendible que soporta DD análisis, tareas de notación múltiple y versionamiento, permitió el reconocimiento de entidades, disambigüedades y titpificación de entidades, pero no pudo integrar pruebas de significancia estadística, crear plug-ins para monitoreo y desarrollar soporte para evaluaciones adicionales (Odoni et al., 2018). Al utilizar jerarquización para agrupación de datos y un modelo de DW híbrido tanto para extracción como análisis, ser resolvieron obstáculos en los algoritmos de procesamiento de DM y *data cube* (AlJanabi y Kadim, 2017).

Otros de los estudios establecieron que al presentarse una arquitectura para resolver consultas BD en depósitos NoSQL que precomputaron los resultados de sectores granulados de colecciones, se demostró la efectividad del modelo al aplicar consultas de DD y *roll-up* aunque la arquitectura propuesta solo fue probada en casos de estudios específicios (Franciscus et al., 2018). Durante la creación de un esquema diseñado con precisión, e proceso de agrupar datos se debió ajustar para entender la relación entre los diferentes orígenes de datos, lo que produjo un diseño que implementó una arquitectura robusta, documentada y actualizable, aunque se debió mantener actualizado dicho esquema para prevenir problemas futuros (Jiménez, 2018). Al generar un gran número de reglas, como primer paso, y después DD, se obtuvo una regla básica usando análisis semántico alternativamente, lo que dio por resultado que la evaluación sistemática de una gran regla básica permitió no solo verificar conocimiento genérico ya conocido sino también encontrar alguno que otro inesperado. Sin embargo, el análisis de conocimiento base, a veces descuidado en la literatura, se convirtió en un paso obligado para mejorar la adopción de estas técnicas (Grabot, 2020). Por último, la utilización de un panel que conduce a la introspección al estado a nivel de instalaciones hizo posible identificar posibles problemas pues ocho indicadores de rendimiento fueron visualizados en varias vistas, las cuales permitieron un análisis DD en datos específicos aunque, para una

mejor eficiencia, no se incluyeron todos los datos (Lechner et al., 2022).

2.1.4. Modelos Probabilísticos y Determinísticos

Los sistemas probabilísticos describen instrucciones que tienen por objeto inferir conclusiones estadísticas desde una mezcla compleja de datos inciertos y observaciones reales (Gilles et al., 2020). Existen dos razones principales para optar un enfoque probabilístico: primero, porque es el enfoque óptimo para la toma de decisiones bajo la incertidumbre y, segundo, los modelos probabilísticos son el lenguaje usado por las áreas en ciencia y tecnología (Murphy, 2022).

Por otro lado, un modelo determinístico es aquel que siempre que se someta a un mismo estímulo reacciona de la misma manera, que carece de incertidumbre y que se puede predecir con certeza mientras su comportamiento se evalúa con medidas de efectividad o eficiencia (Marco Teórico, 2023). Los modelos determinísticos se clasifican en tres métodos de programación que son: programación lineal, lineal entera mixta y algoritmos. Algunas de las técnicas utilizadas en la programación lineal son la identificación de las variables que influyen en las pérdidas de insumos, lineal difusa para mejoras en las cadenas de suministro y lineal integrada en los aspectos heurísticos; sus limitaciones se establecen cuando todos estos métodos tienen como objetivo maximizar las utilidades o minimizar los costos. En el caso de la programación lineal entera mixta, se requiere que las variables tengan tanto valores enteros como valores no negativos con lo que se pueden obtener resultados de coordinación y control para estudios posteriores. Sin embargo, se observa una restricción para dos variables cuando son enteras y se utilizan números binarios. En el caso de los algoritmos, estos se emplean debido a la complejidad existente en los sistemas productivos y el objetivo que se desea alcanzar, siendo una solución a los problemas que no pueden ser desarrollados por métodos convencionales utilizando diferentes tipos como algoritmos multiobjetivos, genéticos y otros (Carabalí et al., 2017).

2.2. Neural Network

Artificial neural network learning algoritm, o ANN, es un sistema de aprendizaje computacional que usa una red de funciones para aprender y traducir la entrada de datos en una salida deseada, usualmente en otro formato. Su concepto está inspirado por la biología humana y la forma en que las neuronas funcionan juntas para entender las entradas de los sentidos humanos (DeepAI, 2023).

Una NN se compone de neuronas, capas, funciones de activación y salidas. Existen varios tipos de neuronas: input (entrada), output (salida), hidden (ocultas), bias (sesgos) y context (series). Las funciones de activación establecen lazos entre las neuronas. Algunas de las más importantes son: linear activation functions, step activation functions, sigmoid activation functions, hyperbolic tangent activation functions, etc. (Heaton, 2015).

Las técnicas de aprendizaje supervisado construyen modelos predictivos al aprender de un gran número de ejemplos donde cada uno tiene una etiqueta indicando una ya comprobada salida (Zhou, 2017), son alimentados de entradas y salidas, además de proveer resultados sobre la precisión de la predicción durante el proceso de entrenamiento (Saravanan et al., 2018).

Por su parte, el aprendizaje no supervisado se realiza frecuentemente como parte de un análisis exploratorio, lo que ocasiona dificultades al intentar obtener productos ya que no hay un mecanismo universalmente aceptado para realizar validación cruzada o validar los resultados en grupos de datos independientes (James et al., 2023) y uno de sus principales problemas es cómo encontrar las estructuras ocultas de datos sin etiquetar (Dike et al., 2018).

2.3. Multilayer Perceptron

MLP es una NN que consiste en tres tipos de capas: de entrada, de salida y oculta. La capa de entrada recibe la señal que será procesada. La tarea de predicción y clasificación es realizada por capa de salida layer. Un arbitrario número de capas ocultas que

están ubicadas entre las dos anteriores son el verdadero motor computacional de MLP. Similar a una red feed forward, o proyectada, en donde se alimentan los valores de una capa previa a la siguiente, en una red MLP los datos fluyen hacia adelante desde la capa de entrada a la capa de salida. Sus neuronas son entrenadas con algoritmos de backpropagation, o retropropagación, y está diseñada para calcular cualquier función continua, pudiendo resolver problemas que no son linealmente separables. Los principales usos de una red MLP son clasificación de patrones, reconocimiento, predicción y aproximación (Abirami y Chitra, 2020).

3. Hipótesis

Los diagnósticos y predicciones de tendencias derivativas determinan el abatimiento de los costos financieros al relacionarse directamente con la reducción del overfitting en los almacenes de datos a través del análisis drill-down basado en multilayer perceptron neural network considerando la variación del precio en el tiempo.

4. Objetivos

En esta investigación se analizará cómo las nuevas tecnologías de hardware y el aumento exponencial de volúmenes de datos han obligado a la utilización de herramientas de AI, cuya implementación es una consecuencia natural del avance de la tecnología de software. Las técnicas de AI pueden producir predicciones y asociaciones más precisas (Achar y Syesh, 2019). El propósito de esta investigación es proponer la implementación de herramientas de multilayer perceptron neural networks (MLP) para potenciar el análisis DD en los DW de las presente y futura generación tecnológicas y así apoyar el diagnóstico y la predicción durante la toma de decisiones con una perspectiva de aprendizaje retroalimentativo, reduciendo el impacto del overfitting.

4.1. Objetivo general

Crear una arquitectura de software que reduzca el problema de overfitting en almacenes de datos financieros DW a través de la implementación de MLP durante el uso de análisis DD.

4.2. Objetivos específicos

- Diseñar una arquitectura de modelado de MLP, llamada MLP-DD, que tenga como variable de control el precio, que reescriba las herramientas de DD en DW, por medio de la construcción de definiciones matemáticas que mejor se adapten a la naturaleza de dichas herramientas.
- Construir un esquema de parámetros que permitan el análisis profundo de diferentes bases de datos, adaptando el modelo MLP-DD a algoritmos operativos de lenguajes de programación como Java o Python.
- Experimentar con los diferentes aspectos del sistema de parámetros, comprobando la eficiencia del modelo MLP-DD, mediante la ejecución de diversas pruebas en bases de datos aleatorias y operacionales de diferente categoría y naturaleza.
- Compilar las métricas obtenidas durante el entrenamiento, constatando la eficiencia del modelo MLP-DD, a través de la depuración de las pruebas realizadas al modelo.

5. Materiales y Métodos

5.1. Diseño de la investigación

La investigación será del tipo cuantitativo y se llevará a cabo de forma deductiva. Con respecto al diseño, será de modelo experimental verdadero de Nivel III, ya que requiere conocimiento considerable previo, comprobar hipótesis o teorías predictivas

y las preguntas responderán sobre la efectividad entre las variables independientes y dependientes (Sousa et al., 2007).

Será de raciocinio deductivo porque iniciará con la estructura establecida de las MLP cuyos parámetros se irán precisando hasta obtener los resultados que satisfagan las variables. Además, su diseño será experimental porque incluirá tareas aleatorias, grupos de control y manipulación de variables independientes y dependientes que serán cuantificadas.

Analizando las limitaciones y problemas no resueltos de la revisión sistemática anterior, se determinaron las siguientes variables independientes y dependientes que habrán de conducir la investigación:

5.2. Variables independientes

De acuerdo al análisis de los antecedentes, se ha determinado que las variables independientes son: *tendencias derivativas* y *costos financieros*, en base a las siguientes causas:

5.2.1. Tendencias Derivativas

En la mayoría de los problemas de optimización producto de aplicaciones científicas, de ingeniería o de AI, las funciones restrictivas y objetivas están disponibles únicamente como resultado sin proveer información derivativa (Larson et al., 2019). Al analizar las propiedades derivativas de pequeños objetivos, las fórmulas primitivas son mejoradas al incorporar información ya derivada, la medida del contraste, por lo tanto, se construye para mejorar un pequeño objetivo y suprimir el ruido en cada subrama derivada (Bai et al., 2018). Al aplicar las técnicas de DD y establecer las tendencias derivativas no establecidas en investigaciones previas, se busca desenfocar la atención de la investigación de sistemas cuya análisis sigue delimitado por paradigmas.

5.2.2. Costos Financieros

Existen dos enfoques para la identificación y la evaluación de finanzas corporativas: la primera se basa en la divulgación de las operaciones en términos de sus dos dimensiones principales como son la cantidad y la cualidad, la segunda se basa en la no divulgación que utiliza los valores de algunas variables observables tales como los desafíos empíricos relacionados con peticiones causales y de cómo son evaluadas en su validez y pragmatismo (Hassan et al., 2019). El rendimiento ha sido usado como indicador de productividad y valor de marca, la correlación entre estas variables y su rendimiento dentro del enfoque de responsabilidad social se ha examinado a través de análisis de correlación y regresión, los resultados han confirmado que su rendimiento tiene una correlación positiva con la productividad y el valor de marca (Cho et al., 2019). La presente investigación busca establecer dichas variables como parámetros de configuración durante el entrenamiento del modelo a seguir.

5.2.3. El precio desde una perspectiva matemática

Existen dos formas de determinar el precio: la forma determinística o *contable* y a través del entrenamiento de una ANN o *trained*, los cuales están definidos por los siguientes teoremas:

Theorem 5.1 (Precio Determinístico (contable)) *Sea el precio definido por la suma de los costos más impuestos multiplicados por el factor de utilidad de acuerdo a la siguiente ecuación*

$$Precio_D = (CB + CO + CF + I)(1 + F) \tag{2}$$

donde CB es costo bruto, CO es costo operativo, CF es costo financiero, I es impuestos y F es factor de utilidad expresado en porcentaje, cuando

Corollary 5.1.1

$$Eficiencia = \begin{cases} 1 & if\ Precio_D\ \in\ Mercado : F\ \exists\ < Bull|Bear > \\ 0 & otherwise \end{cases} \tag{3}$$

donde el precio existe en el mercado y cuyo factor se encuentra en uno de los dominios bull o bear.

Theorem 5.2 (Precio Entrenado (trained)) *Sea el precio definido por los parámetros de la red neural, el entrenamiento y su multiplicación subsecuente por el factor de utilidad de acuerdo a la siguiente ecuación*

$$Precio_T = (MLP + Training)(F)$$

donde MLP es la red neural y F es factor de utilidad expresado en porcentaje, cuando

Corollary 5.2.1

$$Eficiencia = \begin{cases} 1 & if\ Precio_T\ \notin\ Overfitting : F\ \exists\ < Bull, Bear > \\ 0 & otherwise \end{cases} \tag{4}$$

donde el precio no está sobre-entrenado y cuyo factor se encuentra en uno de los dominios bull o bear.

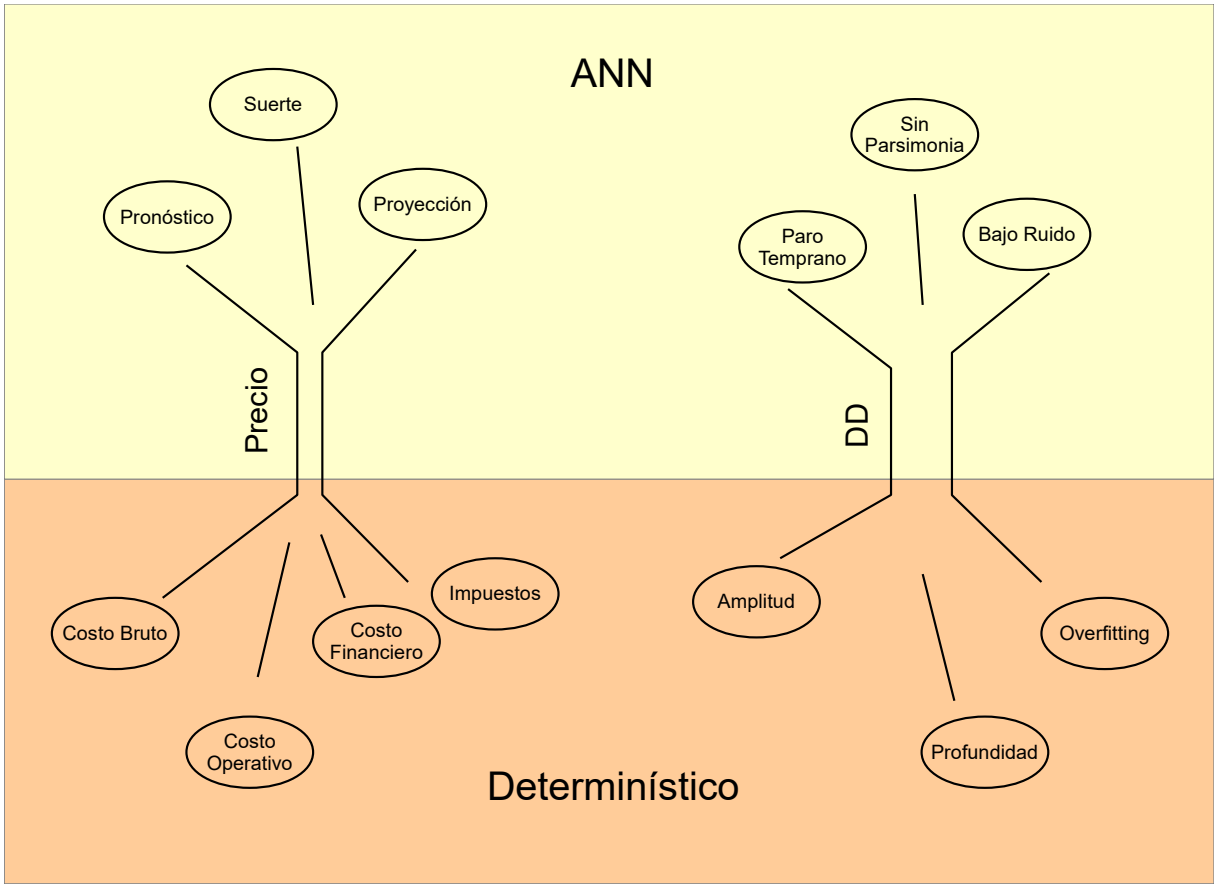
Los cuales se pueden interpretar a través de la Figura 8 que muestra la comparación de ambas problemáticas.

5.3. Variable dependiente

Por consecuencia, a través de la aplicación de la técnica DD sobre las tendencias derivativas y la correlación y regresión de los factores de productividad y marca, se analizará el comportamiento del precio en un esquema de datos entrenados.

Figura 8

Comparación entre las problemáticas del Precio y del análisis DD



5.3.1. Precio en el Tiempo

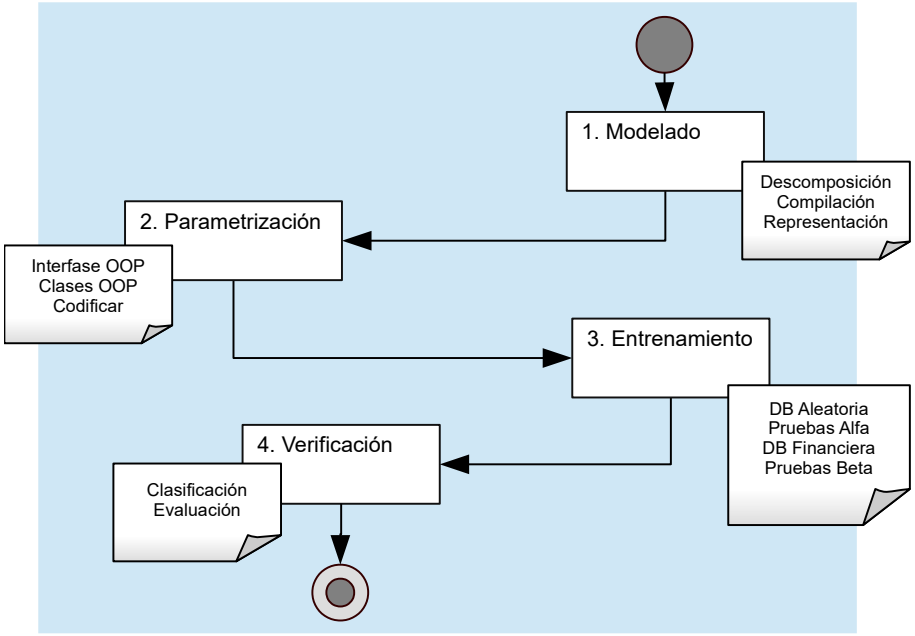
El precio tiene una correlación negativa con los promedios libres de riesgo en depósitos bancarios, pero una correlación positiva en el promedio de préstamos, su análisis estadístico sugiere que existen afectaciones a nivel macroeconómico a causa de riesgos económico, crediticio y políticos (Huy et al., 2020). En el caso de las criptomonedas, éstas están determinadas por el mercado beta (que mide es la sensibilidad de un activo respecto a los movimientos), el volumen de intercambio y la volatilidad en corto y largo plazo, y su atractivo sólo importa en la determinación del precio a largo plazo (Sovbetov, 2023). Se busca identificar a través del precio, como valor de salida (**output**), la comprobación segmentada de la hipótesis en referencia a cada ciclo de entrenamiento.

5.4. Proceso

La Figura 9 muestra el proceso general del proyecto y los subprocesos de cada etapa. El índice tentativo de la tesis, presentado más adelante, se encuentra indexado a este proceso.

Figura 9

Proceso general del proyecto.



5.4.1. Diseño de la arquitectura

Se logrará a partir de la implementación del análisis de las limitaciones de los modelos de DD y de MLP que actualmente se están utilizando tanto en la literatura científica como en los grupos de programadores profesionales, específicamente a nivel código en lenguajes de programación de uso especializado para redes neuronales.

5.4.2. Modelado

A partir de los algoritmos existentes en investigaciones previas, se diseñarán el modelo MLP-DD de acuerdo a las siguientes etapas:

Descomposición Descomponer los términos que componen MLP y DD para establecer sus puntos de coincidencia.

Compilación Compilar los puntos de coincidencia estableciendo dos relaciones: una procesal, para objetos de verificación, y una matemática, para efectos de definición.

Representación Representar el modelo obtenido en formato de matemáticas discretas.

5.4.3. Parametrización

Una vez definido el término MLP-DD, se esquematizarán los parámetros para poder analizar los datos experimentales a profundidad, de acuerdo a los siguientes pasos:

5.4.4. Entrenamiento por k-means

K-means es utilizado para clasificar datos en categorías o clústers llamados k basados en las proximidades de las distancias euclidianas de sus centroides. En esta investigación el entrenamiento se realizó primero seleccionando los valores de k de forma aleatoria a partir del dataset, después cada registro de acuerdo a su distancia euclidiana fue asignado a su correspondiente k y una vez terminado, a través de n épocas el centroide de cada k fue ajustado para obtener el punto óptimo de su posición cartesiana. Si bien el modelo se diseñó para calcular centroides y distancia euclidiana de objetos en tercera dimensión (ejes x,y,z), las dos dimensiones proporcionadas por el dataset se obtuvieron igualando los valores de z a 0.

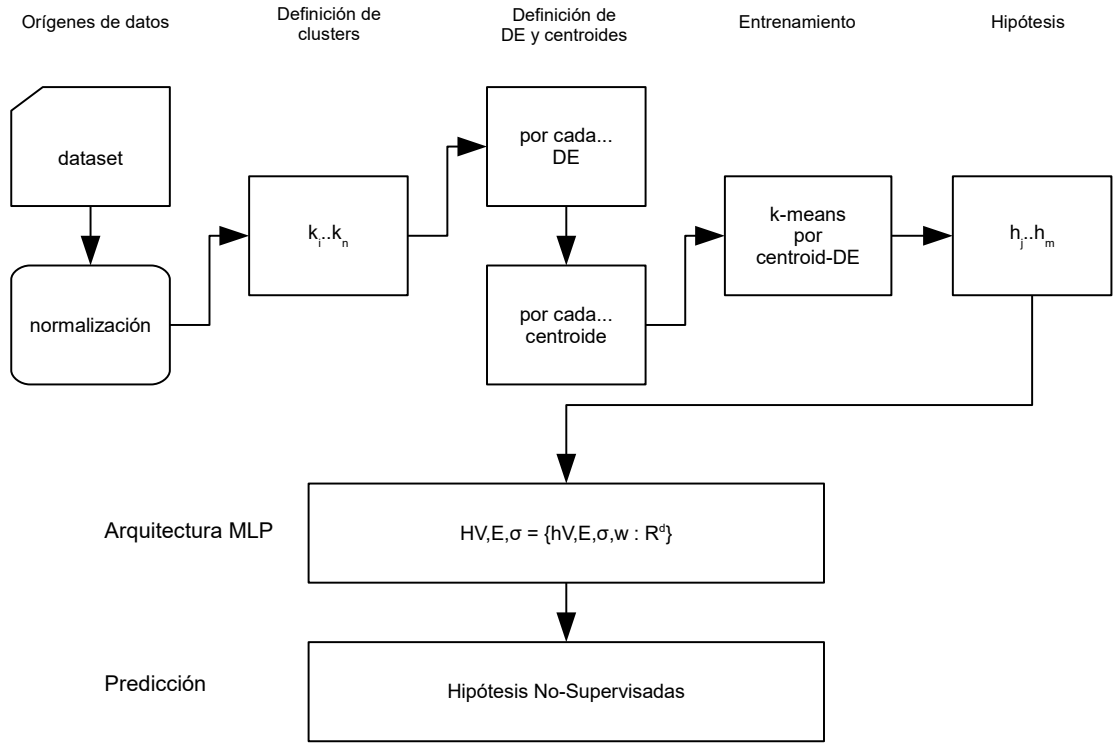
5.4.5. Modelo Parametrizado de MLP y Kmeans

La arquitectura del MLP la ANN se define por la tripleta $(V; E; \sigma)$, en donde V son el número de capas, E son los límites que se corresponden a los pesos w y $(V; E; \sigma)$ es la función de activación. De esta manera, una vez agregado el modelo K-means, el modelo completo queda representado por la Figura 10.

El enlace entre ambos se realiza a partir de la hipótesis de consulta a grupo. Las capas V quedan definidas de la siguiente forma: V_1 es la capa de entradas con 2 neuronas

Figura 10

Arquitectura del modelo integrado MLP-Kmeans.



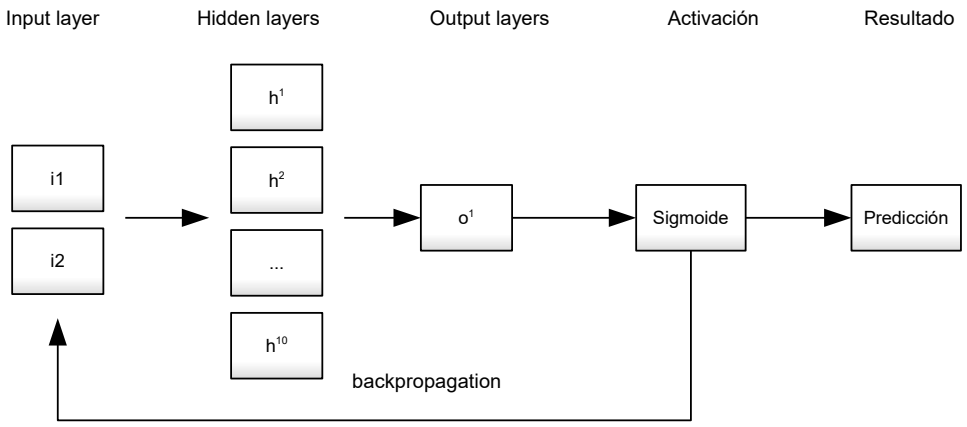
correspondientes a los valores de k recibidos de Kmeans. V_2 es la capa oculta con 10 neuronas y V_3 es la capa de salida con 1 neurona correspondiente a la predicción o hipótesis en particular. El valor de σ corresponde a la función *sigmoide*. Para esta investigación cada neurona requiere 3 valores de entrada $X + 1$ que corresponden a los valores de $W + bias$ con los cuales calcula el *inner product*, define el sigmoide y transfiere el valor a las siguientes capas, como se puede observar tanto en la Figura 11 como en el seudocódigo del algoritmo 1.

enumerar líneas Por lo anterior, la presente investigación se enfoca en generar a través de la clasificación de modelos supervisados patrones de comportamiento en las variables dependientes.

Interfase OOP Diseñar un objeto paternal, con la funcionalidad de interfse, de acuerdo a la filosofía de object oriented programming (OOP), que represente los elementos

Figura 11

Definición de MLP



Algorithm 1 Definición de las capas del MLP enlazadas con Kmeans.

Require: $k > 0$
Ensure: $\sigma < X, V >$
  function LOOP(*epoch*[ ])
    function LOOP(Dataset)
      $loss \leftarrow prediction$
      $error \leftarrow MeanSquareLoss$
    end function
    return $Output \leftarrow prediction$
  end function

del modelo MLP-DD y asigne los atributos, funciones y procedimientos requeridos para la operación de las clases llamadas descendientes o hijas. El nombre piloto de la interfase será MlpddInterface.

Clases OOP Teniendo a la interfase MlpddInterface como padre, se producirán clases descendientes totalmente operacionales, que implementarán las operaciones necesarias del modelo MLP-DD. Las clases, por ejemplo, podrán nombrarse MlpRecopilacionData, MlpClasificacionData, y sus descendientes MlpRecopilar, MlpClasificar, etc.

Codificar Seleccionar el o los lenguajes en que habrán de codificarse interfases y clases – aquellos en paréntesis son los propuestos –, considerando un mínimo de:

1. Lenguaje SQL para consultar la base de datos (PL/SQL),

2. Lenguaje de alto nivel para análisis (Python), y

3. Lenguaje de alto nivel para interacción con el usuario (Java).

5.4.6. Entrenamiento

Para comprobar la eficiencia de la arquitectura desarrollada para el modelo MLP-DD, se llevarán a cabo una serie de pruebas de acuerdo a las siguientes actividades:

DB aleatoria (DBa) Se creará una base de datos aleatoria con una dimensión mínima de 100 mil registros generados a partir de las funciones random, y que ocuparán espacio temporal en la memoria alta del sistema de cómputo objeto.

Pruebas alfa Se ejecutarán los códigos de las clases obtenidas en la codificación sobre los datos de la DBale, manteniendo un registro de las actividades que incluirá una lista de funciones y procedimientos a corregir en el código.

DB financiera (DBf) Se obtendrá una de las bases de datos de información financiera y más de 10 millones de registros disponibles a través del Harvard Dataverse en (Balogh y Attila, 2023).

Pruebas beta Se ejecutarán los códigos ya corregidos sobre la DBf, para la recopilación de actividades y generación de datos estadísticos.

5.4.7. Verificación

Habiendo compilado los datos de las pruebas beta, y obteniendo sus métricas, se clasificarán y evaluarán para la obtención de los resultados.

Clasificación Por cada propuesta de investigación se clasificarán las variables de investigación del dataverso por medio del algoritmo K-means mejorado con *Coordenadas Parametrizadas*, el cual se abordará posteriormente. Una vez realizada la clasificación se aplicará el algoritmo de MLP para su entrenamiento.

Evaluación Cada una de las hipótesis particulares de investigación del dataverso se evaluará de acuerdo a los límites que hayan aportado sus clasificaciones.

5.4.8. Redefinición del Centroide

El centroide se define en su forma matemática como la suma de las distancias euclideanas mínimas para un conjunto de vectores (Slater y Peter, 1978). Se utiliza principalmente en dos ramas: en aplicaciones gráficas y en ciencia de datos. En cuestiones gráficas su determinación poligonal es funcionalmente estándar en software de aplicación para sistemas de información geográfica, Geographic Information System, o GIS. Una razón común para determinar el centroide es para obtener un punto de referencia conveniente de un polígono (Deakin et al., 2002). Se expresa como el centro de gravedad y como seudo-centroide en la solución de problemas de agrupamiento cuando los puntos no tienen coordenadas numéricas (Glover y Fred, 2016) como es el caso de afectaciones por temperatura. En cuestiones de ciencia de datos se utiliza en modelos de aprendizaje automatizado durante el entrenamiento de diferentes tipos de redes y sus aplicaciones van desde la clasificación hasta la detección de malware.

Infinity Type Centroid, ITC, extiende el concepto de seudo-centroide al tomar en cuenta como parámetros del cálculo todos los factores que intervienen en la posición particular del objeto y los integra como dimensiones complementarias a las coordenadas x, y y z que comprenden su cálculo. Desde esta perspectiva, por ende, ITC puede ser utilizado en todas las ramas de la ciencia.

Para obtenerlo el ITC se deben definir los parámetros de acuerdo a la rama de la ciencia que sea materia de estudio como, por ejemplo, en el caso de arquitectura agregar los vectores de gravedad y resistencia de materiales, y en el caso de medicina deportiva el voltaje del objeto en movimiento (el deportista) más su masa y peso que influyen directamente en su particular *centro de gravedad* o centroide, por mencionar algunos ejemplos.

5.4.9. Centroid Geométrico

El centroide geométrico para una figura en el plano cartesiano se define como la intersección de los puntos medios de x (anchura **base**), y (altura **height**) y z (profunidad **width**), siendo $\bar{x}$ el ancho de la figura dividido entre 3, $\bar{y}$ la altura de la figura dividida entre 3, y $\bar{z}$ la profundidad de la figura dividida entre 3, de acuerdo a la ecuación 5.

$$GeometricCentroid_{(\bar{x},\bar{y},\bar{z})} = (base/3, height/3, width/3) \tag{5}$$

El centroide obtenido a partir de los puntos A, B y C definidos en el espacio a partir de sus coordenadas cartesianas, se define para cada coordenada como la media de la suma de las coordenadas *(x, y, z)* dividida entre tres, de acuerdo a la ecuación 6.

$$Centroid = ((A.x + B.x + C.x)/3, (A.y + B.y + C.y)/3, (A.z + B.z + C.z)/3) \tag{6}$$

Que está representado por el teorema 7.

$$Centroid_{(A,B,C)} = (A_{|x,y,z|} + B_{|x,y,z|} + C_{|x,y,z|})/3 \tag{7}$$

La *Euclidean Distance* es la diferencia entre los centroides de dos objetos establecidos para establecer similitudes entre conjuntos de datos. Se define como la raíz cuadrada de la diferencia de los centroides, obtenidos a partir de la diferencia de sus coordenadas cartesianas, como indica la ecuación 8.

$$ED = \sqrt{(x' + y' + z')} \tag{8}$$

donde: ED es *Euclidean Distance*, $x' = (b.x - a.x)^2, y' = (b.y - a.y)^2$ y $z' = (b.z -$

$a.z)^2$.

5.4.10. Centroide Térmico

El centroide no únicamente sirve para calcular posiciones geométricas o de datos sino que también se utiliza en la definición de patrones de comportamiento de objetos en cuestiones físicas o químicas. El *centroide geométrico* tiene características propias que lo difieren del *centroide térmico*, afectado por la temperatura (Fang y Mingqiang, 2006). De la misma forma pueden encontrarse características de diversa índole que afectan la posición del centroide al considerar factores del entorno, por ejemplo, en el caso de la medicina, el centroide dependerá de la presión arterial, el peso y el ritmo cardíaco, o en la aeronáutica de los cambios que la velocidad producen sobre la masa y el peso del proyectil.

En todas las áreas de la ciencia existen afectaciones al centroide por diversos factores complementarios que forman parte integral e incidental en la obtención de resultados ponderables. La Tabla 7 muestra una aproximación de la parametrización a partir de unidades fundamentales del sistema internacional, SI.

Tabla 7

Unidades Fundamentales que pueden afectar el centroide en sus coordenadas x,y,z.

Símbolo	**Unidad**
m	metro
kg	kilogramo
s	segundo
A	amperio
K	kelvin
mol	mol
cd	candela

La Tabla 8 muestra otros ejemplos de disciplinas de la ciencia en cuyos estudios el centroide se ve afectado por diferentes parámetros, coincidiendo únicamente en la posición de las coordenadas gráficas iniciales.

Tabla 8

Factores externos considerados que pueden considerarse como parámetros, agrupados por rama de la ciencia y que se agregan a las coordenadas x,y,z.

Disciplina	**Parámetros**		
Oceanografía	t – temperatura		
Medicina	bp – presión arterial	w – peso	hr - ritmo cardíaco
Aeronáutica	s – velocidad	m – masa	

Para dar un ejemplo de la afectación por factores externos o complementarios, en el caso de la distancia euclideana entre dos planetas el centroide no sólo se determina por su posición con respecto a su estrella dominante sino también por su fuerza de gravedad; otro ejemplo es el del motor de un coche encendido que tendrá una temperatura $t > 0$ que afectará su distancia euclideana con respecto a las llantas antes de iniciar su movimiento ($t = 0$) que será diferente cuando ya estén en su trayectoria ($t > 0$). Para este caso el centroide integrado con la temperatura estaría definido en el teorema expresado por la ecuación 9.

$$Centroid_{(A,B,C)} = (A_{|x,y,z,t|} + B_{|x,y,z,t|} + C_{|x,y,z,t|})/4 \tag{9}$$

Los puntos A, B y C responden al producto de sus matrices donde 4 representa el número de coordenadas o columnas. Las 3 primeras coordenadas es posible representarlas gráficamente por lo que t es una representación no-gráfica. En consecuencia, para este caso en particular la Euclidean Distance quedaría definida por la ecuación 10.

$$ED_t = \sqrt{(x' + y' + z' + t')} \tag{10}$$

donde: ED_t es *Euclidean Distance con el factor de tiempo*, $x' = (b.x - a.x)^2, y' = (b.y - a.y)^2, z' = (b.z - a.z)^2$ y $t' = (b.t - a.t)^2$.

5.4.11. Centroide de Tipo Infinito

Al teorema que calcula el centroide a partir de las coordenadas x, y y z, y que considera un número cuasi-ilimitado de factores externos, o parámetros complementarios, adecuables a la disciplina científica que los utilice durante la resolución de un problema, se le llama *Centroide de Tipo Infinito*, ITC, el cual es un concepto que presenta este trabajo, y está definido por el teorema expresado por la ecuación 11.

$$ITC(A', B', C') = (A_{|x,y,z,[i_1,i_2 \to i_n]|} + B_{|x,y,z,[i_1,i_2 \to i_n]|} + C_{|x,y,z,[i_1,i_2 \to i_n]|})/(3+n) \quad (11)$$

Donde A, B y C son los puntos en el plano cartesiano del objeto, x, y y z sus coordenadas gráficas, i el iterador de los parámetros que se infieren hasta n, n es el número máximo de parámetros solicitados para el problema que se esté resolviendo. $ITC(A', B', C')$ se lee como el centroide ITC para los puntos primos, o derivados, con coordenadas parametrizadas.

Se demuestra en el teorema 11 que en el cálculo del centroide es necesario complementar las coordenadas de cada punto A, B y C con una lista dinámica de parámetros que pueda ser recalculada en tiempo real y que se refleje directamente en su valor. Entendiendo por *dinámica*: *no estática*, en otras palabras, que mientras las coordenadas gráficas existen de forma invariable en todo tipo de problema la otras dimensiones pueden o no pueden pertenecer, y ser definitivas o no definitivas, al comportamiento del objeto. A partir del teorema 11, se establece que a la unión de coordenadas cartesianas y lista dinámica de parámetros se llama *Coordenadas Parametrizadas* y está definida por el corolario expresado por la ecuación 12.

$$CoordenadaParametrizada = x, y, z, [i_1, i_2 \to i_n] : i \ \exists \ X \quad (12)$$

cuando el parámetro i existe en el dataset X, proporcionado por la investigación que se esté desarrollando en la rama de la ciencia que lo requiera.

En consecuencia, el cálculo de la Euclidean Distance entre dos objetos a partir de ITC se define por la ecuación 13.

$$ED_{ITC} = \sqrt{({ITC_1}^2 - {ITC_2}^2)} \tag{13}$$

donde: ED_{ITC} es *Euclidean Distance para ITC*, ITC_1 es el centroide del primer objeto e ITC_2 el centroide del segundo objeto, los cuales se componen de coordenadas parametrizadas de acuerdo al corolario 12 y que detallada quedaría en el corolario 14.

$$ED_{ITC} = \sqrt{ITC_{x'} + ITC_{y'} + ITC_{z'} + ITC_{[i' \to n']}} \tag{14}$$

donde:

- $ITC_{x'} = (ITC_{1.x} - ITC_{2.x})^2$,
- $ITC_{y'} = (ITC_{1.y} - ITC_{2.y})^2$,
- $ITC_{z'} = (ITC_{1.z} - ITC_{2.z})^2$, y
- $ITC_{[i'..n']} = [(ITC_{1.i_1} - ITC_{2.i_1})^2 + (ITC_{1.i_2} - ITC_{2.i_2})^2 \to (ITC_{1.n} - ITC_{2.n})^2]$

5.4.12. Análisis de datos

Históricamente el análisis de datos ha tenido tres vertientes: clásico, EDA y Bayesiano, en este trabajo se propone una cuarta a través de k-means, como se muestra en la Tabla 9. Este dataset, publicado por Harvard Datavers (Balogh y Attila, 2023), contiene actividad comercial interna de empresas de cotización pública sancionadas por la *Securities and Exchange Commission* (SEC) y ha sido analizado previamente para

determinar prospectos y ofertantes (Balogh y Attila, 2023). Sus registros se actualizan desde el origen diariamente y al cierre corresponden a la fecha de 30 de abril de 2024 con un volumen mayor a $10.6x10^6$ registros.

Tabla 9

Métodos históricos para el análisis de datos y propuesta k-means.

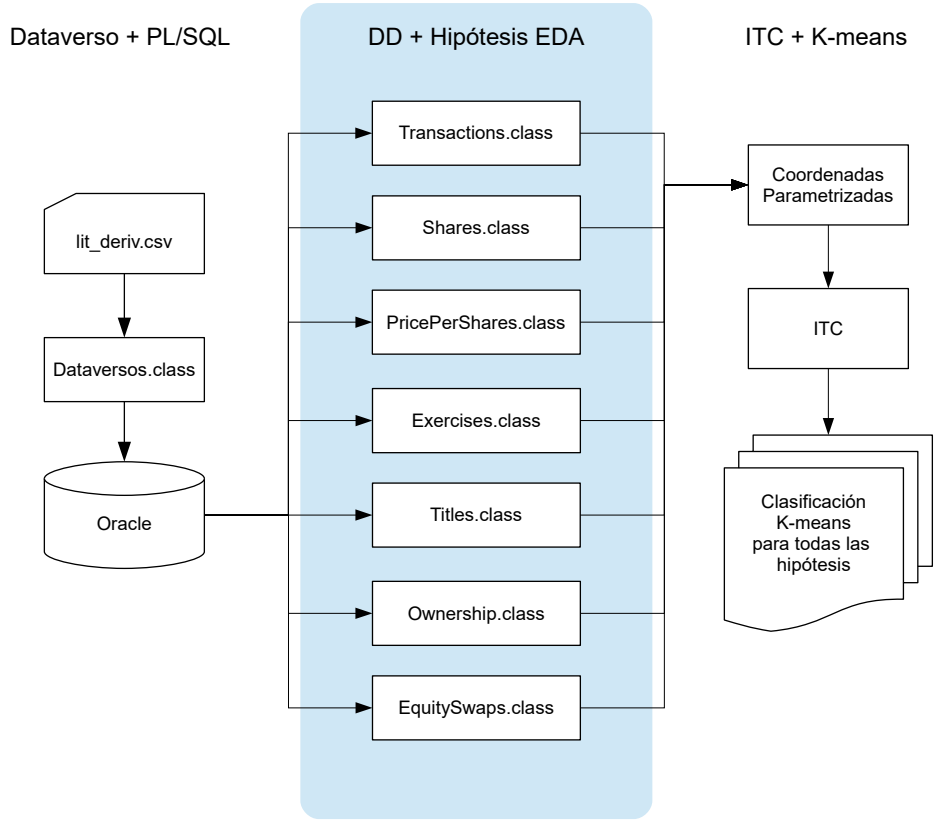
Clásico	**EDA**	**Bayesiano**	**EDA con K-means***
Definición del Problema	Definición del Problema	Definición del Problema	Lectura de Dataset
Colección de Datos	Colección de Datos	Colección de Datos	Convertir de Primitivo a Objeto
Desarrollo del Modelo	Análisis de Datos	Desarrollo del Modelo	Propuesta de Hipótesis DD
Análisis de datos	Desarrollo del Modelo	Distribución Previa	Configuración de Tipos
Resultados Comunicación	Resultados Comunicación	Análisis de Datos	Proceso en DB Implementación
		Resultados Comunicación	K-means Clasificación
			Resultados Comunicación

Nota. Fuente: (Kumar y Ahmed, 2020) *trabajo propuesto.

La arquitectura del software de aplicación utilizado en el proceso se muestra en la Figura 12.

Figura 12

Arquitectura del software de aplication.



5.4.13. Lectura del Dataset

Cuyo origen son los reportes reportes de insider trading utilizados por (Balogh y Attila, 2023) y disponibles a través de *HARVARD Dataverse* en (Harvard, 2023).

5.4.14. Convertir de Primitivos a No-Primitivos (Objetos)

Convertir los datos primitivos del dataset en formato de objectos a través de java, cuyo código está disponible en (Silva, 2024) realizando los siguientes pasos:

1. Extracción de los tipos de datos (Registrosdata.java)
2. Lectura del dataset, depuración y asignación de objetos (Registros.java)
3. Creación de lista de datos específica por hipótesis.

5.4.15. Establecer Hipótesis Matemáticas

Es el diseño del análisis a través de DD y la definición de las hipótesis de investigación por medio de EDA. Donde la hipótesis de investigación tiene una clase *ad hoc* que parte de los campos obtenidos a partir del dataverso. La Tabla 10 muestra las clases y sus objetivos.

Tabla 10

Descripción de las hipótesis de investigación

Hipótesis	**Objetivo**
transactionType	Determinar el tipo de transacción.
transactionShares	Monto de acciones por fecha.
transactionPricePerShare	Precios de la acción código de entrega.
exerciseDateFn	Función y fecha del ejercicio.
securityTitle	Títulos, acciones y seguros subyacentes.
directOrIndirectOwnership	Propiedad directa e indirecta de acciones.
equitySwapInvolved	Futuros y códigos de transacción.

5.4.16. Clasificación K-means

Es la aplicación de las coordenadas parametrizadas y obtención de ITC en la clasificación de K-means, de acuerdo a los valores propios de cada hipótesis, como se muestra en la Tabla 11.

Tabla 11

Parámetros adicionales de las hipótesis (considerando x,y,z como denominador común.

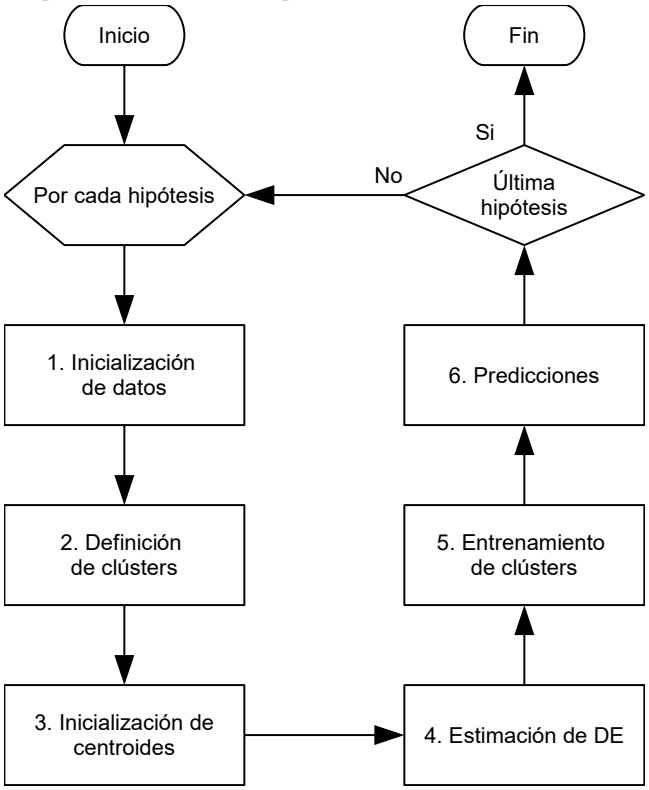
Hipótesis	**Parámetros**
transactionType	transactionType
transactionShares	transactionShares
	Year
transactionPricePerShare	transactionPricePerShare
	transactionPricePerShareMin
	transactionPricePerShareMax
	transactionAcquiredDisposedCode
exerciseDateFn	exerciseDate
	exerciseDateFn
securityTitle	securityTitle
	transactionShares
	underlyingSecurityShares
directOrIndirectOwnership	directOrIndirectOwnership
	transactionShares
equitySwapInvolved	transactionCode
	equitySwapInvolved

5.4.17. Entrenamiento y predicción

El proceso para la implementeación de las clases al proceso de K-means se presenta en la Figura 13.

Figura 13

Implementación del proceso de K-means.



[For each Hypothesis] corresponde a cada clase definida en la Tabla 10 anteriormente presentada. Los pasos se describen a continuación:

1. **Inicializaión de datos.** Los datos son cargados en listas (o datasets) dinámicas compuestas de tres puntos A, B y C de tipo *Coordenadas parametrizadas*.

2. **Definición de clústers.** Son seleccionados como clústers los valores más significativos, ya sea de forma arbitraria o aleadoria, que servirán para calcular los centroides.

3. **Inicializaión de centroides.** Los centroides son calculados para los clústers seleccionados.

4. **Estimación de DE.** Se calcula la distancia euclideana, para todos los puntos del dataset, con respecto a los centroides de los clústers seleccionados.

5. **Entrenamiento de clústers.** Se realiza el entrenamiento de los clústers.

6. **Predicciones.** Se comprueba la eficiencia del modelo al realizar predicciones sobre datos no entrenados.

El Listado 1 muestra la implementación práctica que siguen las hipótesis de investigación, en este caso el cálculo de *Security Titles*, y la solicitud de predicción para datos de prueba tipo beta, todo una vez realizado el entrenamiento. Dentro del código desarrollado para esta investigación, cada una de las propuestas tuvo su propio código *ad hoc*, aunque el mismo código puede es adaptable a cualquier tipo de librería.

Listing 1: Predicción y clasificación a través de centroides ITC y distancia Euclideana

```
private class KmeansForTitles extends Kmeans {
    /**
     * Constructor
     */
    public KmeansForTitles() {
    }
    /**
     * Ejecutar el algoritmo.
     */
    public void Execute() {
        Kmeans.PredictionResult="";
        Kmeans.Print=false;
        //paso 1
        DataInitialization();
        //paso 2
        ClustersDefinition(n_clusters);
        //paso 3
        for(int j = 0; j < Kmeans.Cluster.length; j++) {
            Datasdata ddata=Kmeans.Cluster[j];
            Random rn=new Random();
```

```
            int k=rn.nextInt(0, Titlesdata.listtitles.size());
            Titlesdata tdata=Titlesdata.listtitles.get(k);
            ddata.A.x=tdata.apoint;
            ddata.B.x=tdata.bpoint;
            Kmeans.Cluster[j]=ddata;
            //System.out.println(String.format("k=%d, %.3f, %.3f",k,ddata..
        }
        //step 4
        ComputeEachEuclideanDistance();
        //step 5
        //entrenamiento
        CentroidTraining(n_epochs);
        //step 6
        Datasdata Hy=new Datasdata();
        Hy.A.x=0.03;
        Hy.A.y=0.06;
        Hy.A.z=0.09;
        Hy.B.x=0.04;
        Hy.B.y=0.07;
        Hy.B.z=0.10;
        Hy.C.x=0.05;
        Hy.C.y=0.08;
        Hy.C.z=0.11;
        AssignCathegory(Hy);
        //resultado
        System.out.println(Kmeans.PredictionResult);
    }
    /**
    * Iniciar datos.
    */
```

```
    public final void DataInitialization() {
        Datasdata.Dataset=new ArrayList<>();
        Datasdata midato=new Datasdata();
        Titlesdata tidata=new Titlesdata();
        for (int i=0;i<Titlesdata.listtitles.size();i++) {
            tidata=Titlesdata.listtitles.get(i);
            midato=new Datasdata(tidata);
            Datasdata.Dataset.add(midato);
        }
    } //DataInitialization
} //KmeansForTitles
```

6. Resultados y Discusión

6.1. Resultados

Se presentan los resultados de los análisis de las hipótesis matemáticas de investigación, agrupando para el caso de K-means 2 clústers, seleccionados aleatoriamente a partir de cada muestra, y 10 épocas. Del número total de registros del dataverso se consideraron sólo los más significativos que pudieron normalizarse con respecto a la mediana de su frecuencia, quitando valores que no fueran significativos como por ejemplo el caso de que en un año hubiera menos de 100 registros y al siguiente más de 100,000, se estableció la política de normalización de sólo aquellos superiores a 10,000 registros coincidentes en su perfil.

De acuerdo a la metodología, primero se realizó un análisis DD para definir las hipótesis y seleccionar los campos con el perfil adecuado, después se normalizó el dataset retirando valores extremos, de ahí se procedió en base a EDA a calcular las tablas y graficarlas, a partir de ahí se utilizó el algoritmo de K-means para calcular los centroides dinámicos, produciendo las tablas dispersas, dejando listos los datos para su predicción, ejemplificada con el código del Listado 1.

A continuación, se describen los resultados.

6.1.1. *TransactionType*

Comportamiento por tipo de transacción. Los valores derivados, *Derivative*, son productos financieros que dependen de otro, por ejemplo, las compras fijadas a un precio al futuro. *Holding* es cuando la operación se encuentra en espera y *Transaction* cuando se ejecuta. La Tabla 12 muestra el volumen de registros para cada tipo. Los títulos en *holding* carecen de valor porque no están a la venta.

Tabla 12

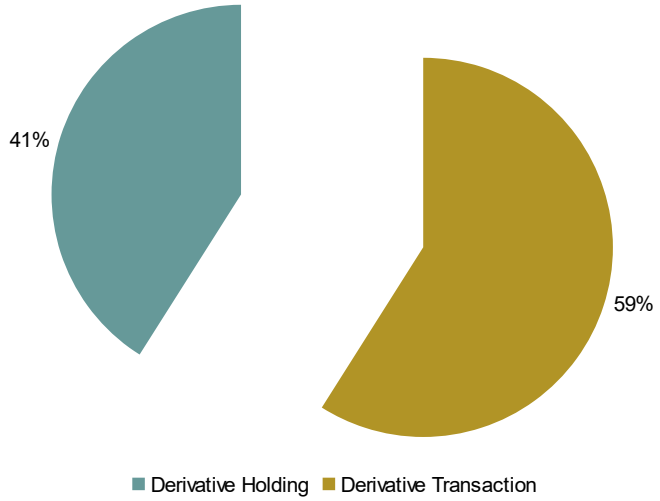
Tipos de transacción para información interna.

Tipo de transacción	**Registros**	**Valor (US$)**
Derivative Holding	$4.3x10^6$	0.00
Derivative Transaction	$6.2x10^6$	$316.6x10^9$

Por parte de EDA, la gráfica de pastel de la Figura 14 muestra el porcentaje de cada tipo de transacción.

Figura 14

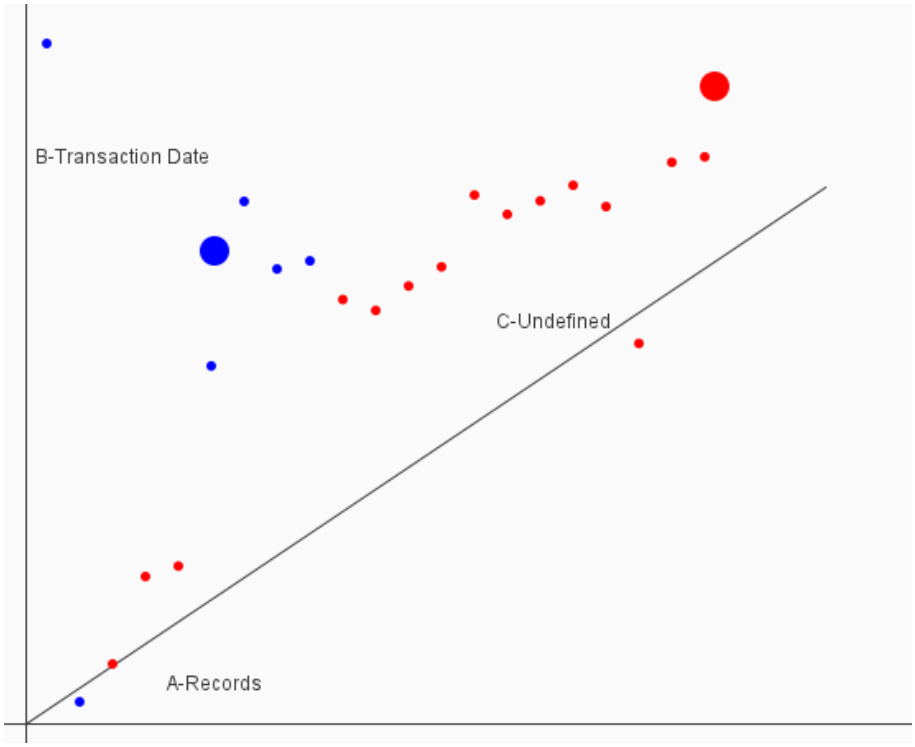
Volumen de transacciones por su tipo.



En el caso de la agrupación por K-means, la Figura 15 muestra dos tipos de agrupamiento por sus centroides por número de registros y por año.

Figura 15

Clasificación K-means para Transaction Type. Los ejes muestran para el Punto A el número de registros y para el Punto B el año, los clústers aleatorios agrupan los registros coincidentes.



6.1.2. *Transaction Shares*

El monto de acciones intercambiadas por período. Se normalizó considerando sólo los valores superiores a $400x10^6$ títulos. La Tabla 13 muestra las operaciones agrupadas por los períodos más significativos. Se debe notar la diferencia entre el comportamiento de los años 2008 y 2020, ya que en el primero se dio la caída de la bolsa a causa, entre otras cosas, de la quiebra del mercado inmobiliario y en el segundo fue el año de la pandemia de Covid-19.

Tabla 13

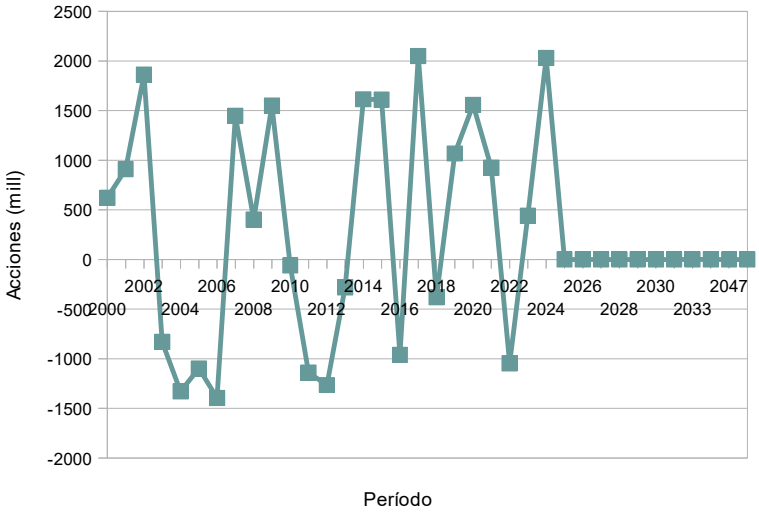
Acciones intercambiadas por año.

Período	**Acciones**	**Período**	**Acciones**
2000	$623.4x10^6$	2015	$1,611.1x10^6$
2001	$910.5x10^6$	2017	$2,049.8x10^6$
2002	$1,862.3x10^6$	2019	$1,068.7x10^6$
2007	$1,446.9x10^6$	2020	$1,558.0x10^6$
2008	$402.4x10^6$	2021	$925.2x10^6$
2009	$1,551.2x10^6$	2023	$441.0x10^6$
2014	$1,614.9x10^6$	2024	$2,030.1x10^6$

Por parte de EDA, la gráfica de la Figura 16 muestra la dispersión de los volúmenes de acciones intercambiados por cada año.

Figura 16

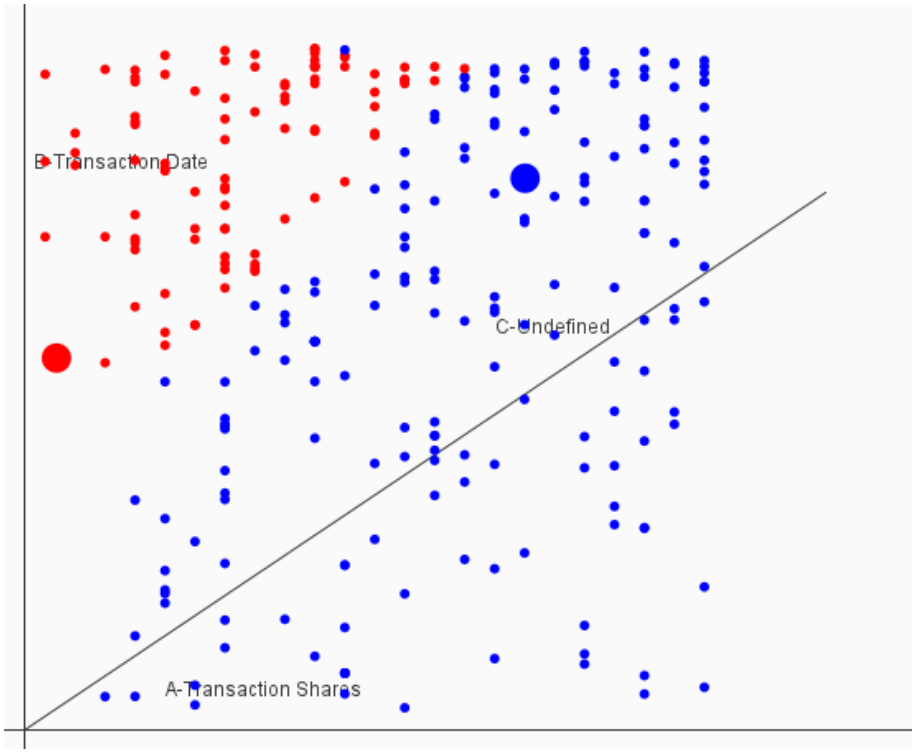
Transacciones por año.



En el caso de la agrupación por K-means, la Figura 17 muestra dos centroides agrupando uno por la coincidencia del año de cada transacción y otro por la coincidencia del monto de operaciones.

Figura 17

Clasificación K-means para Transaction Shares. Los ejes muestran para el Punto A el número de acciones comerciadas y para el Punto B el año, los clústers aleatorios agrupan los registros coincidentes.



6.1.3. *Transaction Price per Share*

Precios de la acción por código de entrega. Determina el precio de las acciones al momento de su adquisición, que se da al momento de asegurar la propiedad, y su disposición, que se da al momento de perderla ya sea por venta o caducidad. La Tabla 14 muestra que el máximo de las operaciones A y D son prácticamente idénticos dado que siempre existen alguien que compra y una contraparte que vende, exceptuando el caso de productos con caducidad o retiro del mercado.

Tabla 14

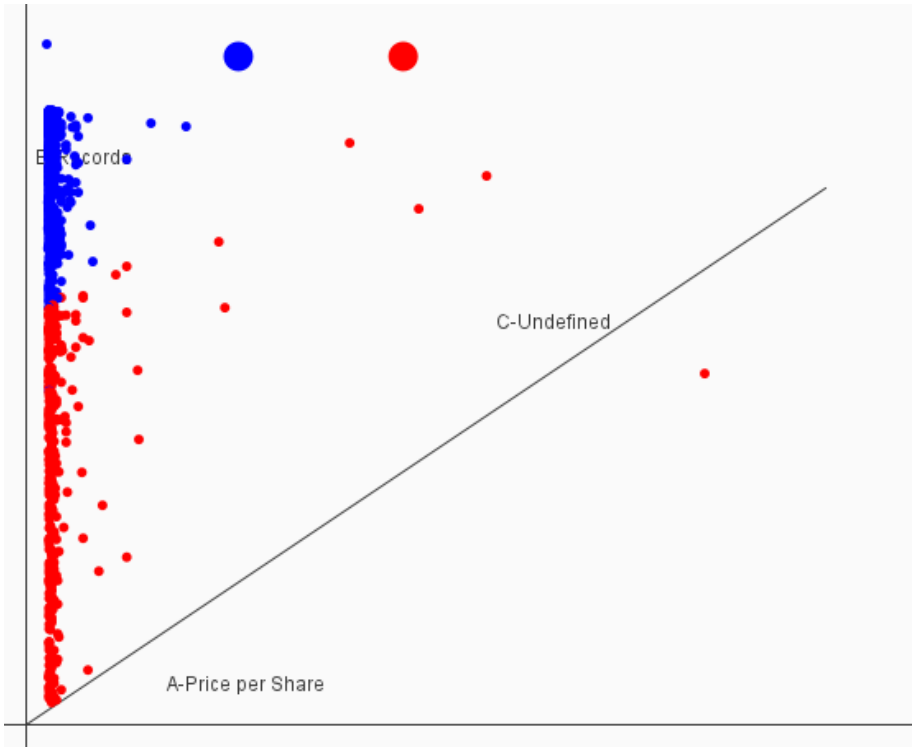
Precio por acción sobre código de adquisición (A) y disposición (D), precio por acción y valores mínimo y máximo.

Código	Precio	Min	Max
A	20.20	0.00	9,650,000.00
D	3.79	0.00	9,642,475.00

Para la agrupación por K-means, la Figura 18 fueron seleccionados de forma aleatoria: precios de alguna acción y el año. Los centroides quedan aislados, especialmente en el caso del año, por la variabilidad del precio. Es posible que una normalización más estricta habría reducido el aislamiento de ambos centroides.

Figura 18

Clasificación K-means para Price per Share. Los ejes muestran para el Punto A el precio de las acciones y para el Punto B el año, los clústers aleatorios agrupan los registros coincidentes.



6.1.4. *Exercise Date Function*

Muestra el volumen de operaciones por monto de capitalización. Por ejemplo, F2 equivale a operaciones alrededor de 2 millones de dólares y F40 a operaciones superiores a los 40 millones de dólares. Las no clasificadas (*Desconocido*) corresponden a cantidades menores a 1 millón de dólares. La Tabla 15 muestra el tipo de función y la cantidad de operaciones registradas.

Tabla 15

Registros por tipo de función.

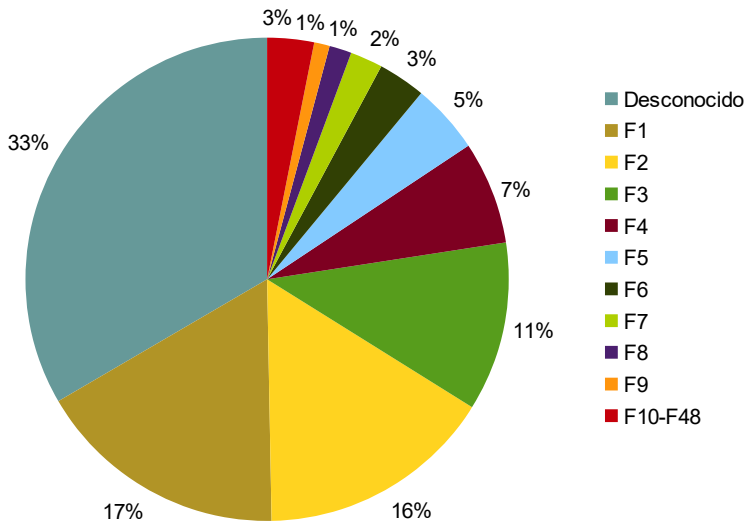
Función	**Registros**	**Función**	**Registros**
Desconocido	$3.6x10^6$	F1	$1.8x10^6$
F2	$1.7x10^6$	F3	$1.2x10^6$
F4	741,305	F5	505,092
F6	341,560	F7	235,228
F8	159,719	F9	112,887
F10	82,111	F11	63,005
F12	44,610	F13	32,688
F14	24,860	F15	19,177
F16	18,102	F17	11,120
F18	8,540	F19	6,050
F20	4,430	F21	3,448
F22	3,433	F23	2,717
F24	2,854	F25	2,669
F26	1,505	F27	1,272
F28	1,118	F29	922
F30	943	F31	405

Función	Registros	Función	Registros
F32	327	F33	300
F34	305	F35	132
F36	74	F37	76
F38	36	F39	50
F40	22	F41	12
F42	2	F43	6
F45	2	F46	2
F48	2		

Por parte de EDA, la gráfica de pastel de la Figura 19 muestra las agrupaciones de los valores de adquisición de forma porcentual.

Figura 19

Distribución por función y período.

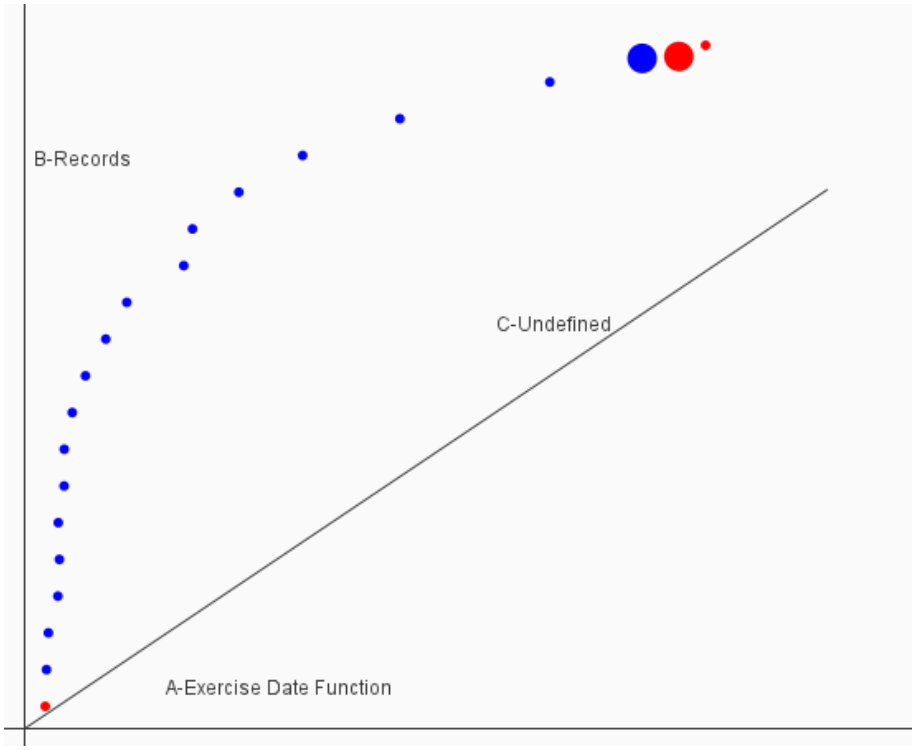


En el caso de la agrupación por K-means, la Figura 20 muestra un centroide para el volumen de capitalización y otro para la cantidad de registros contabilizados. En caso

del punto B, que tiene un valor agrupado cerca de las coordenadas $(0, 0, 0)$ (al centro de la gráfica) estaría sugiriendo similitudes entre valores con el mismo tipo de función y diferente cantidad de registros.

Figura 20

Clasificación K-means para Exercise Date Function. Los ejes muestran para el Punto A el número de la función de capitalización y para el Punto B el número de registros por cada una, los clústers aleatorios agrupan los registros coincidentes.



6.1.5. *Security Title*

Títulos, acciones y seguros subyacentes. El mercado de valores ofrece una gran variedad de productos, desde *commodities* o productos básicos, hasta servicio y productos derivados. La Tabla 16 muestra cómo se encontraron distribuidos y el número de operaciones o registros que estuvieron representados.

Tabla 16

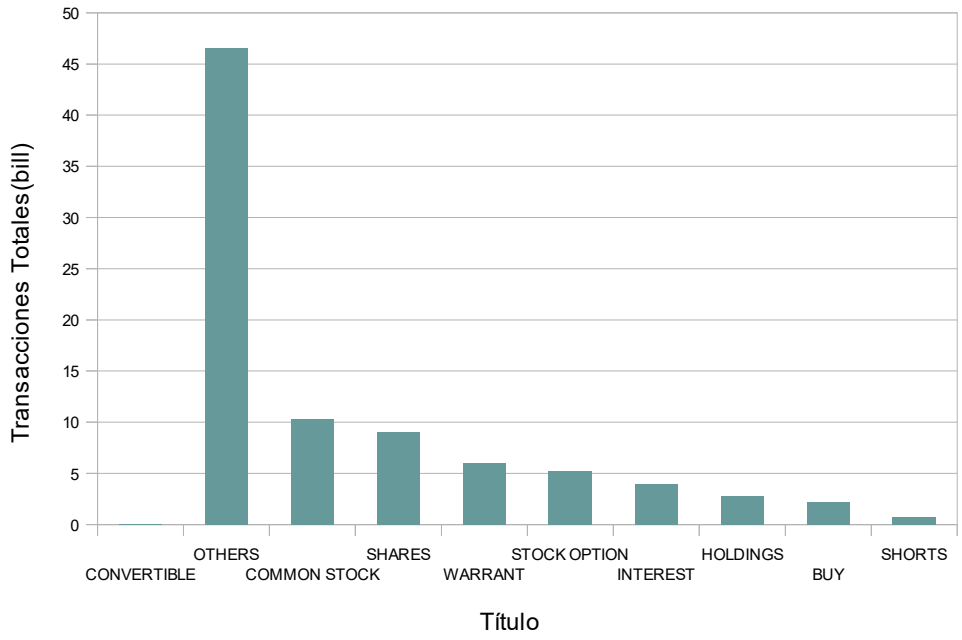
Títulos de seguros por información interna.

Type de título	**Acciones**	**Acciones de seguro**	**Titulos totales**
COMMON STOCK	$2,368.7x10^6$	$47,928.3$	$410,297.1x10^6$
WARRANT	$41,714.0x10^6$	$44,253.5x10^6$	$45,967.5x10^6$
SHARES	$42,867.3x10^6$	$46,132.5x10^6$	$48,999.9x10^6$
RENTS	16,400	16,400	32,800
INCENTIVE	$19.2x10^6$	$36.2x10^6$	$55.5x10^6$
PETROLEUM	10,000	135,000	145,000
STOCK OPTION	$306.8x10^6$	$4,900.8x10^6$	$5,207.6x10^6$
INTEREST	$1,521.9x10^6$	$2,451.1x10^6$	$3,973.1x10^6$
SALARY	$151.4x10^6$	$157.3x10^6$	$308.8x10^6$
INVESTMENT	$2.6x10^6$	$271.9x10^6$	$274.5x10^6$
CALL OPTION	$59.3x10^6$	$289.2x10^6$	$348.5x10^6$
HOLDINGS	$246.2x10^6$	$2,502.9x10^6$	$2,749.2x10^6$
PURCHASE	$126.6x10^6$	$449.9x10^6$	$576.6x10^6$
SELL	$266.0x10^6$	$468.8x10^6$	$734.8x10^6$
BUY	$1580.4x10^6$	$601.2x10^6$	$2,181.7x10^6$
ACQUIRE	$102.2x10^6$	$257.3x10^6$	$359.6x10^6$
CANCELLATION	178,750	120,000	298,750
CONVERTIBLE	$10,093.3x10^6$	$41,534.7x10^6$	$51,628.1x10^6$
RESTRICTED	$106.0x10^6$	$128.3x10^6$	$234.3x10^6$
OTHERS	$6,962.9x10^6$	$39,569.0x10^6$	$46,531.9x10^6$

Por parte de EDA, la Figura 21 muestra la tendencia descendente que el número de acciones está representando. En este caso se podrían haber normalizado los valores quitando los extremos al eliminar los menores de 10^6 (1 millón), sin embargo, como los títulos no pocos, se decidió dejarlos de forma ilustrativa.

Figura 21

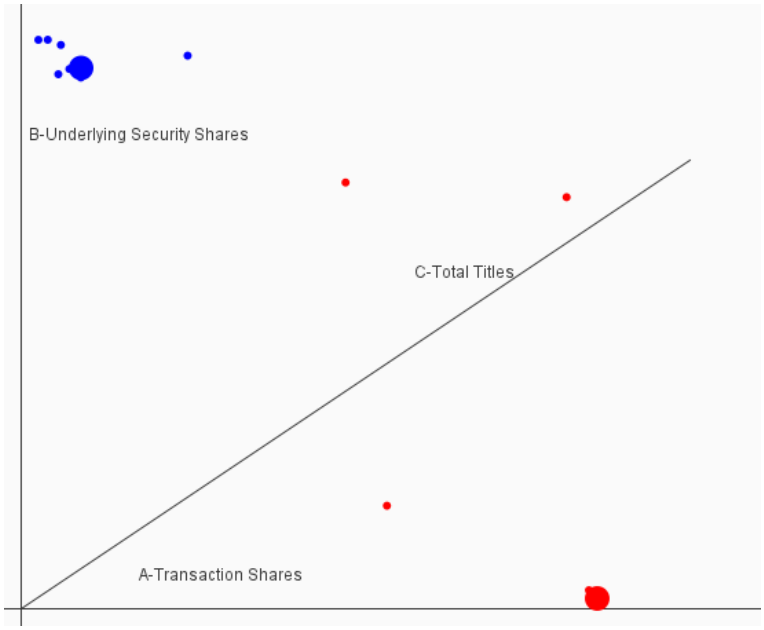
Títulos de seguros por tipo.



En el caso de la agrupación por K-means, la Figura 22 toma como centroides las acciones intercambiadas y las acciones de seguro (columnas 2 y 3 de la Tabla 16).

Figura 22

Clasificación K-mean para Security Title. Los ejes muestran para el Punto A el tipo de acción, para el Punto B el volumen de acciones intercambiadas y para el Punto B el volumen de acciones de seguridad, los clústers aleatorios agrupan los registros coincidentes.



6.1.6. *Direct or Indirect Ownership*

Propiedad directa e indirecta de acciones, muestra si las acciones estaban siendo manipuladas a través del propietario o de un tercero. La Tabla 17 muestra el volumen de acciones, los registros de las operaciones en la base de datos y sus porcentajes.

Tabla 17

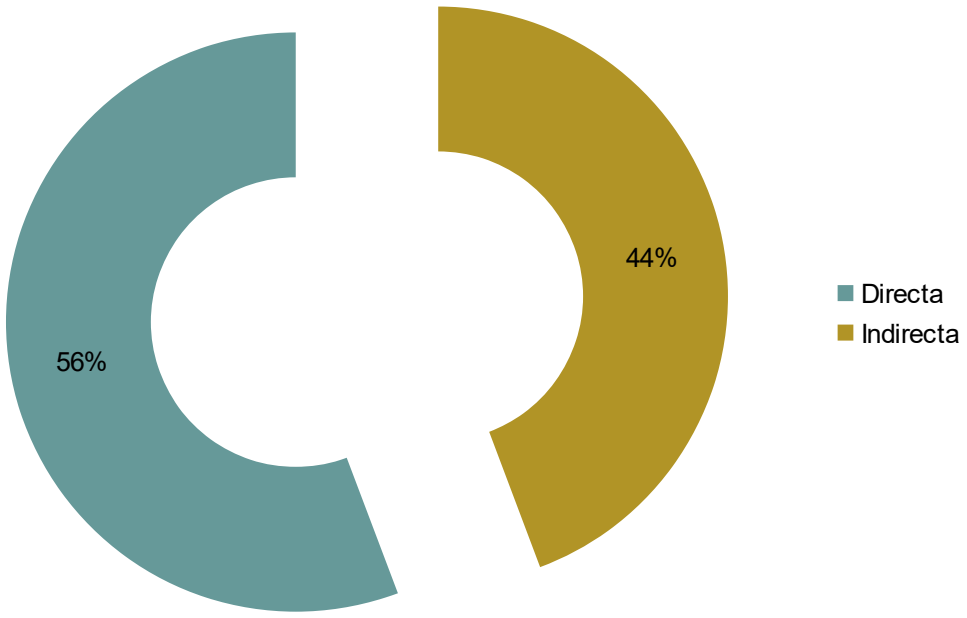
Propiedad directa o indirecta, valor absoluto de acciones y número de registros.

Propiedad	**Acciones**	**%**	**Registros**	**%**
Directa	$1,693.2x10^6$	55.73 %	$9.4x10^6$	89.12 %
Indirecta	$1,345.0x10^6$	44.27 %	$1.1x10^6$	10.88 %

Por parte de EDA, la gráfica de pastel de la Figura 23 muestra el porcentaje de acciones de cada tipo de propiedad, que difiere de la suma total de registros de las operaciones.

Figura 23

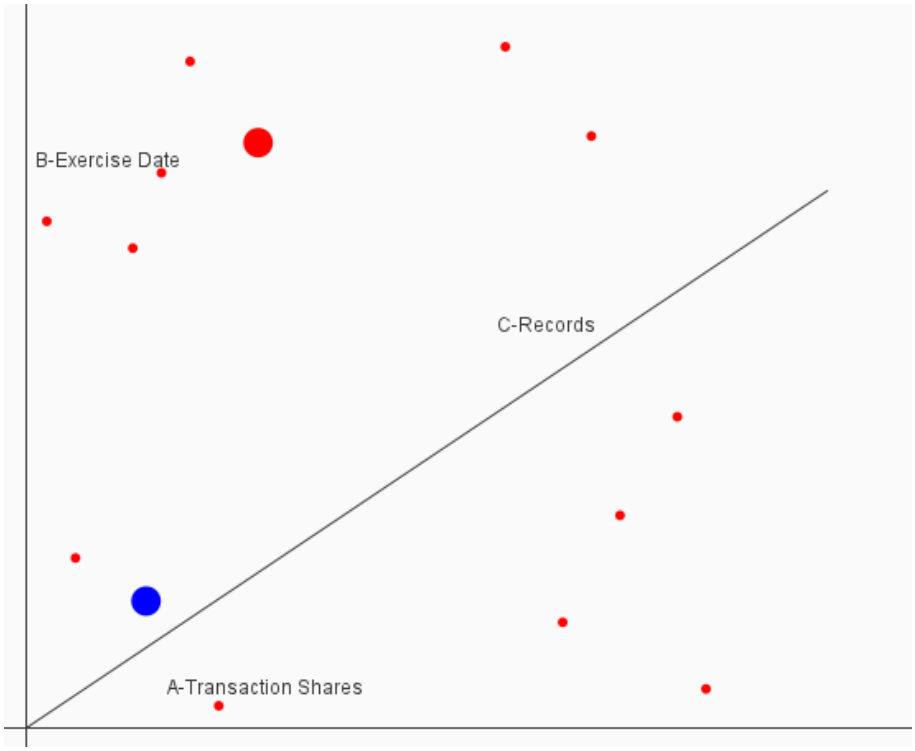
Propiedad Directa o Indirecta.



A diferencia de las definiciones anteriores de K-means, que sólo utilizaron 2 parámetros, la Figura 24 muestra para este caso 3 valores parametrizados de los centroides: para el Punto A el volumen de acciones intercambiadas, para el Punto B el año de la transacción y para el Punto C la cantidad de registros.

Figura 24

Clasificación K-means para Direct or Indirect Ownership. Los ejes muestran para el Punto A el volumen de acciones intercambiadas, para el Punto B el año de la transacción y para el Punto C el número de registros por año, los clústers aleatorios agrupan los registros coincidentes.



6.1.7. *Equity Swap Involved*

Futuros y códigos de transacción. Los códigos A, B, C, etc. se refieren a los flujos de efectivo al momento del intercambio. Su valor se calcula comparando los valores presentes y futuros de la transacción. La Tabla 18 muestra el número de *Equity Swap*, ES, por tipo y la cantidad de registros representados en la muestra.

Tabla 18

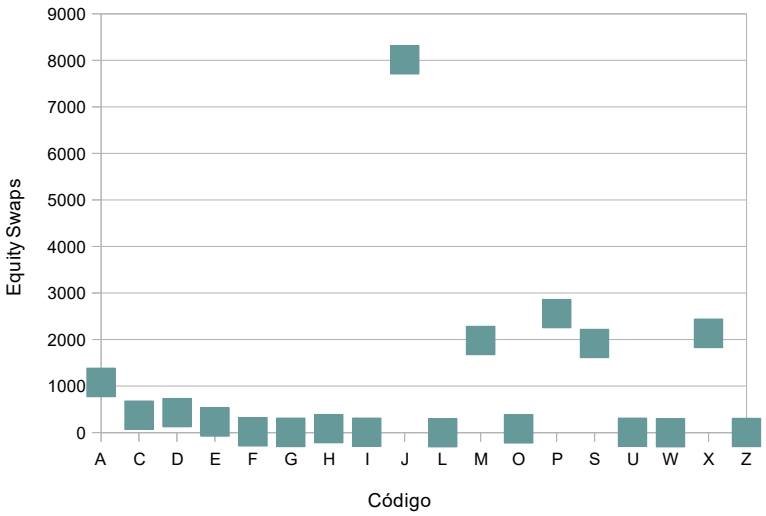
Equity Swaps (ES) que intervienen por código de transacción.

Código	**ES**	**%**	**Registros**	**%**
A	1,079	5.71 %	2,817,928	26.47 %
C	373	1.97 %	294,223	2.76 %
D	428	2.26 %	370,589	3.48 %
E	230	1.22 %	5,367	0.05 %
F	18	0.10 %	14,206	0.13 %
G	4	0.02 %	28,833	0.27 %
H	84	0.44 %	5,682	0.05 %
I	8	0.04 %	10,377	0.10 %
J	8,014	42.37 %	266,805	2.51 %
L	0	0.00 %	1,984	0.02 %
M	1,979	10.46 %	2,178,473	20.47 %
O	80	0.42 %	1,550	0.01 %
P	2,558	13.53 %	108,885	1.02 %
S	1,915	10.13 %	51,389	0.48 %
U	8	0.04 %	8,368	0.08 %
W	0	0.00 %	807	0.01 %
X	2,133	11.28 %	113,378	1.07 %
Z	2	0.01 %	298	0.00 %
No definido	0	0.00 %	4,364,723	41.01 %

Por parte de EDA, la gráfica de la Figura 25 muestra la dispersión del número total de *swaps* con respecto a su código de transacción.

Figura 25

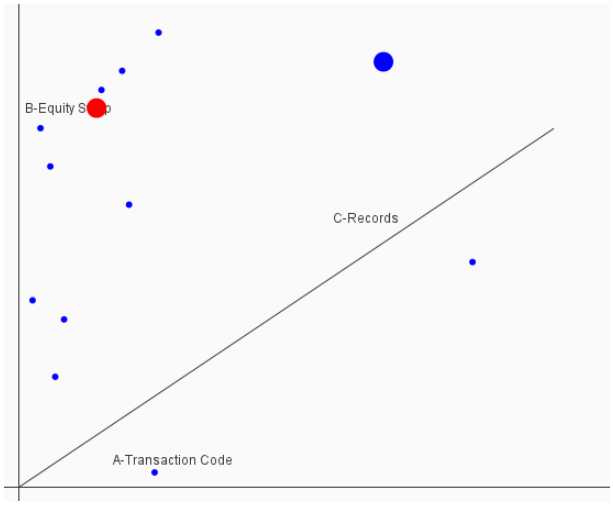
Equity Swaps por código de transacción.



En el caso de la agrupación por K-means, en este caso también se utilizaron 3 parámetros, la Figura 26 muestra para A las agrupaciones por código de transacción, para B el número de *swaps* y para C el número de registros por código.

Figura 26

Clasificación K-means para Equity Swaps. Los ejes muestran para el Punto A el código de transacción, para el Punto B el número de swaps envueltos en la operación y para el Punto C el número de registros por código, los clústers aleatorios agrupan los registros coincidentes.



En todos los casos la selección de los centroides, que agrupan en torno suyo a todos los valores coincidentes, se hizo de manera aleatoria. Para estudios específicos de laboratorio se pueden pre-seleccionar centroides con características deseadas, como sería el caso de, por ejemplo, querer saber el comportamiento de las ventas de acciones previos al cierre por día festivo de Semana Santa.

6.2. Discusión

Con respecto a la obtención de los objetivos propuestos:

La arquitectura creada desde lenguaje Java permitió la implementación del dataverso en una instancia de base de datos, su análisis DD, el EDA, la clasificación por K-means y la posterior predicción a través del algoritmo de MLP. A partir de los teoremas matemáticos deducidos durante la investigación, el código se caracterizó por transformar estructuras de tipo primitivo en objetos y clases, desarrollando una interpretación nativa, esto es, desde código fuente, de los teoremas propuestos la resolución de las operaciones y cálculos que dieron por resultado las tablas, gráficas y ploteos de clasificación.

El esquema de parámetros se construyó a partir de los campos del dataverso y correspondió a las hipótesis del comportamiento del precio que dieron por resultado tablas, gráficas y ploteos posteriores. Dada la robustez del dataverso, éste no sólo se pudo dividir en 7 bases de datos sino que éstas mismas tuvieron que ser normalizadas de acuerdo a sus similitudes.

La extensión del dataverso hizo innecesario utilizar valores aleatorios. Se pudo observar una gran adaptabilidad de la estructura de datos con respecto a la creación de nuevas tendencias en el análisis de datos al haber dejado abiertas las posibilidades de nuevas hipótesis soportadas por la metodología de análisis de datos de DD, EDA, k-means y MLP.

Las diferentes tablas y gráficas obtenidas demostraron que las particularidades de cada tendencia informativa obtenida a través de las hipótesis matemáticas. La arquitectura creada redujo el problema del overfitting a través de una aplicación previa del análisis DD y posterior del EDA, con lo que todos los elementos fueron separados acorde a las solicitudes de las hipótesis matemáticas, compartiendo solamente características indexadas evitando de tal suerte la repetición o sobre-análisis que son comunes al overfitting. En el caso del MLP, el uso del algoritmo de K-means permitió la clasificación precisa de los datos a ser entrenados, llevando con esta a una exclusión del overfitting al transmitir valores precisos y no difusos a la etapa del entrenamiento de datos. El entrenamiento programado de esta forma redujo también las circunstancias donde podría darse un problema de overfitting posterior.

La investigación ha logrado aportar conocimiento a diferentes áreas de la ciencia de datos. La Tabla 19 enumera las principales contribuciones comparadas con respecto a investigaciones similares.

En el futuro, las investigaciones podrán, a partir de los fundamentos aquí planteados, implementar resultados comparativos a través de algoritmos de clasificación tales como KNeighborsClassifier, NearestCentroid o NearestNeighbors, dependiendo el perfil de la hipótesis que se haya propuesto.

Tabla 19

Tabla comparativa de las contribuciones.

Este trabajo	**Otros trabajos**	**Contribuciones: "En esta investigación..."**
Matemáticas: Teorema Coordenadas Parametrizadas	Módulo para mejorar el aprendizaje matemático a través de coordenadas cartesianas. (Laia, 2023)	Las coordenadas y sus parámetros forman en el lenguaje una sola clase
Matemáticas: Teorema Distancia Euclideana (DE) dinámica	Establecimiento de un núcleo central para análisis de poblaciones a través de ED (Weidong y Wanlu, 2022)	La DE incluiría metadatos de cada individuo
Matemáticas: Teorema Centroide ITC	Diagramas interactivos inducidos por centroides móviles (Chiorean, 2024)	Los parámetros del centroide incluiría el movimiento como valor
Ciencia de Datos: EDA a través de información de *Insider Trading*	EDA a través de sicología de toma de decisiones(Kruszewski y Michalak, 2024)	La precisión del EDA se forma a partir de la robustez del dataset y no de propuestas subjetivas
Machine Learning: Cálculo de k-means con ITC integrando la lista de parámetros	Reducción de la pérdida con respecto a los parámetros de cada vista, presenta un k-means más amigable (Liu et al., 2022)	Las vistas y sus parámetros forman parte del centroide y por lo tanto de k-means

7. Conclusiones

Las tendencias en el mercado financiero están relacionadas incidentalmente al comportamiento de las acciones, cuyos precios fluctúan dependiendo al tipo, duración y fortaleza de la información contenida en sus estructuras de datos. Los planteamientos originales que ofrece el análisis DD producen frecuentemente un grado amplio de overfitting el cual se reduce al hacer uso de EDA y se minimiza cuando se aplican modelos de clasificación K-means, para después ser entrenados por medio del MLP. Como resultado, se puede observar que un único dataset con suficientes tipos de campos, de tipologías diversas, y suficiente profundidad, por la cantidad de registros (llamado este fenómeno *robustez*), permite normalizar y racionalizar los datos de tal suerte que la experimentación permite observar resultados a partir de diferente enfoques y perfiles. El precio como factor fundamental se ve potenciado por el tipo de transacción, su manipulación, su temporalidad y su disponibilidad.

El hecho de aplicar la metodología aquí propuesta: análisis DD + EDA + K-means + MLP reduce el overfitting y apunta hacia una economización de los procesos y con ello a un mejor control de la naturaleza del precio que desde, la perspectiva del comprador, se convierte en el costo de la inversión proyectada.

8. Referencias

.

Abirami, S. y Chitra, P. (2020). Chapter Fourteen - Energy-efficient edge based real-time healthcare support system *Computers* https://doi.org/10.1016/bs.adcom.2019.09.007

Achar, S. (2019). Early Consequences Regarding the Impact of Artificial Intelligence on International Trade *American Journal of Trade and Policy* https://ideas.repec.org/a/ris/ajotap/0133.html

Adler, Y., Farchi, E., Klausner, M., Pelleg, D., Raz, O. y Shochat, M. (2009). Automated substring hole analysis *31st International Conference on Software Engineering* https://doi.org/10.1109/ICSE-COMPANION.2009.5070982

Afzal, S., Ghani, S., Jenkins-Smith, H.C., Ebert, D.S., Hadwiger, M. y Hoteit, I. (2020). A Visual Analytics Based Decision Making Environment for COVID-19 Modeling and Visualization *IEEE Visualization Conference* https://doi.org/10.1109/VIS47514.2020.00024

Agrawal, K., Mehta, V., Renganathan, S., Acharyya, S., Padmanabhan, V.N. y Kotipalli, C. (2021). Monitoring Cloud Service Unreachability at Scale *IEEE Conference on Computer Communications* https://doi.org/10.1109/INFOCOM42981.2021.9488778

Akbary, P., Ghiasi, M., Pourkheranjani, M. R. R., Alipour, H. y Ghadimi, N. (2017). Extracting Appropriate Nodal Marginal Prices for All Types of Committed Reserve *Computational Economics* https://doi.org/10.1007/s10614-017-9716-2

Albayati, H., Kim, S. K. y Rho, J. J. (2020). Accepting financial transactions using blockchain technology and cryptocurrency: A customer perspective approach *Technology in Society* https://doi.org/10.1016/j.techsoc.2020.101320

AlJanabi, K. y Kadim, R. (2017). A Hybrid Data Warehouse Model to Improve Mining Algorithms *Journal of Kufa for Mathematics and Computer* https://journal.uokufa.edu.iq/index.php/jkmc/article/view/2099/1952

Allwin, M., Tanoto, A., Forbes, W. y Ashari (2019). Forecasting Determination by Ousing

Development Schedule Using Learning Machine Approach Using Clustering Method *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1230/1/012019

Amigo, J. M. (2021). Data Mining, Machine Learning, Deep Learning, Chemometrics *Brazilian Journal of Analytical Chemistry* https://doi.org/10.30744/brjac.2179-3425.ar-38-2021

Angryk, R.A. y Petry, F.E. (2005). Mining Multi-Level Associations with Fuzzy Hierarchies *The 14th IEEE International Conference on Fuzzy Systems* https://doi.org/10.1109/FUZZY.2005.1452494

Annanth, K., Abinash, M. y Rao, L. (2021). Intelligent manufacturing in the context of industry 4.0: A case study of siemens industry *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1969/1/012019

Arvanitidis, A. I., Bargiotas, D., Daskalopulu, A., Kontogiannis, D., Panapakidis, I. P. y Tsoukalas, L. H. (2022). Clustering Informed MLP Models for Fast and Accurate Short-Term Load Forecasting *Energies* https://doi.org/10.3390/en15041295

Ashraf, B. N. (2020). Economic impact of government interventions during the COVID-19 pandemic: International evidence from financial markets *Journal of Behavioral and Experimental Finance* https://doi.org/10.1016/j.jbef.2020.100371

Attaran, M., Stark, J. y Stotler, D. (2018). Opportunities and challenges for big data analytics in US higher education: A conceptual model for implementation *Industry and Higher Education* https://doi.org/10.1177/0950422218770937

Augustine, V., Hudepohl, J., Marcinczak, P. y Snipes, W. (2018). Deploying Software Team Analytics in a Multinational Organization *IEEE Software* https://doi.org/10.1109/MS.2017.4541044

Bai, X. y Bi, Y. (2018). Derivative Entropy-Based Contrast Measure for Infrared Small-Target Detection *IEEE Transactions on Geoscience and Remote Sensing* https://doi.org/10.1109/TGRS.2017.2781143

Balogh, A. (2025). Layline insider trading dataset https://doi.org/10.7910/DVN/VH6GVH

Balogh, A. (2023). Insider trading *Scientific Data* https://doi.org/10.1038/s41597-023-

02147-6

Baltzer, O., Dehne, F. y Rau-Chaplin, A. (2013). OLAP for moving object data *International Journal of Intelligent Information and Database Systems* https://doi.org/10.1504/IJIIDS.2013.051745

Baresi, L. y Guinea, S. (2013). Event-Based Multi-level Service Monitoring *IEEE 20th International Conference on Web Services* https://doi.org/10.1109/ICWS.2013.21

Barral, M. (2018). Herman Hollerith el ordenador del censo https://www.heraldo.es/noticias/sociedad/2018/12/09/herman-hollerith-ordenador-del-censo-1281679-310.html

Barthe, G., Katoen, J.-P. y Silva, A. (2020). *Foundations of probabilistic programming* Cambridge University Press. https://doi.org/10.1017/9781108770750

Basat, R.B., Shahout, R. y Friedman, R. (2018). Frequent elements on query defined ranges *IEEE Conference on Computer Communications Workshops* https://doi.org/10.1109/INFCOMW.2018.8406919

Bianchi, R.G., Hatano, G.Y. y Lopes, T.L. (2013). On the performance and use of spatial OLAP tools *XXXIX Latin American Computing Conference* https://doi.org/10.1109/CLEI.2013.6670652

Brasil, M., Serique, B. y Magalhães, J. (2016). Temporal Data Visualization Technique Based on Treemap *20th International Conference Information Visualisation (IV)* https://doi.org/10.1109/IV.2016.65

Brastama, A., Mukaromah, S., Agussalim, Lusiarini, Y., Ibnu, M. y Yunifa, P. (2019). Design and Development Executive Information System Application with Drilldown and What-If Analysis features *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1569/2/022050

Buaton, R., Mawengkang, H., Zarlis, M., Effendi, S., Manaor, A., Maulita, Y., Fauzi, A., Novriyenni, N., Sihombing, A. y Lumbanbatu, K. (2019). Decision Tree Optimization in Data Mining with Support and Confidence *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1255/1/012056

Buccafurri, F. y Lax, G. (2010). Approximating sliding windows by cyclic tree-like histograms for efficient range queriesXXX ERROR UNICODE X2606XX *Data & Knowledge Engineering*

https://doi.org/10.1016/j.datak.2010.05.002

Bukhari, A. H., Raja, M. A. Z., Sulaiman, M., Islam, S., Shoaib, M. y Kumam, P. (2020). Fractional Neuro-Sequential ARFIMA-LSTM for Financial Market Forecasting *IEEE Access* https://doi.org/10.1109/ACCESS.2020.2985763

Campos, O. (2011). Alan Turing, padre de la informática moderna y paria social https://www.genbeta.com/desarrollo/alan-turing-padre-de-la-informatica-moderna-y-paria-social

Cañas, P. (2000). Aspectos Jurídicos del Censo Romano *Boletín de la Facultad de Derecho UNED* https://dialnet.unirioja.es/servlet/tesis?codigo=40384

Cariou, V., Cubille, J., Derquenne, C., Goutier, S., Guisnel, F. y Klajnmic, H. (2009). Embedded indicators to facilitate the exploration of a data cube *International Journal of Business Intelligence and Data Mining* https://doi.org/10.1504/IJBIDM.2009.029083

Carnagie Mellon University (Ed.) (2022). Artificial Intelligence Engineering https://www.sei.cmu.edu/our-work/artificial-intelligence-engineering/

Castro, H. (2010). Primer Censo de la Nueva España 1790

Chang, C., Chen, R. y Zhuo, Y. (2005). The case study for building a data warehouse in semiconductor manufacturing *International Journal of Computer Applications in Technology* https://doi.org/10.1504/IJCAT.2005.008265

Chassagnon, G., Vakalopolou, M., Paragios, N. y Revel, M.-P. (2019). Deep learning: definition and perspectives for thoracic imaging *European Radiology* https://doi.org/10.1007/s00330-019-06564-3

Chavalier, M., El Malki, M., Kopliku, A., Teste, O. y Tournier, R. (2016). Document-oriented data warehouses: Models and extended cuboids, extended cuboids in oriented document *IEEE Tenth International Conference on Research Challenges in Information Science* https://doi.org/10.1109/RCIS.2016.7549351

Chen, J., Long, T. y Deng, K. (2008). The Consistency of Materialized View Maintenance and Drill-Down in a Warehousing Environment *The 9th International Conference for Young Computer Scientists* https://doi.org/10.1109/ICYCS.2008.212

Chen, Q., Li, Z., Pong, T.C. y Qu, H. (2019). Designing Narrative Slideshows for Learning

Analytics *IEEE Pacific Visualization Symposium* https://doi.org/10.1109/PacificVis.2019.00036

Chen, Y., Yang, B. y Wang, W. (2017). NetFlowMatrix: a visual approach for analysing large NetFlow data *International Journal of Security and Networks* https://doi.org/10.1504/IJSN.2017.088115

Chiorean, C. G. (2024). Computational issues in biaxial bending capacity assessment of RC and composite cross-sections exposed to fire *Computers & Structures* https://doi.org/10.1016/j.compstruc.2024.107477

Cho, S. J., Chung, C. Y. y Young, J. (2019). Study on the Relationship between CSR and Financial Performance *Sustainability* https://doi.org/10.3390/su11020343

Conklin, N., Prabhakar, S. y North, C. (2002). Multiple foci drill-down through tuple and attribute aggregation polyarchies in tabular data *IEEE Symposium on Information Visualization* https://doi.org/10.1109/INFVIS.2002.1173158

Connor, J. y Robertson F. (2001). Max August Zorn. School of Mathematics and Statistics https://mathshistory.st-andrews.ac.uk/Biographies/Zorn/

Culkin, R. y Sanjiv, R. (2017). Machine learning in finance: the case of deep learning for option pricing *Journal of Investment Management*

De Haan, J., Schoenmaker, D. y Wierts, P. (2020). *Financial markets and institutions: A European Perspective* Cambridge University Press.

De Melo, T., Rocha, R., Garcia, T. y De Castro, J. (2021). Spatial data cubes based on shared dimensions and neighbourhood relationship concepts *International Journal of Business Information Systems* https://doi.org/10.1504/IJBIS.2021.116084

Deakin, R. E., Bird, S. C. y Grenfell, R. I. (2002). The Centroid? Where would you like it to be be? *Cartography* https://doi.org/10.1080/00690805.2002.9714213

Dhyani, B., Kumar, M., Verma, P. y Jain, A. (2020). Stock Market Forecasting Technique using Arima Model *International Journal of Recent Technology and Engineering (IJRTE)* https://doi.org/10.35940/ijrte.f8405.038620

Dike, H. U., Zhou, Y., Deveerasetty, K. K. y Wu, Q. (2018). Unsupervised Learning Based On Artificial Neural Network: A Review https://doi.org/10.1109/CBS.2018.8612259

Ecured (Ed.) (2011). Censo de Cuba https://www.ecured.cu/Censo

Egeland, R., Wildish, T. y Huang, C.H. (2010). PhEDEx Data Service *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/219/6/062010

El Naqa, I. y Murphy, M. J. (2015). What Is Machine Learning? *Machine Learning in Radiation Oncology* https://doi.org/10.1007/978-3-319-18305-3_1

Enlyft (Ed.) (2019). Companies using Oracle ERP. USA https://enlyft.com/tech/products/oracle-erp

Euston (Ed.) (2019). Máquina Tabuladora. U.S.A https://www.euston96.com/maquina-tabuladora/

Fairbairn, A. (2005). A history of agricultural production at Neolithic Çatalhöyük East Turkey *World Archaeology* https://doi.org/10.1080/00438240500094762

Fang, M. (2006). A thermal centroid or a mass centroid? *Journal of Oceanography* https://doi.org/10.1007/s10872-006-0093-z

Feng, Y., Agrawal, D., El Abbadi, A. y Metwally, A. (2004). Range cube: efficient cube computation by exploiting data correlation *20th International Conference on Data Engineering* https://doi.org/10.1109/ICDE.2004.1320035

Fine, S., Singer, Y. y Tishby, N. (1998). The Hierarchical Hidden Markov Model: Analysis and Applications *Machine Learning* https://doi.org/10.1023/A:1007469218079

Fortier, P. y Michel, H. (2003). Computer Systems Performance Evaluation and Prediction *Elsevier*

Franciscus, N., Ren, X. y Stantic, B. (2018). Precomputing architecture for flexible and efficient big data analytics *Vietnam Journal of Computer Science* https://doi.org/10.1007/s40595-018-0109-9

Franklin, P. (2021). Solving Problems with Rapid Data Discovery *Annual Reliability and Maintainability Symposium* https://doi.org/10.1109/RAMS48097.2021.9605783

Friedman, M. (1963). *Price Theory* Routledge. https://doi.org/10.4324/9781315127378

Fung, C.C. y Thanadechteemapat, W. (2010). Discover Information and Knowledge from Websites Using an Integrated Summarization and Visualization Framework *Third International Conference on Knowledge Discovery and Data Mining* https://doi.org/10.1109/WKDD.2010.109

Gadicha, A., Gadicha, V. y Obaid, A. (2021). A Novel approach towards Implicit Authentication System by using Multi-share visual key Cryptography Mechanism *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1963/1/012141

Gartner (Ed.) (2022). Hyperautomation https://www.gartner.com/en/information-technology/glossary/hyperautomation

Genevay, A., Dulac-Arnold, G. y Vert, J.-P. (2019). Differentiable Deep Clustering with Cluster Size Constraints *CoRR* https://doi.org/10.48550/ARXIV.1910.09036

Geymayer, T., Lex, A., Streit, M. y Schmalstieg, D. (2011). Visualizing the Effects of Logically Combined Filters *15th International Conference on Information Visualisation* https://doi.org/10.1109/IV.2011.52

Glover, F. (2016). Pseudo-centroid clustering *Soft Computing* https://doi.org/10.1007/s00500-016-2369-6

Gonzalez, L., Powell, J.G., Shi, J. y Wilson, A. (2005). Two centuries of bull and bear market cycles *International Review of Economics & Finance* https://doi.org/10.1016/j.iref.2004.02.003

Grabot, B. (2020). Rule mining in maintenance: Analysing large knowledge bases *Computers & Industrial Engineering* https://doi.org/10.1016/j.cie.2018.11.011

Greig, J. (2021). Generative AI autonomic systems hyperautomation and more top Gartner list of top tech trends in 2022 https://www.zdnet.com/article/generative-ai-autonomic-systems-hyperautomation-and-more-top-gartner-list-of-top-tech-trends-in-2022/

Gu, S., Kelly, B. y Xiu, D. (2020). Empirical Asset Pricing via Machine Learning *The Review of Financial Studies* https://doi.org/10.1093/rfs/hhaa009

Gutiérrez, P. (2017). Tipos de criptografía: simétrica asimétrica e híbrida. https://www.genbeta.com/desarrollo/tipos-de-criptografia-simetrica-asimetrica-e-hibrida

Hanna, A. J., Turner, J. D. y Walker, C. B. (2020). News media and investor sentiment during bull and bear markets *The European Journal of Finance* https://doi.org/10.1080/1351847X.2020.1743734

Hartono, W.S. y Widyantoro, D.H. (2016). Fisheye zoom and semantic zoom on citation

network visualization *2016 International Conference on Data and Software Engineering (ICoDSE)* https://doi.org/10.1109/ICODSE.2016.7936109

Hassan, O. A. G. y Marston, C. (2019). Corporate Financial Disclosure Measurement in the Empirical Accounting Literature: A Review Article *The International Journal of Accounting* https://doi.org/10.1142/s1094406019500069

Hawkins, D. (2003). The Problem of Overfitting *Journal of Chemical Inofrmation and Computer Sciences* https://doi.org/10.1021/ci0342472

He, C., Micallef, L., He, L., Peddinti, G., Aittokallio, T. y Jacucci, G. (2021). Characterizing the Quality of Insight by Interactions: A Case Study *IEEE Transactions on Visualization and Computer Graphics* https://doi.org/10.1109/TVCG.2020.2977634

Heaton J. (2015). Artificial Intelligence for Humans Volume 3: Deep Learning and Neural Networks

Herbert, Putro, B., Putra, R. y Fitriasari, N. (2019). Learning Management System (LMS) model based on machine learning supports 21st century learning as the implementation of curriculum 2013 *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1280/3/032032

Hong, Y. (2020). Analysis of Database Programming Technology in Computer Software Engineering *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1651/1/012070

Hosch, W. (2019). Edgar Frank Codd https://www.britannica.com/biography/Edgar-Frank-Codd

Hussain, S.J., Hernandez, J.A., Rehman, M.U., Al-Yahyaee, K.H. y Zakaria, M. (2018). A global network topology of stock markets: Transmitters and receivers of spillover effects *Physica A: Statistical Mechanics and its Applications* https://doi.org/10.1016/j.physa.2017.11.132

Huy, D. T. N., Nhan, V. K., Bich, N. T. N., Hong, N. T. P., Chung, N. T. y Huy, P. Q. (2020). Impacts of Internal and External Macroeconomic Factors on Firm Stock Price in an Expansion Econometric model-A Case in Vietnam Real Estate Industry *Data Science for Financial Econometrics* https://doi.org/10.1007/978-3-030-48853-6_14

IBM (Ed.) (2022). What is a data fabric? https://www.ibm.com/topics/data-fabric

IBM (Ed.) (2001). Sabre. The First Online Reservation System https://www.ibm.com/history/sabre

IBM (Ed.) (2022). Drilling up and drilling down. IBM Planning Analytics with Watson https://www.ibm.com/docs/en/planning-analytics/2.0.0?topic=data-drilling-up-drilling-down

Ikeda, R., Cho, J., Fang, C., Salihoglu, S., Torikai, S. y Widom, J. (2012). Provenance-Based Debugging and Drill-Down in Data-Oriented Workflows *IEEE 28th International Conference on Data Engineering* https://doi.org/10.1109/ICDE.2012.118

Ilyas, Q.M., Ahmad, M., Zaman, N., Alshamari, M.A. y Ahmed, I. (2022). Localized Text-Free User Interfaces *IEEE Access* https://doi.org/10.1109/ACCESS.2021.3139525

Imran, S., Mahmood, T., Morshed, A. y Sellis, T. (2021). Big data analytics in healthcare XXX ERROR UNICODE X2212XX A systematic literature review and roadmap for practical implementation *IEEE/CAA Journal of Automatica Sinica* https://doi.org/10.1109/JAS.2020.1003384

INE (Ed.) (2013). Instituto Nacional de Estadística https://www.ine.es/explica/docs/historia_estadistica.pdf

Inmon, W. (2005). Building the Data Warehouse https://books.google.com.mx/books/about/Building_the_Data_Warehouse.html?id=duRQAAAAMAAJ&redir_esc=y

Intellectual Point (Ed.) (2022). What is Cybersecurity Mesh? https://intellectualpoint.com/what-is-cybersecurity-mesh/

Istiake Sunny, M.A., Maswood, M.M.S. y Alharbi, A.G. (2020). Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model https://doi.org/10.1109/NILES50944.2020.9257950

James, G., Witten, D., Hastie, T., Tibshirani, R. y Taylor, J. (2023). Unsupervised Learning *Springer Texts in Statistics* https://doi.org/10.1007/978-3-031-38747-0_12

Jiao, S., Song, J. y Liu, B. (2020). A Review of Decision Tree Classification Algorithms for Continuous Variables *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1651/1/012083

Jin, W. (2020). Research on Machine Learning and Its Algorithms and Development *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1544/1/012003

Jin, Z., Yang, Y. y Liu, Y. (2019). Stock closing price prediction based on sentiment analysis and LSTM *Neural Computing and Applications* https://doi.org/10.1007/s00521-019-04504-2

Joglekar, M., Garcia-Molina, H. y Parameswaran, A. (2019). Interactive Data Exploration with Smart Drill-Down *IEEE Transactions on Knowledge and Data Engineering* https://doi.org/10.1109/TKDE.2017.2685998

Johansson, J., Treloar, R. y Jern, M. (2004). Integration of unsupervised clustering, interaction and parallel coordinates for the exploration of large multivariate data *Eighth International Conference on Information Visualisation* https://doi.org/10.1109/IV.2004.1320124

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms* Wiley-IEEE Press.

Keynes, J.M. (1930). *The Pure Theory of Mone* A Treastise on Money. Cambridge University Press. http://tankona.free.fr/keynescw5.pdf

Khosravi, H., Shabaninejad, S., Bakharia, A., Sadiq, S., Indulska, M., y Gašević, D. (2021). Intelligent Learning Analytics Dashboards: Automated Drill-Down Recommendations to Support Teacher Data Exploration *Journal of Learning Analytics* https://doi.org/10.18608/jla.2021.7279

Kim, J.H., Yoon, S.H. y Kim, M.S. (2012). Study on traffic classification taxonomy for multilateral and hierarchical traffic classification *14th Asia-Pacific Network Operations and Management Symposium (APNOMS)* https://doi.org/10.1109/APNOMS.2012.6356105

Klimentov, A., Nevski, P., Potekhin, M. y Wenaus, T. (2011). The ATLAS PanDA Monitoring System and its Evolution *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/331/7/072058

Kole, E. y Van Dijk, D. (2016). How to Identify and Forecast Bull and Bear Markets? *Applied Econometrics* https://doi.org/10.1002/jae.2511

Kotamsetty, R. y Govindarasu, M. (2016). Adaptive Latency-Aware Query Processing on Encrypted Data for the Internet of Things *25th International Conference on Computer*

Communication and Networks (ICCCN) https://doi.org/10.1109/ICCCN.2016.7568488

Kritzinger, L.M., Krismayer, T., Vierhauser, M., Rabiser, R. y Grünbacher, P. (2017). Visualization support for requirements monitoring in systems of systems *32nd IEEE/ACM International Conference on Automated Software Engineering* https://doi.org/10.1109/ASE.2017.8115700

Kruszewski, T. y Michalak, J. (2024). Emotional Markers As Indicators of Investor Attitudes: EDA Sub-process Proposal *Studies in Classification* https://doi.org/10.1007/978-3-031-55917-4_22

Kumar, S. y Ahmed, U. (2020). *Hands-On Exploratory Data Analysis with Python* Packt Publishing.

Kusi-Sarpong, S., Orji, I., Gupta, H. y Kunc, M. (2021). Risks associated with the implementation of big data analytics in sustainable supply chains *Omega* https://doi.org/10.1016/j.omega.2021.102502

Laia, M.F. (2023). Development of a Cartesian Coordinate Module to Improve the Ability to Understand Mathematical Concepts *Jurnal Pendidikan Matematika* https://doi.org/10.57094/afore.v2i2.1129

Larson, J., Menickelly, M., y Wild, S. M. (2019). Derivative-free optimization methods https://doi.org/10.1017/S0962492919000060

Lechner, C., Rumpler, M., Dorley, M. C., Li, Y., Ingram, A. y Fryman, H. (2022). Developing an Online Dashboard to Visualize Performance Data—Tennessee Newborn Screening Experience *International Journal of Neonatal Screening* https://doi.org/10.3390/ijns8030049

Lee, D., Dev, H., Hu, H., Elmeleegy, H., y Parameswaran, A. (2019). Avoiding drill-down fallacies with VisPilot *Proceedings of the 24th International Conference on Intelligent User Interfaces* https://doi.org/10.1145/3301275.3302307

Lee, J., Grossman, D., Frieder, O. y McCabe, M.C. (2000). Integrating structured data and text: a multi-dimensional approach *Proceedings International Conference on Information Technology: Coding and Computing* https://doi.org/10.1109/ITCC.2000.844234

Lee, J.K., Yang, H., Park, K.H., Lee, S.Y. y Choi, S.G. (2018). The flow-reduced malware

detection system by controlling inactive/active timeout *20th International Conference on Advanced Communication Technology (ICACT)* https://doi.org/10.23919/ICACT.2018.8323759

Lex, A., Streit, M., Partl, C., Kashofer, K. y Schmalstieg, D. (2010). Comparative Analysis of Multidimensional, Quantitative Data *IEEE Transactions on Visualization and Computer Graphics* https://doi.org/10.1109/TVCG.2010.138

Li, B. y Tsui, R. (2020). How to Improve the Reuse of Clinical Data– openEHR and OMOP CDM *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1624/3/032041

Li, X., Xu, X. y Malik, T. (2016). Interactive provenance summaries for reproducible science *2016 IEEE 12th International Conference on e-Science (e-Science)* https://doi.org/10.1109/eScience.2016.7870920

Li, Z., Ao, Z., y Mo, B. (2021). Revisiting the Valuable Roles of Global Financial Assets for International Stock Markets: Quantile Coherence and Causality-in-Quantiles Approaches *Mathematics* https://doi.org/10.3390/math9151750

Liao, S.-H., Chu, P.-H. y Hsiao, P.-Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011 *Expert Systems with Applications* https://doi.org/10.1016/j.eswa.2012.02.063

Liu, J., Cao, F. y Liang, J. (2022). Centroids-guided deep multi-view K-means clustering *Information Sciences* https://doi.org/10.1016/j.ins.2022.07.093

Mangtani, A. (2021). Everything You Need To Know About Composable Applications https://ashley-mangtani.medium.com/everything-you-need-to-know-about-composable-applications-49814806ee81

Mathrani, S. (2021). Critical business intelligence practices to create meta-knowledge *Iternational Journal of Business Information Systems* https://doi.org/10.1504/IJBIS.2021.112413

McGuffin, M., Davison, G. y Balakrishnan, R. (2004). Expand-Ahead: A Space-Filling Strategy for Browsing Trees *IEEE Symposium on Information Visualization* https://doi.org/10.1109/INFVIS.2004.21

McNally, S., Roche, J. y Caton, S. (2018). Predicting the Price of Bitcoin Using Machine Learning https://doi.org/10.1109/PDP2018.2018.00060

Mensi, W., Rehman, M.U. y Vo, X.V. (2021). Risk spillovers and diversification between oil and non-ferrous metals during bear and bull market states *Resources Policy* https://doi.org/10.1016/j.resourpol.2021.102132

Mercado, A. M., Facio, M. M., Flores, F. F., y Moya, A. G. (2016). Historia Y Evolución De La Industria De Semiconductores Y La Integración De México En El Sector *European Scientific Journal* https://doi.org/10.19044/esj.2016.v12n18p65

Mioduchowska-Jaroszewicza, E. (2022). Use of A Deterministic Cash Flow Model To Support Manager Decisions *Procedia Computer Science* https://doi.org/10.1016/j.procs.2022.09.198

Miranda, B., Amorim, V. y Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction *Expert Systems with Applications* https://doi.org/10.1016/j.eswa.2019.01.012

Miret, J. (2013). Alan Turing: El descrifrado de la máquina Enigma https://blogs.elpais.com/turing/2013/06/alan-turing-el-descifrado-de-la-maquina-enigma.html

Mita, M., Ito, K., Ohsawa, S. y Tanaka, H. (2019). What is Stablecoin?: A Survey on Price Stabilization Mechanisms for Decentralized Payment Systems https://doi.org/10.1109/IIAI-AAI.2019.00023

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. y Anastasiu, D. C. (2019). Stock Price Prediction Using News Sentiment Analysis https://doi.org/10.1109/BigDataService.2019.00035

Morris, A. (2021). Data Drilling Defined: Drill Down Analysis for Business *Oracle Netsuit* https://www.netsuite.com/portal/resource/articles/data-warehouse/data-drilling.shtml

Murari, A., Lungaroni, M. y Gelfusa, M. (2019). Testing the consistency of multimachine databases for physical studies of regression *Nuclear Fusion* https://doi.org/10.1088/1741-4326/ab4285

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. y Yu, B. (2023). Definitions *Proceedings*

of the National Academy of Sciences https://doi.org/10.1073/pnas.1900654116

Murphy, K.P. (2022). *Probabilistic Machine Learning* Massachusetts Institute of Technology. https://lccn.loc.gov/2021027430

Nelsen, R.B. (2005). 14 - Copulas and quasi-copulas: An introduction to their properties and applications https://doi.org/10.1016/B978-044451814-9/50014-8

Nemeth, M., Borkin, D., Nemethova, A. y Michalconok, G. (2021). Deep drill-down analysis for failures detection in the production line *23rd International Conference on Process Control (PC)* https://doi.org/10.1109/PC52310.2021.9447500

Nemeth, M. y Michalconok, G. (2017). The initial analysis of failures emerging in production process for further data mining analysis *21st International Conference on Process Control* https://doi.org/10.1109/PC.2017.7976215

Nettayanun, S. (2023). Asset pricing in bull and bear markets *Journal of International Financial Markets* https://doi.org/10.1016/j.intfin.2023.101734

Nivethithaa, K. y Vijayalakshmi, S. (2021). Survey on Data Mining Techniques *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1947/1/012052

O Connell, M. y Molloy, K. (2023). Farming and Woodland Dynamics in Ireland during the Neolithic *Biology and Environment: Proceedings of the Royal Irish Academy* http://www.jstor.org/stable/20500109

Odoni, F., Kuntschik, P., Braşoveanu, A. y Weichselbraun, A. (2018). On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance *Procedia Computer Science* https://doi.org/10.1016/j.procs.2018.09.004

Oracle (Ed.) (2022). What is Cloud Native? https://www.oracle.com/cloud/cloud-native/what-is-cloud-native/

Orlando, G., Bufalo, M. y Stoop, R. (2022). Financial markets' deterministic aspects modeled by a low-dimensional equation *Scientific Reports* https://doi.org/10.1038/s41598-022-05765-z

Ostmann, F. y Dorobantu, C. (2023). AI in Financial Services https://doi.org/10.5281/zenodo.4916041

Pagan, A.R. y Sossounov, K.A. (2002). A simple framework for analysing bull and bear markets

Journal of Applied Economics https://doi.org/10.1002/jae.664

Palo Alto Networks (Ed.) (2022). What Is a Distributed Enterprises and Why Does Cybersecurity Matter for Branch Offices? https://www.paloaltonetworks.com/cyberpedia/what-is-distributed-enterprises-and-why-does-cybersecurity-matter#:~:text=A%20distribu 20enterprise%2C%20such%20as,gain%20access%20to%20sensitive%20data

Palza, E., Fuhrman, C. y Abran, A. (2003). Establishing a generic and multidimensional measurement repository in CMMI context *28th Annual NASA Goddard Software Engineering Workshop* https://doi.org/10.1109/SEW.2003.1270721

Pandora FMS (Ed.) (2025). ¿Sabes quién fue ENIAC? https://pandorafms.com/blog/es/eniac/

Peng, G., Huiming, Z. y Wanhai, Y. (2018). Asymmetric dependence between economic policy uncertainty and stock market returns in G7 and BRIC: A quantile regression approach *Finance Research Letters* https://doi.org/10.1016/j.frl.2017.11.001

Peng, L., Liu, S., Liu, R. y Wang, L. (2018). Effective long short-term memory with differential evolution algorithm for electricity price prediction *Energy* https://doi.org/10.1016/j.energy.2018.05.052

Pourabbas, E. y Shoshani, A. (2010). Improving estimation accuracy of aggregate queries on data cubes *Data & Knowledge Engineering* https://doi.org/10.1016/j.datak.2009.08.010

Prat, N., Comyn-Wattiau, I. y Akoka, J. (2011). Combining objects with rules to represent aggregation knowledge in data warehouse and OLAP systems *Data & Knowledge Engineering* https://doi.org/10.1016/j.datak.2011.03.004

Psiuk, M., Bujok, T. y Zieliński, K. (2012). Enterprise Service Bus Monitoring Framework for SOA Systems, *IEEE Transactions on Services Computing* https://doi.org/10.1109/TSC.2011.32

Quitaleg, A. y Ortiz, M. (2020). Design and Development of Data Warehouse Framework of Highland Vegetable Crops for Benguet *IOP Conference Series: Materials Science and Engineering* https://doi.org/10.1088/1757-899X/803/1/012035

Ragavi, V. y Geetha, N.K. (2021). A drill down analysis of the pandemic COVID-19 cases in

India using PDE *Materials Today: Proceedings* https://doi.org/10.1016/j.matpr.2020.05.595

Razzaq, A., Sharif, A., An, H. y Aloui, C. (2022). Testing the directional predictability between carbon trading and sectoral stocks in China: New insights using cross-quantilogram and rolling window causality approaches *Technological Forecasting and Social Change* https://doi.org/10.1016/j.techfore.2022.121846

Real Academia Española (2023). finanza https://dle.rae.es/finanza

Rijmenants, D. (2008). Historia de la Máquina de Cifrado Enigma https://jcampos220687.files.wordpress.com/2013/11/la-maquina-de-enigma.pdf

Robinson, A.J., Rahayu, W.J. y Dillon, T. (2009). WAD Workflow System: Data-Centric Workflow System *Australian Software Engineering Conference* https://doi.org/10.1109/ASWEC.2009.26

Rodríguez, E. (2008). Forma y raíces del microprocesador http://www.maestrosdelweb.com/historia-de-los-microprocesadores/

Rodríguez, J., Díaz, C. y Galindo, J. (2017). Herramientas cuantitativas para la planeación y programación de la producción: estado del arte. Ingeniería Industrial. Actualidad y Nuevas Tendencias https://www.redalyc.org/pdf/2150/215052403008.pdf

Saini, G., Seema y Mor, K. (2021). Machine Learning and Prophecy of Behavior: A Breakthrough in Artificial Intelligence *IOP Conference Series: Materials Science and Engineering* https://doi.org/10.1088/1757-899X/1099/1/012026

Santra, A., Komar, K., Bhowmick, S. y Chakravarthy, S. (2022). From base data to knowledge discovery - A life cycle approach - Using multilayer networks *Data & Knowledge Engineering* https://doi.org/10.1016/j.datak.2022.102058

Saravanan, R. y Sujatha, P. (2018). A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification https://doi.org/10.1109/ICCONS.2018.8663155

Schär, F. (2021). Decentralized Finance: On Blockchain- and Smart Contract-Based Financial Markets *FRB of St. Louis Review* https://doi.org/10.20955/r.103.153-74

Sen, S., Chaki, N. y Cortesi, A. (2009). Optimal Space and Time Complexity Analysis on the

Lattice of Cuboids Using Galois Connections for Data Warehousing *Fourth International Conference on Computer Sciences and Convergence Information Technology* https://doi.org/10.1109/ICCIT.2009.185

Sen, S. y Chaki, N. (2011). Efficient Traversal in Data Warehouse Based on Concept Hierarchy Using Galois Connections *Second International Conference on Emerging Applications of Information Technology* https://doi.org/10.1109/EAIT.2011.69

Shabaninejad, S., Khosravi, H., Indulska, M., Bakharia, A. y Isaias, P. (2020). Automated insightful drill-down recommendations for learning analytics dashboards *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* https://doi.org/10.1145/3375462.3375539

Shah, D., Isah, H. y Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques *International Journal of Financial Studies* https://doi.org/10.3390/ijfs7020026

Shailak J. (2018). The growth of cryptocurrency in India: Its challenges and potential impacts on legislation https://doi.org/10.13140/RG.2.2.14220.36486

Shalev-Shwartz, S. y Ben-David, S. (2014). *Understanding Machine Learning From Theory to Algorithms* Cambridge University Press. https://www.cs.huji.ac.il/~shais/UnderstandingMa

Shende, S. W. (2021). Artificial intelligence and machine learning for internet of things *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1913/1/012151

Shi, K., Irani, P. y Li, B. (2005). An evaluation of content browsing techniques for hierarchical space-filling visualizations *IEEE Symposium on Information Visualization* https://doi.org/10.1109/INFVIS.2005.1532132

Shiller, R.J. (2023). *Finance and the Good Society* Princeton University Press.

Shmueli, G., Bruce, P.C. y Patel, N.R. (2016). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner* Wiley.

Silva, V.H. (2024). EDA Methods for Insider Trading GitHub Repository https://github.com/victorhugosilvablancas/EDA_MethodsForInsiderTrading/

Sirignano, J. y Cont, R. (2019). Universal features of price formation in financial markets: perspectives from deep learning *Quantitative Finance* https://doi.org/10.1080/14697688.

2019.1622295

Sitanggang, I., Trisminingsih, R., Khotimah, H. y Syukur, M. (2019). Usability testing of SOLAP for Indonesia agricultural commodity *IOP Conference Series: Earth and Environmental Science* https://doi.org/10.1088/1755-1315/299/1/012054

Slater, P. J. (1978). Centers to centroids in graphs *Journal of Graph Theory* https://doi.org/10.1002/jgt.3190020304

Smartz Solutions (Ed.) (2020). What is Privacy-enhancing Computation? https://smartz-solutions.com/what-is-privacy-enhancing-computation/

Soulis, K., Varlamis, I., Giannakoulopoulos, A. y Charatsev, F. (2013). A tool for the visualisation of public opinion *International Journal of Electronic Governance* https://doi.org/10.1504/IJEG.2013.058404

Sousa, V. D., Driessnack, M., y Mendes, I. (2007). An overview of research designs relevant to nursing: Part 1: quantitative research designs *Revista Latino-Americana de Enfermagem* https://doi.org/10.1590/S0104-11692007000300022

Soussi, N. (2021). Big-Parallel-ETL: New ETL for Multidimensional NoSQL Graph Oriented Data *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1743/1/012037

Sovbetov, Y. (2023). Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin *Journal of Economics and Financial Analysis* https://ssrn.com/abstract=3125347

Splechtna, R., Gračanin, D., Todorović, G., Goja, S., Bedić, B. y Hauser, H. (2023). Interactive Visual Analysis of Structure-borne Noise Data *IEEE Transactions on Visualization and Computer Graphics* https://doi.org/10.1109/TVCG.2022.3209478

Sumathi, S. y Sivanandam, S.N. (2006). *Data Mining Tasks Techniques and Applications* Introduction to Data Mining and its Applications. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-34351-6_7

Supriana, C. (2020). Designing database lecture model in informatics engineering study program *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1516/1/012002

Talpsepp, T., Liivamägi, K. y Vaarmets, T. (2020). Academic abilities *Journal of Banking &*

Finance https://doi.org/10.1016/j.jbankfin.2020.105848

Tamilselvi, R. y Kalaiselvi, S. (2013). An Overview of Data Mining Techniques and Applications *International Journal of Science and Research* https://www.ijsr.net/getabstract.php?paperid=IJSROFF2013059

Tan, F., Cascante, P., Guo, X., Wu, H., Feng, S. y Ordonez, V. (2019). Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries *33rd Conference on Neural Information Processing Systems, Vancouver, Canada.* https://proceedings.neurips.cc/paper/2019/hash/471c75ee6643a10934502bdafee198fb-Abstract.html

Tang, M. (2021). Design of Library Mobile User Behavior Analysis model for Personalized Information Service *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1982/1/012179

Tang, S., Yang, J., Liu, Y., Wu, Z. y Chen, B. (2007). An Energy Efficient Design of Multi-resolution Storage for Ubiquitous Data Management *IFIP International Conference on Network and Parallel Computing Workshops* https://doi.org/10.1109/NPC.2007.170

Tang, Y. y Lan, Y. (2021). Design of University Financial Decision-Making Platform Based on Data Mining *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1881/4/042063

Taylor, R. (2002). Tiber river bridges and the development of the ancient city of Rome *The Waters of Rome* https://waters.iath.virginia.edu/Journal2TaylorNew.pdf

Thampanya, N., Nasir, M.A. y Duc Huynh, T.L. (2020). Asymmetric correlation and hedging effectiveness of gold & cryptocurrencies: From pre-industrial to the 4th industrial revolutionXXX ERROR UNICODE X2730XX *Technological Forecasting and Social Change* https://doi.org/10.1016/j.techfore.2020.120195

Trindade, A., Uryasev, S., Shapiro, A. y Zrazhevsky, G. (2007). Financial prediction with constrained tail risk *Journal of Banking & Finance* https://doi.org/10.1016/j.jbankfin.2007.04.014

United States Census Bureau (2021). 1790 https://www.census.gov/history/www/through_the_decades/overview/1790.html

United States Census Bureau (bis) (2021). By Decade (1870-1880) https://www.census.gov/programs-surveys/decennial-census/decade.1880.html

United States Congress (2002). Sarbanes-Oxley Act https://www.govinfo.gov/content/pkg/COMPS-1883/pdf/COMPS-1883.pdf

Vassiliadis, P., Marcel, P. y Rizzi, S. (2019). Beyond roll-up s and drill-down s: An intentional analytics model to reinvent OLAP *Data & Knowledge Engineering* https://doi.org/10.1016/j.is.2019.03.011

Veras, R.; Marques, A.G., Santiago, N.J., Simões, Meiguins, A.S. y Meiguins, B.S. (2011). Design Considerations for Drill-down Charts *15th International Conference on Information Visualisation* https://doi.org/10.1109/IV.2011.65

Vieira, M., Chino, F., Traina, C. y Traina, A. (2010). A visual framework to understand similarity queries and explore data in Metric Access Methods *International Journal of Business Intelligence and Data Mining* https://doi.org/10.1504/IJBIDM.2010.036125

Walkme (Ed.) (2022). What is Total Experience? https://www.walkme.com/glossary/total-experience/

Wang, C., Yu, F., Liu, Y., Li, X., Chen, J., Thiyagalingam, J. y Sepe, A. (2021). Deploying the Big Data Science Center at the Shanghai Synchrotron Radiation Facility: the first superfacility platform in China *Machine Learning: Science and Technology* https://doi.org/10.1088/2632-2153/abe193

Wang, H., Lu, S. y Zhao, J. (2019). Aggregating multiple types of complex data in stock market prediction: A model-independent framework *Knowledge-Based Systems* https://doi.org/10.1016/j.knosys.2018.10.035

Wang, H., Wang, C., Liu, K., Meng, B. y Zhou, D. (2004). VisDM-PC: a visual data mining tool based on parallel coordinate *Proceedings of 2004 International Conference on Machine Learning and Cybernetics* https://doi.org/10.1109/ICMLC.2004.1382382

Wang, M. y Iyer, B. (1997). Efficient roll-up and drill-down analysis in relational database *Workshop on Research Issues on Data Mining and Knowledge Discover* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b32f291d9bf60a5827ac3580a7

Weidong, J. y Wanlu, N. (2022). A Dynamic Control Method of Population Size Based on Euclidean Distance *Journal of Electronics & Information Technology* https://doi.org/10.11999/JEIT210322

Wetzstein, B., Leitner, P., Rosenberg, F., Brandic, I., Dustdar, S. y Leymann, F. (2009). Monitoring and Analyzing Influential Factors of Business Process Performance *IEEE International Enterprise Distributed Object Computing Conference* https://doi.org/10.1109/EDOC.2009.18

Wright, C. y Hunt, W. (1900). The History and Growth of the United States Census https://www.census.gov/library/publications/1900/dec/history-growth-census.html

Xie, S., Hu, Q., Zhang, J. y Yu, P.S. (2015). An effective and economic bi-level approach to ranking and rating spam detection *IEEE International Conference on Data Science and Advanced Analytics* https://doi.org/10.1109/DSAA.2015.7344794

Yang, X. y Luo, Y. (2014). Rumor Clarification and Stock Returns: Do Bull Markets Behave Differently from Bear Markets? *Emerging Markets Finance and Trade* https://doi.org/10.2753/REE1540-496X500111

Yazlyuk, B., Guley, A., Brukhanskyi, R., Shovkoplias, H. y Shvydka, T. (2018). Basic principles of financial markets regulation and legal aspects of the legislative requirements *Central and Eastern European Online Library* https://www.ceeol.com/search/article-detail?id=741354

Yin, J., Zhang, Q. y Karunanithi, M. (2015). Unsupervised daily routine and activity discovery in smart homes *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* https://doi.org/10.1109/EMBC.2015.7319636

Ying, X. (2019). An Overview of Overfitting and its Solutions *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1168/2/022022

You, J., Xi, J., Zhang, P. y Chen, H. (2008). A Parallel Algorithm for Closed Cube Computation *Seventh IEEE/ACIS International Conference on Computer and Information Science* https://doi.org/10.1109/ICIS.2008.63

Yu, X. (2021). The Application of Data Warehouse in Teaching Management in Colleges and Universities *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-

6596/1738/1/012090

Yunita, A., Santoso, H.B. y Hasibuan, Z.A. (2021). Research Review on Big Data Usage for Learning Analytics and Educational Data Mining: A Way Forward to Develop an Intelligent Automation System *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1898/1/012044

Zaric, S. (2022). Drill Down vs. Drill Through Reportes https://databox.com/drill-down-vs-drill-through-report#drill-down

Zhang, D., Tang, S., Yang, D. y Jiang, L. (2007). An Effective Drill-Down Paths Pruning Method in OLAP *Fuzzy Systems and Knowledge Discovery, Fourth International Conference* https://doi.org/10.1109/FSKD.2007.148

Zhang, L., Qin, H., Liu, K. y Wu, T. (2012). System composition and multidimensional analysis tools of the Multidimensional Hyperspectral Database for Rocks and Minerals *4th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing* https://doi.org/10.1109/WHISPERS.2012.6874253

Zhang, M., Peng, H. y Yan, X. (2020). Improved algorithm of decision tree based on neural network *Journal of Physics: Conference Series* https://doi.org/10.1088/1742-6596/1693/1/012081

Zhang, Y., Chan, S., Chu, J. y Sulieman, H. (2020). On the Market Efficiency and Liquidity of High-Frequency Cryptocurrencies in a Bull and Bear Market *Journal of Risk and Financial Management* https://doi.org/10.3390/jrfm13010008

Zheng, X., Zhu, M., Li, Q., Chen, C. y Tan, Y. (2019). FinBrain: when finance meets AI 2.0 *Frontiers of Information Technology & Electronic Engineering* https://doi.org/10.1631/FITEE.1700822

Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning *National Science Review* https://doi.org/10.1093/nsr/nwx106

Ziegler, C.N., Skubacz, M. y Viermetz, M. (2008). Mining and Exploring Unstructured Customer Feedback Data Using Language Models and Treemap Visualizations *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* https://doi.org/10.1109/WIIAT.2008.69

Zou, B., You, J., Ding, J. y Sun, H. (2019). TAVO: A Tree-like Analytical View for OLAP *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing* https://doi.org/10.1109/PACRIM47961.2019.8985106

9. Anexos

9.1. Anexo 1.

Artículo en revista indexada:

Silva-Blancas, V. H., Álvarez-Alvarado, J. M., Herrera-Navarro, A. M. y Rodríguez-Reséndiz, J. (2023). Tendency on the Application of Drill-Down Analysis in Scientific Studies: A Systematic Review. *Technologies*, 11(4), 112. `https://doi.org/10.3390/technologies11040112`

Figura 27

Carátula del artículo indexado número 1.

9.2. Anexo 2.

Artículo en revista indexada:

Silva-Blancas, V. H., Jiménez-Hernández, H., Herrera-Navarro, A. M., Álvarez-Alvarado, J. M., Córdova-Esparza, D. M. y Rodríguez-Reséndiz, J. (2024). A Clustering and PL/SQL-Based Method for Assessing MLP-Kmeans Modeling. *Computers*, 13(6), 149. `https://doi.org/10.3390/computers13060149`

Figura 28

Carátula del artículo indexado número 2.

9.3. Anexo 3.

Artículo en revista indexada:

Silva-Blancas, V. H., Jiménez-Hernández, H., Herrera-Navarro, A. M., Álvarez-Alvarado, J. M., Córdova-Esparza, D. M. y Rodríguez-Reséndiz, J. (2025). Infinite Type Centroid Java Library: An Implementation of Parameterized Coordinates for an Enhanced Centroid Calculation during K-means Classification. *Software Impacts*. `https://doi.org/10.1016/j.simpa.2025.100751`

Figura 29

Carátula del artículo indexado número 3.