UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

COMPARACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA DESCUBRIR INFORMACIÓN RELEVANTE DE VENTAS DE UNA MYPE COMERCIAL

PARA OPTAR EL TÍTULO PROFESIONAL DE

INGENIERO DE SISTEMAS

Autor:

Bach. Carlos Johny Miñano Sanchez

ORCID:

https://orcid.org/0000-0003-4273-9383

Asesor:

Mg. Mejia Cabrera Heber Ivan

ORCID:

https://orcid.org/0000-0002-0007-0928

Línea de Investigación:

Infraestructura, Tecnología y Medio Ambiente

Pimentel – Perú 2022

APROBACIÓN DEL JURADO

TITULO

_____

Bach. Miñano Sánchez Carlos Johny

Autor

_____

Mg. Mejía Cabrera Heber Iván

Asesor

_____ _____

Mg. Atalaya Urrutia Carlos William

Presidente de Jurado

Mg. Diaz Vidarte Miguel Orlando

Secretario de Jurado

_____

Mg. Mejía Cabrera Heber Iván

Vocal de Jurado

Dedicatorias

A ti mujer que me entregaste el alma y corazón. Sofia.
A mis padres Amelia y Carlos
A mi hijo Jesús, y
A mi abuelita Arminda

Me inspiraron a luchar para alcanzar
mis sueños y metas.

Agradecimientos

Al ing. Heber Iván Mejía Cabrera que con sus enseñanzas y asesorías pude ver consolidado este trabajo.

Al C.P.C. César A. Puicón Estrada y Comercial Damián EIRL por su predisposición, apertura y confianza.

Resumen

Perú aplicó Inteligencia Artificial (IA) en empresas de envergadura, constituyó apalancamiento para la productividad y se estimó que impulsaría un crecimiento de hasta 6% en PBI al 2028. Muchos emprendimientos no fueron sostenibles ante la falta de herramientas tecnológicas como minería de datos. Se han realizado soluciones de negocio en tecnología de información e inteligencia artificial, para superar incidencias en fraude electrónico, toma de decisiones, soluciones para ventas y otros que no han sido suficiente por alto costo que representan, para las Medianas y Pequeñas Empresas MYPE´s acceder a tales herramientas tecnológicas, para fortalecimiento de capacidades. En ese sentido se desarrolló un método que permita conocer que técnicas de minería de datos existentes, proporcionan mejor desempeño, para descubrir información relevante de ventas que permita apuntalar sus objetivos de negocio y proporcione confiabilidad y eficiencia. Este método comprendió elegir una MYPE comercial en virtud al área de influencia de la Universidad señor de Sipán que proporcionó los datos y se construyó un data set al cual se le aplicó normalización de variables de entrada haciendo uso de la técnica de escalado de variables de Min y Max, procesándose 5,522 registros y a éstos se les aplicó las técnicas de minería seleccionadas por su eficiencia y rendimiento concorde a la investigación de las bases de datos Ieeexplore, Scopus y Science Direct. Posteriormente haciendo uso de librerías contenidas en la suite Anaconda Navigator, junto a Python como herramienta de programación y Jupyter como editor, se logró resultados que evidencian que regresión logística es la técnica eficiente en tanto que las demás no ofrecen óptimos resultados en indicadores tiempo de respuesta y precisión; concluyendo que la técnica de clasificación en lo concerniente a regresión logística es la más eficiente con un promedio de tiempo de respuesta de 0.0620 segundos, nivel de precisión (P) de 99.93%, consumo de CPU 4.6 Gb; consumo de memoria de 6.13; error cuadrático medio (ECM) de 0.00090 y desviación absoluta media (MAD) 0.000898.

Palabras Clave: Minería de Datos, precisión, tiempo de respuesta, técnica eficiente, MYPE.

Abstract

Peru applied Artificial Intelligence (AI) in large companies, leveraged productivity and was estimated to drive up to 6% GDP growth by 2028. Many ventures were not sustainable due to the lack of technological tools such as data mining. Business solutions in information technology and artificial intelligence have been developed to overcome incidences of electronic fraud, decision making, sales solutions and others that have not been sufficient due to the high cost they represent for medium and small MSEs to access such technological tools for capacity building. In this sense, a method was developed to find out which existing data mining techniques provide better performance in order to discover relevant sales information to support their business objectives and provide reliability and efficiency. This method involved choosing a commercial MSE in the area of influence of the Universidad Señor de Sipán, which provided the data, and a data set was constructed to which input variable normalization was applied using the Min and Max variable scaling technique, processing 5,522 records and applying the mining techniques selected for their efficiency and performance according to the research of the Ieeexplore, Scopus and Science Direct databases. Subsequently, using libraries contained in the Anaconda Navigator suite, together with Python as a programming tool and Jupyter as an editor, results were obtained that show that logistic regression is the efficient technique, while the others do not offer optimal results in terms of response time and accuracy indicators; concluding that the classification technique concerning logistic regression is the most efficient with an average response time of 0. 0620 seconds, precision level (P) 99.93%, CPU consumption 4.6 Gb; memory consumption 6.13; mean square error (MSE) 0.00090 and mean absolute deviation (MAD) 0.000898.

Keywords: Data mining, accuracy, response time, efficient technique, MSE.

Índice

I. INTRODUCCIÓN

1.1. Realidad Problemática.

Latino América (LAM) quedó rezagada con relación al empleo de la Inteligencia Artificial (IA) y Machine Learning. Perú lo implementó a través de grandes corporaciones como algunos bancos mediante aplicaciones de robots online. Su aplicación constituyó un apalancamiento para fortalecer su productividad y sin embargo existió una baja capacidad de adopción de la tecnología, tal como lo precisó el banco mundial. Así mismo se previó un crecimiento del PBI donde existió intervención de la IA en escenarios negativos, neutral y positivo que podría repercutir en el nivel porcentual de su crecimiento, si se tiene en cuenta la tendencia histórica desde 1990, que se representa en la Figura Nº 1, con una proyección al 2028 donde el crecimiento económico de Perú sin poner en práctica el uso de IA sería un escenario negativo con un nivel porcentual de 3,7%, en tanto que se evidencia de que respecto a la intervención o no de la IA en modo neutral, el crecimiento del PBI sería de 4.9% y de promoverse el uso de la IA, el crecimiento sería positivo y alcanzaría el 6%. (Albrieu, Rapetti, López, & Larroulet, 2018).

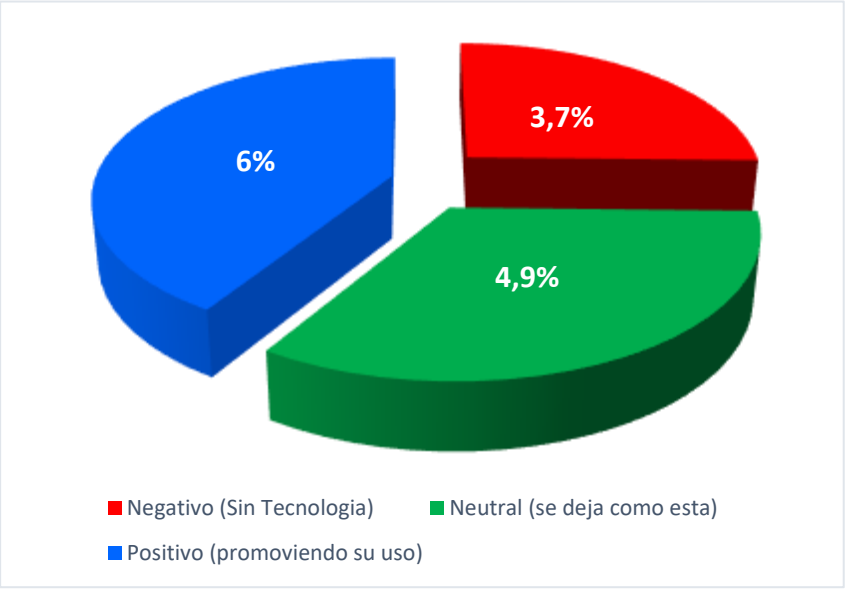


Figura 1. Representación de proyección de crecimiento del PBI (2018-2028) con o sin tecnología. Fuente: Informe CIPPEC – 2018.

Tabla 1.

Nivel porcentual de Comercios y el uso de TIC por Sector Económico, Perú-2018

Actividad económica	% Usan computador	% Usan teléfono
Agrícola y Ganadero	56,7	96,0
Pesquero	88,5	86,6
Minero	79,6	82,1
Manufacturero	76,2	78,9
Construcciones	94,9	82,6
Comercializaciones	61,0	76,5
Servicios	80,1	77,3

Nota: (Ministerio de la Producción, 2020). p.12-29.

El bajo crecimiento de las economías emergentes como lo es la economía LAM se debe a la escasa transformación digital, que la crisis y el efecto disruptivo del SARS-CoV-2 obligó a que las PYME´s opten por la transformación tecnológica, que redireccione sus ventas a través del uso de las tecnologías e internet. (Heredia, 2020). Perú a través de uno de sus ministerios (MEF), instituyó políticas, para que el estado en general (todo nivel de gobierno) y en el afán de reactivar la economía aceleren la adquisición de bienes a las MYPE, hasta S/ 736 millones de soles, donde el sector comercio es un sector beneficiado. (Ministerio de Economía y Finanzas, 2020).

La MYPE’s en Perú sostuvieron una marcada brecha respecto a la Implementación de tecnologías de Información. Solo el 29,7% contó con sitio web en el año 2018, para ofertar sus productos, 2.4% lo realizaba en línea y sobre del 35% no usaba la banca en modo electrónico. La innovación, pudo coadyuvar a reducir esa brecha que también apuntaló la salida a la crisis y a la adaptación a cambios disruptivos que nos llevó el estado económico por el SARS-CoV-2. (Diaz, Deza, & Moreno, 2020).

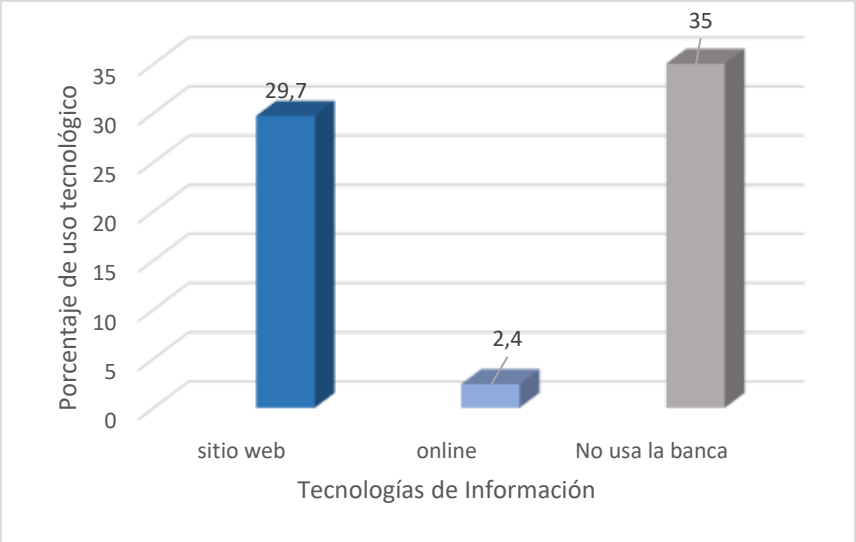


Figura. 2 Representación de uso tecnológico en PYMES peruanas. Fuente: Informe BID Desafíos del desarrollo en el Post Covid–19. *(Diaz, Deza, & Moreno, 2020)*.

Uno de los factores de frenaje respecto al empleo de la data mining en las MYPES en Perú fue porque en Latinoamérica su uso fue considerado como tecnología emergente debido a las brechas existentes. (Heredia, 2019). No obstante CEPAL y UNCTAD en el estudio Global Initiative towards post-Covid-19 resurgence of the MSME sector, primer diálogo regional LAM entre representantes de las microempresas, pymes y de defensa de la competencia en Agosto del 2020, con la finalidad de promover generación de oportunidades e intercambio entre entidades que permitan apuntalar el mutuo aprendizaje; concluyeron que fue necesario que las pymes tuvieran como aliadas a las Tecnología digitales y la colaboración regional, para la apertura de una nueva economía digital con importancia en la economía de datos en un marco de normas y estrategias, para el logro, consolidación y generación de nuevas oportunidades de negocio de la mano con nuevo diseño de políticas, dinámicas descentralistas con preminencia en la protección de datos debido a que el uso de los datos, tuvo fuerte impacto en el crecimiento. (Dine & Nuñez, 2020).

En Perú al primer trimestre del 2021 existía un total de 2 millones 838 mil 494 empresas que con relación al año anterior creció en 2,6%, respecto al 2020 y se dieron de baja 8,087 empresas y se crearon 68,811 nuevas empresas

proyectando un 2,4% de crecimiento y una tasa de mortalidad de empresas de 0.3%. (INEI, 2021).

Una propuesta de método ágil de arquitectura empresarial (AE), para pequeñas organizaciones LAM, afirmaba que las TI no están alineadas con los objetivos del negocio y representan un problema relevante para las empresas; esto debido a que a las MYPE's les demanda festinar recursos y al no estar alineadas a sus objetivos, fueron un factor de frenaje para alcanzar sus metas puesto que el análisis de datos o gestión de datos, ocupó el tercer lugar con un 23,4%, en una escala de 10 principales problemas relevantes de las organizaciones LAM (Mejía, Tuesta, Samillan, & Forero, 2021).

La alta brecha tecnológica que existió respecto a las pymes los compradores mensuales en promedio que hasta antes de la pandemia bordearon los 5.1 millones cerró en el 2020 con 11.8 millones de compradores con un nivel alcanzado del 36% de la población; no obstante, éstas ventas estuvieron dirigidas a e-commerce retails, con un crecimiento del 400%. Así también sucedió lo mismo con Mercado Libre que duplicó sus ventas. (CAPECE, 2021).

La estrategia nacional de IA en Perú empleó a enero del 2021 alrededor del 20% de las instituciones públicas que resolvieron la encuesta nacional; solo el 7% confirmó que empleaban la Inteligencia Artificial. De 375 gobiernos locales peruanos, solo 29 usaron la Inteligencia Artificial; de 28 Gobierno regionales solo 2 la usó. Respecto al poder ejecutivo de 17 que desarrollaron la encuesta también solo 6 usaron Inteligencia Artificial y de las universidades nacionales existentes en nuestro país de 6 que desarrollaron también la referida encuesta; solo 1 usó IA a excepción de las 12 empresas del estado que desarrollaron dicha encuesta 12 si la usaron. (Presidencia del Concejo de Ministros, 2021).

Diversas soluciones de ingeniería se desarrollaron en LAM, investigaciones realizadas nos dan muestras de la precariedad con que muchos emprendimientos iniciaron, que pudieron ser sostenibles si hubiesen dejado de lado la gestión intuitiva, para convertirse en actividades económicas

apalancadas con herramientas tecnológicas como el uso de la minería de datos, que pudo contribuir al fortalecimiento de las MYPE´s con redundancia en el mejoramiento de la contribución de impuestos a favor de cada país de la región. Colombia desarrolló un estudio para apoyar a una asociación productora de café denominada las Rosas en el que basó su estudio en productividad y fortalecimiento personal de las mujeres emprendedoras que conformaron dicha organización, donde se encontró factores que condicionaron la rentabilidad por hectárea de café sembrada y que su crecimiento personal y empresariales se consolidó en seis años de trabajo. (Martinez & Palencia, 2021).

Perú fue uno de los países con la más alta incidencia en fraude electrónico, realizó soluciones al respecto para superar dichos factores perturbadores del mercado donde tuvo que hacer uso del data mining, para la configuración de un modelo mecánico que clasifique transacciones de banca electrónica (internet y banca móvil), cuyo problema fue la clasificación respecto a transacciones íntegras o fraudulentas, basada en árboles de decisión y teniendo como fuente la tabla de registro LOG de las transacciones electrónicas, con un data set de 4980 operaciones genuinas concernientes a 7 tipos de transacciones como envío de efectivo móvil, transferencia cuentas de terceros, pago tarjeta de crédito entre otros, entre otras transacciones. Las conclusiones fueron que fue una técnica eficiente, para descubrir nuevo conocimiento y poder identificar si una operación fue o no fraudulenta en virtud al modelo de aprendizaje del 99.02% de exactitud. (Ñaupas, 2016).

El empleo o uso de minería de datos, para la resolución de la problemática en toma de decisiones de la empresa Computer House en la ciudad de Lima- Perú, cuya metodología se solventó en el uso de la metodología Ralph Kimball, haciendo uso además de la encuesta predominando el diseño pre-experimental y teniendo como base a una fuerza laboral de 30 trabajadores, obtuvieron los resultados en la cual el BI influyó y apuntaló la gestión administrativa, fortaleciéndola y consolidando mejores resultados. Después de aplicada la t-student con $\alpha = 0,05$ se concluyó que la práctica en uso de minería de datos a través del BI mejoró la toma de decisiones al alcanzar un nivel de satisfacción

de 18% y el tiempo de generación de reportes menor donde el 1.67 aplicando el BI y 2.81 sin el BI. (Carhuana & Carhuana, 2019).

La aplicación de Minería de Datos en un negocio de venta de suplementos nutricionales y accesorios que comercializa 137 productos clasificados que busca determinar patrones de consumo y aumentar cross selling, debido a en el negocio existen productos con mayor rotación que otros y en la esperanza de mejorar sus ventas y campaña de marketing de la empresa Lab Nutrition, que empleo la metodología estructurada exploratoria CRISP-DM por su neutralidad y facilidad de implementación que tuvieron como resultados el patrón de consumidores por producto, el perfil de los consumidores por cada producto expendido y concluyeron que es probable que del 100% de sus clientes en edad de 25-32 que realizan ejercicios físicos compren determinado producto; clientes mujeres que hacen ejercicios con elíptica adquieran un producto determinado también; los que realizan kickboxing en mismo nivel porcentual, tienen menos de 2 hijos y que clientes que tienen de 3 a más hijos tienen un peso de entre 80,4 - 91,5 y consuman un producto determinado también. (Grández Márquez, 2017)

1.2. Trabajos previos.

Eti & İnel (2020) At a Research on the Comparison of Classification Algorithm In Finance, realizado en Turquía, cuyo estudio estuvo basado en solventar el problema de clasificación en el sentido de que si una entidad fue rentable o no. Hecho que derivó en examinar los algoritmos de clasificación en un marco de la minería de datos. Razones que condujeron a los autores a la ejecución de la metodología de data mining y aprendizaje automático con la finalidad de extraer la información solventándose en conceptos de 5V, como volumen, velocidad, variedad, verificación y valor; además de Algoritmos de aprendizaje automático que a su vez se sostuvieron en categorías técnicas de aprendizajes supervisados y no supervisados. El método que propusieron fue técnicas data mining y dividieron en forma aleatoria los datos de manera que 50% fue para datos de aprendizaje y 50%, para prueba; aplicando ambas segmentaciones a la regresión, análisis de separación, redes neuronales, vectores de soporte, regla difusa, algoritmo de bosque aleatorio y modelo de árboles de decisión,

utilizando Knime y Matlab. Los resultados obtenidos se compararon con datos reales y considerando tasas precisas de predicción, se determinó el éxito de los algoritmos de clasificación, finalidad que demostró la precisión de si una empresa es o no rentable. Los resultados del estudio reflejaron en una tabla de ratios de precisión porcentual y se obtuvo que la máquinas de vectores soporte otorgó 85,714%, árboles de decisión 94,286, regresión logística 91,429%, redes neuronales artificiales 94,286%, reglas difusas (lógica) 91,429%, algoritmo de bosque aleatorio 88,571%, análisis discriminante 90.00%; que reflejó que el algoritmo de árboles de decisión nos da la correcta clasificación con un 94,286% en los datos de prueba (33 de 35 empresas seleccionadas); concluyeron que el mejor rendimiento lo determinó las redes neuronales y árboles de decisión con la finalidad de conocer si una empresa fue o no rentable.

Pabón, Torres & Bucheli (2020), en la Investigación denominada: Un Enfoque de Análisis Inteligente de Datos para apoyar la Relación con los Clientes, efectuada en España, para tratar el problema de datos no estructurado, mediante la extracción de indicadores de desempeño del servicio al cliente a la cual alinearon técnicas de Bussines Intelligent and procesamiento natural del lenguaje (NLP). El método que se propuso se basó en combinación de técnicas de inteligencia de negocios y procesamiento de lenguaje natural a efecto de extraer indicadores de desempeño de diversas fuentes de datos; que se apoyó en el procesamiento de datos estructurados, así como también no estructurados para la construcción de una ETL y finalmente procesada, donde los resultados a los que se llegó fue que el algoritmo de árboles de decisión tiene mejor rendimiento y la dificultad hallada fue en los textos largos, conforme se evidenció en dos experimentos; donde con la métrica F-score mayor al 85% demostró que la técnica NLP sirve para extraer información de los clientes, conforme a los resultados del experimento que lo demuestran. Finalmente arribaron a las conclusiones de que el procesamiento de datos, tanto estructurados como no estructurados y la combinación de técnicas de inteligencias de negocios, permiten obtener indicadores útiles, para mejorar la relaciones con los clientes.

Palakshappa, A. & Patil, M. (2019) En el estudio de investigación RFM (Reciente, Frecuencia y Monetario), Model For Customer Purchase Behavior Using K-Means Algorithm, realizado en la India, cuyo problema fue la toma de decisiones respecto al marketing, ventas efectivas, determinación de estrategias, comportamiento del cliente, patrones adoptados y mejoramiento del servicio al cliente con redundancia en beneficio a la empresa. La metodología aplicada se sostuvo en el uso del algoritmo que permiten la segmentación como K-means Clustering, por ser rápida, mejores resultados y reducir la tasa de clasificación errónea que consistió en efectuar una análisis de exploración y proceso de los datos, realizar el análisis RFM (Recency, Frequency and Monetary), por antigüedad, periodicidad e importe monetario y utilizando el algoritmo, cálculo de la puntuación de silueta y evaluación de clústeres; con lo cual se evidencia que haciendo uso del RFM, se contribuyó a obtener resultados óptimos respecto a la taxonomía ya sea de clientes o productos relativos a las transacciones de ventas por la comparación de varios parámetros de ventas como ventas recientes, ventas por frecuencia y volumen de ventas. Los resultados a los que se arribó es que con relación al RFM donde se calculó para el número de clúster representado por K; para K=3 y K=5, en el caso de K=5 son menos óptimos si se compara con K=3, aplicando el análisis a los valores monetarios y recientes; y aplicando la partición de clientes respecto al monto generado con transacción reciente y agrupación de clientes por importe generado con transacciones frecuentes; donde sus resultados en el último caso mostró que en el clúster 0 se obtiene en registro de actual 161.191479, en registro de frecuencia 16.761959 y en registro de cantidad 291.852580; para el clúster 1 se obtiene en registro actual de 11.373230, en registro de frecuencia 209.371490 y en registro de cantidad 5316.800437; en tanto que para el clúster 2, con el cual se llega al K=3, se tiene en registro de actual 20.323096, en registro de frecuencia 48.877509 y en registro de cantidad 894.321423; arribando a la conclusión de que este tipo de segmentación, permitió ponderar resultados otorgándole al negocio una óptima solución en cuanto a la toma de decisiones respecto a productos con similares características, ventas recientes, frecuencia de ventas, preferencia de productos y valor monetario de las ventas.

Silva, J., Varela, N., Borrero, L. & Rojas, R. (2019). En el estudio Association Rules Extraction For Customer Segmentation In The Smes Sector Using The A priori Algorithm, el cual fue efectuado en Colombia, para solucionar la escasez de estrategias que permitan la extracción de la información de los datos con la finalidad de determinar características contables de las empresas más rentables, perfil de clientes mediante segmentación para la construcción de estrategias diferenciadas. La metodología utilizada fue la de CRISP-DM que se basa en comprensiones del negocio, datos y preparación de ellos, representación o modelado, evaluación e implementación. Estas fases pueden accionarse en forma indistinta respecto a su orden de prelación debido a que cada fase se divide en niveles y tareas generales; y, SEMMA que consiste en la discriminación, exploración y representación de ingentes cantidades de datos, para develar los estándares comerciales desconocidos y en la aplicación del algoritmo CLARA. Los resultados obtenidos de su aplicación derivaron en grupos de clientes basados en atributos RFM cada grupo con su respectiva etiqueta que describía fidelización de clientes, mediante la cual se generó las reglas de clasificación, para evaluar los métodos de segmentación utilizada y se descubrió los niveles de lealtad grupo 1 alto, grupo 2 bajo; grupo 3 medio y grupo 4 muy bajo, que identificó que en niveles de fidelización baja y muy baja se distribuyen más del 50% de los clientes; donde las conclusiones fueron que los segmentos generados por el algoritmo CLARA de k-medoides otorga mayor precisión y los grupos de clientes de pymes revelaron nivel de lealtad como alto, medio, bajo y muy bajo; que permitió a las empresas desarrollar estrategias que les permitan retener a clientes.

Varela, Cabrera, Lopez, Vilora, Gaitán& Ardila (2019), en la investigación denominada: Methodology for the Reduction and Integration of Data in the Performance Measurement of Industries Cement Plants in International Conference on Data Mining and Big Data, efectuado en Cuba ante la carencia de instrumentos que permitan la administración control y reducción de datos, para las mediciones de desempeño empleando la estandarización en cuanto a criterios y datos que permitan comparar con otras mediciones de empresas similares; la metodología en la que se solventó el estudio realizado fue la

técnicas de preprocesamiento sobre una agrupación de datos, para generar un tipo representativo con más beneficios; y, también se utilizó la creación de indicadores integrales a efecto de evaluar la efectividad del sistema. Los resultados obtenidos fue que se logró determinar el grado de efectividad de la empresa de cemento con un nivel de eficiencia en el proceso con el estándar de Alta Eficiencia de 95-100; mediana eficiencia entre el 85 – 94 puntos y un estándar de baja eficiencia que llegó a un nivel de 85 puntos. Finalmente concluyeron que usó índices sintéticos, para que varios indicadores se conviertan en un solo valor, muestra un camino óptimo para determinar el grado de eficiencia en tiempo real.

Martínez, Carrasco, García, Gallego, & Herrera (2019) en el estudio Comparison Between Fuzzy Linguistic RFM Model And Traditional RFM Model Applied to Campaign Management, donde el problema fue la segmentación de cliente basada en la frecuencia de valor monetario (RFM) en la que se evaluó el comportamiento de compra, para el diseño o determinación de una campaña exitosa, que debió enfocarse por ser más rentable. La metodología se enmarcó en la focalización de un segmento de clientes específicos y rentables (clientes con compras en los últimos 12 meses) y la aplicación del modelo de 2 tuplas (quintiles) al cálculo RFM que se afianzó en algoritmos de agrupación en clústeres de k-medias ya que proporciona mayor posibilidad de resultados interpretables y el cálculo de valores sin perjudicar o alterar la información alcanzando mayor precisión. Los resultados obtenidos de su aplicación, permitió efectuar comparación de los centroides de los grupos de datos y visualizar cuan escrupulosa fue la información como para compararlo con el modelo tradicional. Por los datos obtenidos desde el 2014, se clasificó a los clientes sobre la base de 4 categorías desde la A hasta la D, teniendo en cuenta que A es mejores clientes, y D los peores clientes; arribando a conclusiones en las que el modelo de 2 tuplas identificó como peor al 94% y al 86% como de baja introducción en tanto que el método RFM tradicional nos identificó solo al 71% como clientes inactivos y 17 de clientes nuevos. Una de las conclusiones relevantes a las que se arribó es que este método identificó a un 56% de nuevos en tanto que la metodología tradicional solo identifica al 35%, hechos que evidenció que ambos

métodos son de mucha utilidad; sin embargo, la aplicación del modelo RFM respecto a enfoque lingüístico difuso de 2 tuplas es más eficiente que el tradicional, porque permitió identificar agrupaciones más brillantes e indudables de su representación de variables R, F, M.

Del Moral, Chiclana, Tapia, Tapia, & Herrera (2019) en la investigación denominado: A Comparative Analysis Between Two Statistical Deviation–Based Consensus Measures In Group Decision Making Problems, realizado en España, donde el problema fue la toma de decisiones grupales y la metodología usada, para su resolución fue el consenso como estado de acuerdo absoluto o no y medidas de consenso blando que se sustentó en la similitud entre preferencias, cálculo y agregación de magnitudes de distancia que representa la contigüidad preferente ya sea mediante variabilidad estadística que se sostiene en la desviación media absoluta, desviación estándar y en el estudio comparativo que consiste en la comparación de la hipótesis haciendo uso de una prueba estadística y planteándolas como consenso solventadas en índices SDC (Desviación Standard) y MAD (Desviación Absoluta Media), que no producen diferencias significativas respecto a una distancia; dentro de un marco de la generación de 50 problemas GDM aleatorios con 3 expertos, 4 alternativas y bajo el operador OWA para distancias ingresadas en la sección 2 con el operador promedio de distancia (Euclidiana, coseno, y la distancia. Jaccard); debido a que en este tipo de operador los pesos son iguales. Los resultados a los que se arribó reveló que el consenso de los índices SDC y MAD se acercan a los obtenidos por funciones de distancia, en un ámbito de preferencias difusas no obstante en mediciones SDC y MAD son marcadamente diferentes con relación a la propuesta de la aplicación de las funciones de distancia que nos entrega cálculos distintos pero de nivel aceptable con valores cercanos para ciertos casos; siendo que fue precisamente que SDC y MAD que entregaron medición de consenso el cual puede utilizarse en las relaciones de preferencia difusa en casos de problemas respecto a decisiones grupales distintas aplicables las funciones de distancias, resultados que condujeron al establecimiento de clasificaciones de acuerdo a las medidas y a la comparación del problema en distinto estado de consenso en intervalos ordenados de mayor

a menor, concluyendo que a mayor porcentaje es mayor el consenso. Las conclusiones a las que arribó fue que el consenso derivados de índices como SDC y MAD provenientes de la variabilidad estadística aplicando además las funciones de distancia haciendo además uso del operador agregador y realizando el comparativo revelaron que SDC y MAD tienen comportamiento similar a lo que te arrojan las funciones de distancia. Esto pudo también tomarse en cuenta para la resolución, ya que dio una nueva y más sencilla forma de consenso cálculo. Así mismo SDC y MAD fueron diferentes entre que poseen funciones de distancia que usualmente son utilizadas, que proporciona distintos cálculos alternativos. Ergo fueron herramientas alternativas a ser aplicadas en el cálculo de consensos cuando se desee estudiar relaciones de preferencia difusas en caso de toma de decisiones grupales.

LI, Wu & Chen (2019) en la investigación Time is Money: Dynamic-Model-Based Time Series Data-Mining for Correlation Analysis of Commodity Sales, desarrollado en China, cuyo problema fue la resolución de las correlaciones de ventas entre diferentes productos básicos en series de tiempo, cuya metodología usada fue la correlación de ventas con la aplicación del factor tiempo, con la finalidad de descubrir patrones de ventas a fin de afianzar la calidad del marketing cruzado que coadyuve a la toma de decisiones, ésta consistió en la aplicación de un algoritmo de agrupamiento de series de tiempo por afinidad (AP) y K es el clúster más cercano, mediante el cual se integró la clasificación de vecino (KNN), es decir un símil combinado con modelo dinámico. Los resultados alcanzados luego de la puesta en práctica de los algoritmos minería de reglas de asociación (ARM), que descubre los patrones más concurrentes en el registro de transacciones del conjunto de datos, el Vecino más cercano (KNN), algoritmo denominado de agrupación que engloba a un grupo de clúster de AP y los KNN que sirven para evidenciar patrones y conocimiento, algoritmo DM^{w}, en relación con el tiempo y los patrones de ventas mejoran el marketing cruzado calidad; que sean más sensibles en el sentido de que pueden irradiar características específicas de los que conformen el clúster que apoyen en el descubrimiento de tendencias de ventas de productos. Arribando a la conclusión de que haciendo uso de la metodología aplicada y el

uso de modelo dinámico de minería de datos en series de tiempos, se descubre correlaciones en ventas de productos básicos; así mismo se demostró que mediante la ventana de tiempo de observación se determinó cuánto tiempo puede mantenerse la correlación de ventas de productos básicos en tanto que el punto de tiempo de observación nos mostró en qué momento ocurre la correlación, lol cual es importante para el marketing cruzado, considerando que la metodología tradicional nos da información básica.

Jiachen, Yaling & Yuea (2020) en la investigación realizada denominada: Applying clustering and Co-occurrence Methods to Identifying Key Events and Their Relations in Chinese Stock Market, realizado en China, para el problema relacionado a la influencia de las noticias en el impacto inmediato y mensurable causando fluctuaciones de precios en el mercado de valores, cuya metodología se sustentó en el recojo de información de una muy elevada cantidad de textos de informes de notas periodísticas financieras en tiempo determinado en el internet, dicho método se solvento en: 1. Filtrado de la información a través de un procesamiento de textos, 2. Participio chino, que concernía a la segmentación de palabras y 3 eliminación de palabras vacías e etiquetado del mensaje. Posteriormente se efectúo la detección del evento utilizando el VSM y agrupación de clústers; es decir se hizo uso de dos fórmulas: 1. V (d) = {w1, w2, …, Wnorte. Cada elemento del vector en la fórmula representa el número total de palabras cada uno con su peso respectivo. El método para el cálculo se basó en la palabra Frecuencia inversa de palabras de “en documento D, es la frecuencia del documento de” (TF-IDF) con la fórmula 2. wid = tfidentificación * idfid = tfidentificación * log (N / dfD). Los resultados a los que se arribó fueron que se identificó palabras clave donde por ejemplo en el primer ítem en un Clúster signado con el número 2 se identificó como datos de talla 333, datos de rango 4, palabra clave extraída Cotizaciones, acero, licor, papas fritas, en el rubro eventos basados en palabras clave: La existencia de chapa de acero, segmento de licor y concepto de chips tuvieron buen desempeño, en el rubro de eventos resumidos en la red: se identificó Acciones de Guizhou se abrieron paso 700. En el ítem 2 con el clúster signado con el número 4 en datos de talla se identificó 397, en datos de rango 3, en palabras clave extraídas: Corretaje,

activo gestión, investigación, en eventos basados en palabras clave: Nuevas regulaciones para activos administración y en el rubro Eventos resumidos por medios de la red se identificó: Las nuevas regulaciones para la gestión de activos comenzaron a solicitar opiniones de revisión. En la segunda tabla respecto a resultados de cálculo de la correlación de eventos se identificó en el primer ítem de la columna Par de eventos se identificó: Evento 1 y Evento 2, en la columna Número de palabras de co-ocurrencia se identificó 3, en la columna Índice de correlación de eventos A, C 335.7009; en el ítem Nº 2 se identificó: Evento 1 y Evento 3, en la columna Número de palabras de co-ocurrencia 5, en la columna de correlación A, C 332.1145. Lo que demostró que en el evento 2 y 3 tienen la correlación más alta y más significativa. La conclusión de que la metodología funcionó en forma efectiva, para extraer información de eventos respecto a noticias financieras en forma objetiva, útil para las investigaciones de inversores por la mejora de eficiencia en recabar información.

Morente, Ríos, González & Herrera. (2020) en su investigación Using clustering Methods to Deal With High Number of Alternatives on Group Decision Making realizado en España cuyo problema fue la ingente cantidad de datos por procesar, para la toma de decisiones el cual fue solventado haciendo uso de métodos de agrupación y la combinación por similitud y la toma de decisiones grupal. Su metodología propuso el enfoque GDM basados en métodos de agrupamientos E= {e1,. .. , minorte} y X = {x1,. . . ,Xmetro}, P = {pk $_{ij}$}, que genera y da al sistema por los expertos pag$_{kij}$ que refiere cuan experto mi$_{k}$ Preferencia X_i encima X_j posteriormente se realizó la proporción de preferencias, se obtuvo la matriz de preferencias colectivas, Clasificación de alternativas según preferencias y finalmente se realizó el cálculo del consenso debiendo en forma iterativa volver a procesar el GDM. Los resultados obtenidos al seguir la metodología para el cálculo de alternativas temporales fueron: Para la fila GDD, columna X_7 el valor de 0.888; para Columna X_9 0.983; columna X_{12} fue de 0.111; en tanto que para la fila GNDD columna X_7 fue 1; para la Fila GNDD Columna X_9 el valor fue de 0.722; para la final GNDD, columna X_{12} el valor fue de 0.666; para la fila Valor de clasificación se obtuvo en la columna X_7 el valor de 0.944; en la columna X9 0.402; en la columna X_{12} el valor de 0.388. Así mismo en el caso del cálculo

medidas de consenso se obtuvo lo siguiente: X_7 0,55; X_9 1; X_{12} 0.55. La conclusión fue que el nuevo método que abarca entornos con muchos conjuntos de alternativas, puede reducir ingentes cantidades de información y luego de un acuerdo previo clasificar utilizando un procedimiento GDM.

Pan & Zhou (2020) en su investigación Study on convolutional neural network and its application in data mining and sales forescasting for E-commerce, realizado en China y cuyo finalidad fue solucionar un problema predicción del volumen de ventas de productos básicos de comercio electrónico utilizando redes neuronales convolucionales, su metodología estuvo basado en la configuración de un marco de datos solventado en combinación de atributos de las materias primas con los datos de registro originales, que toman al volumen de ventas, precio, cantidad de visitas, visitantes, búsquedas, cantidad de recolectores y otros indicadores por periodos prolongados de lo ya registrado. Posteriormente aplicaron la red neuronal convolucional extrayendo características efectivas del marco de datos y predecir posteriormente las ventas de bienes. Sus resultados arrojaron que si se trabaja una longitud de datos corto la información es insuficiente con un efecto pobre; en tanto que, si el cuadro de datos es demasiado largo, al tener información no válida se obtiene predicción deficiente, demandando incluso adicionales recursos informáticos. Ergo; les resultó pertinente una selección de un cuadro de datos corto pues garantizó mayor precisión en el sentido de que a menor puntuación del error cuadrático medio (MSE) mejor la estabilidad del algoritmo. Es decir, en un cuadro de datos de 40, el MSE en la región 1 era 121, en la región 2 el MSE fue 86, en la región 3 el MSE 139, en la región 4 el MSE fue 311, en la región 5 el MSE 68. Este método se solventó en tomar como datos de entrada los atributos de productos básicos y datos de registros originales. En tanto que como salida en un periodo de tiempo futuro y el volumen de ventas de productos básicos. Es decir, forma una caja de datos mediante combinación de información de atributos del producto en periodos de tiempo y mediante red neuronal convolucional se extrajo características efectivas. Finalmente concluyen que la red neuronal se usa para obtener características efectivas de los datos en tiempos estructurados utilizando el método en la predicción de ventas

evidenciando que el algoritmo que propusieron, mejoró efectivamente la precisión de la previsión de ventas.

Afifi (2020) en su estudio titulado Demand Forecasting of Short Life Cycle Products Using Data Mining Techniques realizado en Egipto cuya problemática concernía al pronóstico de la demanda de productos con ciclo de vida corto y su metodología se amparaba en el k-algoritmo de agrupamiento de medios y la clasificación de inducción de reglas en el objeto de pronosticar la demanda de productos en ciclos de vida corto donde utilizaron técnicas alternativas de data mining, con el objeto de mejorar el pronóstico de la demanda a una empresa minorista, mediante la técnica de agrupamiento incremento*k* y haciendo uso del algoritmo, aprendizaje de reglas de clasificación (RULES-6); obteniendo como resultado un cuadro de errores de pronóstico de los perfiles de ventas donde destaca el método propuesto en el presente estudio siendo que en se obtuvo en Raridad un MAE de 70.7, un MAPA% de 5.2, un RMSE de 84.3; en Bayesiana ingenua se obtuvo en MAE 63.2, en MAPA 4.5, en RMSE 78.9;K-vecinos más cercanos se obtuvo en MAE 51.3, en MAPA 3.7, en RMSE 66.8; REGLAS-6 Propuestas se obtuvo en MAE 5.9, en MAPA 1.8 y en RMSE 7.3 evaluando que el método propuesto fue efectivo en descubrir agrupaciones sugestivas de datos históricos. Así mismo que por el agrupamiento y sus atributos la predicción en nuevos productos se efectuó fácilmente obteniendo resaltantes pronósticos precisos. Finalmente concluyeron que el uso combinado del método incremental K-medios y las REGLAS-6 fueron válidos para estimar perfiles de ventas de productos nuevos en empresas minoristas que no poseen datos históricos de ventas.

Wisesa, Adriansyah, & Khalaf (2020) en su estudio titulado Prediction analysis sales for corporate services telecommunications company using gradient boost algorithm, realizado en Indonesia cuyo problema fue como determinar el tipo de análisis de predicción al comparar 4 técnicas de aprendizaje automático con relación a las ventas B2B de los años 2016 al 2018. Su metodología se basó en la aplicación de algoritmos apoyado en el trabajo previo consistente en los procesos de minería de datos como el planeamiento, simulación, evaluación e

implementación, calculó y evaluó el MAD, MSE, RMSE Y MAPE, donde además aplicó un nuevo árbol al modelo F(x) completo después de la ronda T-1 y H(x) resultando un árbol potenciado $MAPE = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$ considerando At representa al valor actual, Ft valor esperado y la brecha existente entre ambos se divide en At. Donde prevalece como valor absoluto y se calcula y divide con n. Sus resultados resaltan en la precisión, exactitud en cada clase y la matriz de confusión que precisa el número de predicciones que miden los errores: cuadrático medio, el absoluto y el error promedio; siendo que los mismos alcanzaron MSE en la columna Manual el valor de 122,883,547,626,822,000,000,000.00 y en GBT 880,640,000,000,000,000,000.00 en tanto que RMSE en la comuna de manual alcanzo el valor de 350,547,496,962.71. en tanto que en la columna de GBT alcanzó 157,299.08; y en MAPE en la columna de manual alcanzó el valor de 350,547,496,962.71 y en GBT el valor de 0.18; concluyendo que el resultado de este estudio se determina a través de la comparación de enfoques de aprendizaje automático, donde prevalece la confiabilidad y la precisión basado en los resultados de RMSE, MSE y MAPE con las mismas escalas.

Mardiantien, Atastina & Asror (2020) en su estudio denominado Product segmentation based on sales transaction data using agglomerative hierarchical clustering and FMC model (Case Study: XYZ Company), realizado en Indonesia cuyo problema fue evaluar la ingentes cantidad de datos de transacciones de una determinada empresa con análisis conglomerados aplicando el modelo FMC la metodología Agglomerative Hierarchical Clustering (AHC), donde se buscaba la mejor calidad de los mismos. Su metodología se sostuvo en evaluar técnicas de minería de datos cuyo objetivo fue agrupar productos por periodos de tiempo incluso, usando el método de vinculación de ward´s, a través de las características de frecuencia(F), Monetario (M) y Variedad de Clientes; después de haber explorado y extraído los datos, procesamiento de los datos y agrupamiento de los mismos y determinación de los clústeres así como la normalización de los datos en la que se utilizó la normalización min-max con valores de rango [0, 1]. Los resultados que se obtuvieron de extraer las características en el modelo FMC con 5 muestras de productos fueron los

siguientes para el producto de código PC0383 su frecuencia fue de 8052, El importe monetario fue de 8,527,435,620 y la variedad de clientes fue de 4856; código de producto PC0027 su frecuencia fue de 4064, El importe monetario fue de 12,407,631,722 y la variedad de clientes fue de 2469; código de producto PC0298 su frecuencia fue de 2189, El importe monetario fue de 1,602,328,945 y la variedad de clientes fue de 1727; código de producto PC0295 su frecuencia fue de 2123, El importe monetario fue de 3,327,482,015 y la variedad de clientes fue de 1402 y para el código de producto PC0391 su frecuencia fue de 4456, El importe monetario fue de 2,010,245,825 y la variedad de clientes fue de 2971; situación que los llevó a determinar que existen productos con mayor auge en ventas por año y determinados clientes que pueden incluso repetir en su compra de determinado producto y que ha contribuido notablemente en el fortalecimiento de negocio determinado. Así mismo obtuvieron resultados con clústeres con características sobresalientes al haberse combinado con todas las características FMC. Estos resultados sufrieron una variabilidad de acuerdo al periodo de año, tipo de producto relacionado con las características FMC. La conclusión a la que arribaron fue en el sentido de que este conocimiento puede otorgar información a la empresa para determinar fehacientemente que productos requieren mayor atención por asegurar mayor rentabilidad lo que además puede determinar en estrategia de marketing.

Alessandro, Luciano; Palesi, Lusiano Alessandro Ipsaro; Nesi, Paolo & Pantaleo, Gianni (2022) en su investigación Multi Clustering Recommendation System for Fashion Retail, realizado en Italia que concierne a la resolución de problemas de comercio en tiendas minoristas donde existe escasez de datos ante regulaciones que genera escasez de datos y estacionalidad de artículos de venta (Vida comercial entre 6 meses o 1 año); encontrando asidero esta solución en las técnicas de minería de datos cuya metodología aplicada se sostuvo en aplicar técnicas de data mining, clasificando según sus categorías de acuerdo a sus fuentes como demográficos, contenidos, filtros colaborativos, comunitarios e híbridos y conocimiento donde prevaleció el algoritmo de clasificación a través de la aplicación del algoritmo de Vecino más cercano, árboles de decisión, k-means, redes neuronales; Reglas de asociación; cuyos

resultados optimizan tiempos de compras de los clientes, mejor nivel de precisión que redundan en una mejor toma de decisiones. Este estudio concluye que su propuesta generó un aumento de compra del 3.48%, ser soluciones funcionales, permitiendo incluso predicciones de ventas y comportamiento de clientes.

1.3. Teorías relacionadas al tema.

1.3.1. Data Mining

Consiste en cernir datos y comprimir registros de datos con el objeto de revelar información relevante a través de una representación que facilite la comprensión al usuario. Palma, J. & Marín,R. (2008) p.461. Así mismo el data mining o procesamiento de datos, se compone de cuatro partes como el agrupamiento, la clasificación, minería de patrones de asociación y análisis de valores atípicos, mediante el cual se obtiene datos útiles, conocimiento de estos datos que actualmente han alcanzado el orden de petabytes o exabytes como por ejemplo los datos de la Web como documentos entre otros, transacciones financieras, interacciones del usuario y tecnología de sensores e internet de las cosas. (Aggarwal, 2015). p.1-3.

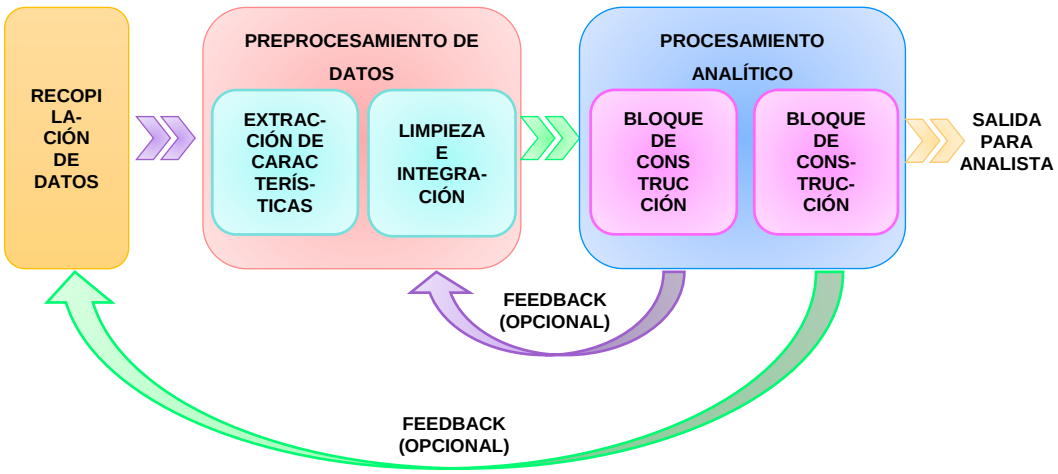


Figura 3. Representación del Data Mining. Fuente: Aggarwal, C. *(2015)*. Data Mining.

1.3.2. Datos etiquetados y no etiquetados

Se denominan datos etiquetados a aquellos datos que tienen el atributo designado donde el objetivo es hacer uso de esos datos para predecir el valor de ese atributo en instancias no identificadas; que en minería de datos se usa para aprendizaje no supervisado y dependiendo de que si el atributo es o no categórico la actividad se le denominaría clasificación. En tanto que si el atributo elegido es numérico la actividad estaría dentro del marco de lo que es regresión.

Los datos que no poseen atributo alguno son aquellos que se denominan sin etiquetar y su extracción es bien conocido como datos no etiquetados conocidos como aprendizaje no supervisado, donde el objetivo es extraer mayor cuantía de información. (Bramer, 2016). p. 4.

1.3.3. Aprendizaje supervisado: clasificación.

Se conceptualiza así cualquier técnica de aprendizaje automático en la que, a través de algoritmo, aprenden del entrenamiento de etiquetas y se orienta a resultados correctos haciendo uso de recursos externos (Kopec, 2019) p.187. La clasificación es la más común que se usa en la cotidianidad de la minería de datos, para clasificaciones las mismas que se aplican sobre la base de métodos y reglas. . (Bramer, 2016). p.7.

1.3.4. Aprendizaje supervisado: predicción numérica

Consiste en la clasificación como un tipo de predicción, donde el valor que se vaticinaba es la etiqueta. A este tipo de predicción además se le denomina regresión que puede ser utilizado, para predecir un valor numérico y que es precisamente muy usado a nivel empresarial mediante la aplicación de redes neuronales complejas. . (Bramer, 2016). p.5 y 6.

1.3.5. Aprendizaje no supervisado: reglas de asociación

Se denomina aprendizaje no supervisado cuando el proceso carece de base de datos previa o especial para "enseñar" al modelo con relación a la noción de agrupación apropiada. (Aggarwal, 2015). p.19; y muy bien podría

además definirse a cualquier técnica que se ejecuta por si sola sin conocimiento previo, para arribar a conclusiones. (Kopec, 2019) p.190. Se refiere precisamente a la acción efectuada por los algoritmos que aprenden, haciendo uso del conjunto de datos de entrenamiento en la que se desea encontrar alguna relación que exista entre los valores de variables aplicando criterios estadísticos, geométricos o de similitud y aplicando distintas asociaciones de reglas como agrupación, reducción de dimensionalidad, detección de valores atípicos, detección de novedades (Igual & Seguí,, 2017). p.67 y 115.

1.3.6. Neural Network

Es el símil al sistema nervioso humano que está conformado por nodos a los que se les denomina neuronas, se interconectan entre sí, las que son unidades de cálculo que recepcionan información de otras neuronas a las que la cataloga como entrada (cálculo definida por pesos en las conexiones), las vuelve a calcular y generando nueva información y a vez alimentan otras neuronas. Estos cálculos se basan en pesos que vienen a ser los datos de entrenamiento que pueden ser modificadas de acuerdo a la necesidad o al detectar predicciones incorrectas. Su eficacia depende de la arquitectura usada existiendo gran variedad (simple perceptrón de una capa hasta complejas de múltiples capas), para efectuar las conexiones entre nodos. (Bell, 2015).

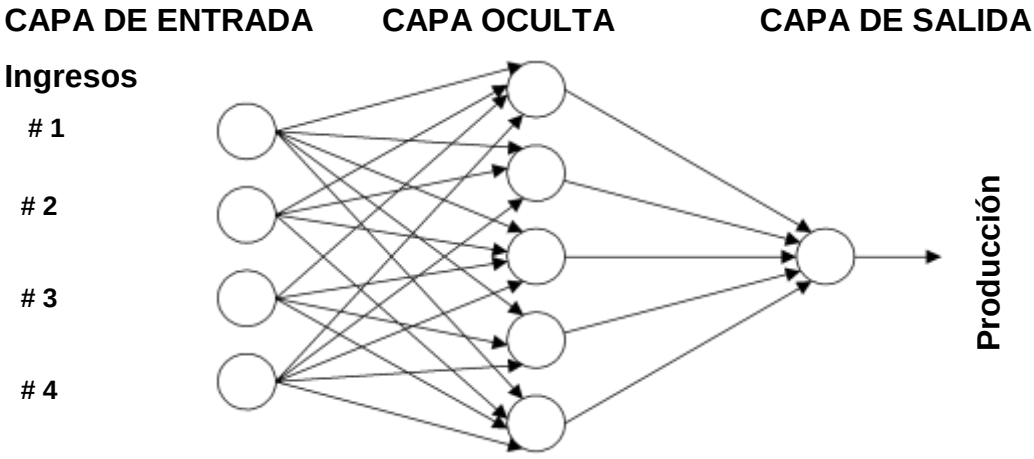


Figura 4. Representación de la Red Neuronal. Fuente: *(Bramer, 2016)*. p.7.

1.3.7. Datos para minería de datos

Estos datos están conformados por archivos electrónicos alimentado por usuarios humanos, información comercial en SQL o cualquier formato de base de datos regular, información automática y electrónicamente registrada o flujos de datos binarios recibos del satélite.

1.3.8. Formulación estándar

Se tiene entendido que se tiene un universo de objetos que bien contienen variables y que éstas a su vez tienen la denominación de atributos; siendo que al conjunto de valores de variables se le denomina registro o instancia. El conjunto de datos es representado por una tabla, donde cada fila representa una instancia y cada columna el valor de una variable o atributo. (Bramer, 2016). p.9.

1.3.9. Tipos de variable.

Los tipos de variables según Bramer, M. (2016) p.10. Son las siguientes:

Variable Nominal: Se usa es para asignar objetos una categoría que puede ser numérico.

Variables Binarias: Éstas vienen a ser caso especial de la anterior que puede tomar 2 valores siendo estos verdadero o falso o también 1 o 0.

Variables Ordinales: Parecidas a las variables nominales que se forman bajo un orden significativo.

Variables Enteras: Son los valores enteros

Variables de escala de intervalo: Son aquellas que toman valores numéricos, medibles a intervalos iguales desde un punto determinado que puede ser el cero o el origen.

Variables de escala de razón: Es un símil a la variable precedente siendo que el cero expresa carencia del rasgo medido.

1.3.10. Teoría de la clasificación

Comprende la división de un objeto la misma que para que a su vez se le fije a una de varias categorías sean estas exhaustivas y exclusivas (cada objeto

a una clase no a más de una; tampoco a ninguna clase), a las que además se le denomina clases. (Bramer, 2016) p.21

1.3.11. Clasificación de Bayes

Proviene del reverendo Thomas Bayes (1702 – 1761), primer matemático en usar la probabilidad en forma inductiva. Esta clasificación concierne a la probabilidad de un evento que por lo general, puede además ser un conjunto alternativo de eventos posibles; excluyentes y exhaustivos, teniendo como la posibilidad de que ocurre un solo evento, donde la suma de la probabilidad de un conjunto de eventos a los que nos referimos en el presente párrafo, debe ser siempre 1. (Bramer, 2016). p.22. En esta misma línea. (Aggarwal, 2015). p. 306., afirma que este tipo de clasificación se utiliza, para que al conjunto de variables se le pueda modelar la probabilidad de cada uno de sus valores de variables.

1.3.12. Clasificación del vecino más cercano.

Este tipo de clasificación es modelada cuando los valores de los atributos son continuos, no obstante, se puede efectuar cambios para relacionar atributos categóricos. Se permite la clasificación de instancia más cercana de un registro invisible en el modo que se requiera definir. Bramer, M. (2016) p. 29. Éste clasificador es válido para ser usado con cualquier tipología de datos si una función de distancia esté disponible. El enfoque en el que se desarrolla, se solventa en el caso de datos multidimensionales, donde se debe de determinar sus k vecinos más cercanos en datos del entrenamiento, debiéndose informar como label de clase relevante; donde el valor k puede resolverse mediante validación cruzada de dejar uno fuera. (Aggarwal, 2015). p. 332.

1.3.12.1. Medidas de distancia.

Existen infinidad formas de medirla a dos instancias de n valores de atributos en un espacio n-dimensional. Como la notación d (X,Y).

a. Distancia de sí mismo (A, A) = 0
b. Distancia (A, B) = d (B, A)

c. Condición de desigualdad triangular (Geometría Euclidiana - distancia corta entre puntos). La condición dice que para los puntos A, B y Z: d (A, B) ≤ d (A, Z) + d (Z, B).

La distancia euclidiana mide dos instancias en forma natural.

1.3.12.2. Normalización.

Sustenta la aplicación de la fórmula de la distancia euclidiana u otras medidas cuando lo valores son grandes que opacan a los pequeños siendo viable que una instancia invisible tenga valor de A menor que mínimo o también podría ser mayor que máximo, para cuando se desee establecer los números en forma ajustada. (Bramer, 2016). p.35.

1.3.12.3. Manejo de atributos categóricos.

La debilidad del enfoque vecino más cercano, es que no hay forma exitosa de tratar los atributos categóricos. Probablemente el modo de decir que la diferencia entre valores semejantes del atributo es cero y la diferencia entre dos valores diferentes es uno. Lo que podría afirmarse que para un atributo de color rojo – rojo = 0; en tanto que rojo – azul = 1 ó azul – verde = 1. (Bramer, 2016). p.36.

1.3.13. Uso de árboles de decisión para la clasificación.

Esta representación de datos tiene la ventaja comparativa respecto a otras por su grado de significancia y facilidad de interpretación se utilizan como el mecanismo para generar reglas de clasificación gracias al algoritmo simple TDIDT. Los árboles de decisión se utilizan ampliamente como medio para generar reglas de clasificación debido a la existencia de un algoritmo simple Bramer, M. (2016). p.39 y 45.

1.3.13.1. Reglas de decisión y árboles de decisión

Existen grandes cantidades de ejemplos en diferentes campos de estudio y que bien podría utilizarse, para la aplicación de reglas de decisión bajo la tutela del enfoque estándar con la finalidad de obtener reglas de los expertos, que puede en muchos más no en todos los casos efectuarse

combinaciones favorables para la formación de una estructura de árbol. (Bramer, 2016). p.39.

1.3.14. TDIDT

Este algoritmo tiende a inducir árboles de decisión de arriba – abajo y desde 1960 forma parte relevante de los sistemas de clasificación. Los más conocidos y utilizados en minería de datos son ID3 [3] y C4.5 [2], ya que origina reglas de decisión dividiendo valores de los atributos en forma reiterativa y que denominan particionamiento recursivo. Se solventa sobre la base del conjunto de datos de entrenamiento, donde cada dato corresponde a un segmento de objetos y que su descripción son los valores del conjunto de tributos categóricos.

1.3.15. Reglas de decisión y árboles de decisión.

Técnica realista de clasificación o reglas de decisión del enfoque estándar del sistema experto que puede combinarse de forma más conveniente, para obtención de reglas que permite desarrollar sistemas expertos convencionales MYCIN y XCON. (Bramer, 2016). p.39

1.3.16. El conjunto de datos de grados.

Consiste en la serie de ramas, donde cada uno termina en un nodo con una clasificación válida. Así mismo cada rama sostiene una ruta desde la raíz a hasta el nodo hoja. El nodo raíz atañe al conjunto original. En tanto que los demás nodos pertenecen al subconjunto del entrenamiento. El caso de los nodos hoja cada instancia o registro del subconjunto pertenecen a la misma clasificación. Si existe un número n determinado de nodos, es la misma cantidad de ramas que debe corresponderle donde cada rama corresponde a una regla de clasificación.

1.3.17. K-means

Técnica computacional que sirve para dividir puntos del conjunto de datos en grupos, que se relacionan entre sí y que requieren verificación humana para casos donde las relaciones son significativas. Cada agrupación de un

punto de datos no está predeterminado, debido a que se consolida ese resultado a través del algoritmo absolutamente sin intervención adicional de ninguna forma. Situación que lo hacen denominarse en el aprendizaje automático como método no supervisado. El algoritmo intenta agrupar puntos de datos en las agrupaciones de datos dependiendo de la distancia relativa de cada punto al centroide; situación que se repite en forma iterativa hasta que los centroides dejen de moverse, siguiendo la forma de evaluación de datos en relación con su puntuación standard dependiendo de los demás valores del mismo tipo. (Kopec, 2019). p. 112 – 115. Esta representación evidencia de que el algoritmo K-means agrupa en forma iterativa simple dividiendo N puntos de datos en K subconjuntos disjuntos S_j con lo cual disminuye la configuración de suma de cuadrados, que viene a ser la distancia euclidiana al cuadrado y ésta a su vez la media más cercana. En ese sentido se puede afirmar fehacientemente que su algoritmo de aprendizaje no supervisado de cuantificación vectorial define k centroides, los que participaran como prototipo en sus grupos pertinentes, donde cada objeto es agrupado a un grupo determinado considerando el centroide más cercano si se mide con la métrica determinada de distancia. Este procedimiento concluye cuando todos los objetos fueron asignados, se reitera el proceso recalculando los centroides, hasta que no haya algún movimiento de centroides de algún grupo. Este algoritmo basa su consistencia en el proceso de asignación de puntos de datos al grupo y actualización, donde cada centroide de grupo se calcula en el centro de todos los datos, alternándose hasta que se detenga y no haya más alteraciones respecto a la configuración de puntos de datos. (Award & Khana, 2015). p.10, 27.

$$J= \sum_{j=1}^{k} \quad \sum_{n \in S_j} \left| X_n - \mu_j \right|^2$$

x_n = vector que representa el n-ésimo punto de datos

μ_j = centroide geométrico de los puntos de datos en S_j

1.3.18. Clustering de k-means

La clasificación demanda clases de etiquetas establecida por personas, que motiva a separar datos en grupos de características símiles, donde cada grupo puede ser una clase, teniendo en cuenta que los clústeres sirven para análisis de datos exploratorios a modo de aprendizaje no supervisado en vez de análisis para realizar predicciones específicas. Este tipo de agrupación K-means, es la más adecuada, flexible y eficiente, para el análisis de datos en forma experimental, cuando se configura los patrones, como datos sin clases o etiquetas y grupos generados que podrían ser usados para la agrupación doce se relacionen de acuerdo a sus símiles. (Dinov, 2018).

1.4. Formulación del Problema.

¿Cuál es la técnica de minería de datos más eficiente, para descubrir información relevante de ventas de una MYPE comercial?

1.5. Justificación e importancia del estudio.

MYPE´s en Perú conforman el conglomerado del 95% de empresas que le proporcionan empleo al 47.7% de la PEA, llegando a representar el nivel porcentual de 19,3% del PBI. El crecimiento al año 2019 con relación al 2018 asciende al 6% y su informalidad es de 83.3%, el mismo año se toma de referencia. (Ministerio de la Producción, 2020). p.12-29.

Razones que motivan el presente estudio que se solventa sobre la comparación de técnicas que permitan elegir la que mejor se podría alinear con alguna estrategia de negocio, para el descubrimiento de información relevante de ventas. Así mismo en el contexto actual de pandemia, el regreso a la nueva normalidad en pandemia demandó de tecnología que este orientada a las MYPE’s, que generan más empleo, por lo que la pertinencia de la aplicación de la metodología data mining, impulsó la economía, la producción y generación del empleo articulando las TIC´s con la dinamicidad del conocimiento científico data mining, que apuntaló el redireccionamiento de la producción, mejora del marketing e incremento de las ventas.

1.6. Hipótesis.

La Técnica Data Mining de clusterización es la más eficiente para descubrir información relevante de ventas.

1.7. Objetivos.

1.7.1. Objetivo general.

Comparar técnicas de minería de datos, para descubrir información relevante de ventas de una MYPE comercial.

1.7.2. Objetivos específicos.

a) Seleccionar una MYPE comercial como caso de estudio para establecer los requisitos de información.
b) Preparar el conjunto de datos desde las bases de datos disponibles en caso de estudio.
c) Seleccionar técnicas de data mining con mejor desempeño en el procesamiento de ventas.
d) Aplicar técnicas de minería en la base de datos del caso de estudio previamente seleccionado.
e) Evaluar el desempeño de las técnicas de minería de datos aplicadas.

II. MATERIAL Y MÉTODO

2.1. Tipo y Diseño de Investigación.

Tipo

Se sostuvo en metodología cuantitativa que, haciendo uso de técnicas computacionales, estadísticas y matemáticas se demostró resultados de la presente investigación.

Diseño

El diseño es cuasiexperimental en la que el investigador a través de la manipulación de las variables (dependiente e independiente), mediante cambios de valor de variable independiente, se obtuvo como efecto o consecuencia otro valor en la variable dependiente. (Hernández, Fernández, & Baptista, 2014). p. 184.

2.2. Población y muestra.

Población.

Objeto del estudio e investigación que está compuesta por 15 técnicas de minería de datos que forman parte del análisis del problema y se detallan en el Anexo 4.

Muestra

Con relación al muestreo y considerando que es un muestreo no probabilístico se decidió establecer un muestreo no estadístico por conveniencia debido a las ingentes técnicas existentes de minería de datos, por lo que se tomó en cuenta aquellas que se encuentran más ligadas al ámbito de las MYPE's, precisamente porque en Perú aproximadamente el 95% de empresas son MYPE's (Ministerio de la Producción, 2020). p.12-29; y hay evidencia que para permanecer en el mercado, acrecentar el negocio o expandir se deben de afianzar en la tecnología que les redunde beneficios en esas líneas intrínsecas e inherentes a los objetivos propios de las MYPE´s (Mejía, Tuesta, Samillan, & Forero, 2021).

Finalmente, las técnicas que se alinean en el sentido de las necesidades de la PYMES conciernen a:

Regresión Logística.

Regresión lineal simple

Super Vector Machine

Árboles de decisión

Neural Network.

Nearest Neighbors

K-mean

Clustering.

Razones que motivaron que el presente estudio se solvente sobre la comparación de técnicas que permitan elegir la que mejor se podría alinear a la estrategia del negocio en relación al descubrimiento de información relevante de ventas; donde se consideró el criterio de clientes y ventas.

2.3. Variables, Operacionalización.

Variables	Dimensión	Indicador	Ítem	Técnica e instrumentos de recolección de datos
Técnica de Minería de Datos	Consumo de recursos	a. Promedio de tiempo de respuesta	$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$	Registro electrónico
		b. Grado de consumo de CPU	$Uc = \sum_{j}^{n} uc_j / n$	
		c. Grado de consumo de memoria (RAM)	$Cm = \sum_{j}^{n} cm_j / n$	
Información relevante de ventas de una MYPE.	Exactitud	a. Error cuadrático medio	$ECM = \frac{\sum_{t=1}^{n} (F_t - Y_t)^2}{n}$	
		b. Precisión	$P = \frac{VP}{FP + VP}$	
		c. Desviación absoluta media (MAD)	$MAD = \frac{\sum_{t=1}^{n} \lvert F_t - Y_t \rvert}{\mathrm{n}}$	

2.4. Técnicas e instrumentos de recolección de datos, validez y confiabilidad.

Se previó un método de instrumento mecánico electrónico, para la extracción de los datos relevantes de ventas; que fue trasladado a contenedores de datos, para ser preparados y luego ser procesados efectuando el uso del registro electrónico que se solventó sobre la base del registro automático de los resultados de los procedimientos y entrenamiento que se previeron implementar, para las mediciones. El registro de los datos se efectuó en los formatos de registro de matriz de confusión que obra en Anexo Nº 3, donde se registró resultados reales de clasificación, de consultas, de mediciones y rendimientos detallándose los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Así mismo se elaboró el registro de medidas de exactitud, precisión, velocidad para considerar el consumo de recursos que forman parte del Anexo Nº 3.

Los instrumentos electrónicos además se solventaron bajo la consistencia del software mediante el cual se extrajo los datos, para la construcción de un data set que fue procesado a través de la aplicación de librerías existentes contenidas en el Software denominado Anaconda Navigator, Python y jupyter para el modelado que nos entregó resultados para el análisis e interpretación de los mismos.

2.5. Procedimiento de análisis de datos.

Este procesamiento implicó extraer datos haciendo usos de la técnica mecánico o electrónico (extracción), efectuar el registro electrónico (previa aplicar la limpieza y modelado de datos), ejecutando el código computacional, logrando descubrir la información relevante que es de utilidad y de alto interés para la MYPE. El total de datos obtenidos sufrió alteración o transformación en información que pudo ser utilizada para el análisis y conclusiones.

La variable técnica de minería de datos estuvo dimensionada por el consumo de recursos lo cual tiene como indicadores de medición, promedio de tiempo de respuesta, consumo de memoria y consumo de CPU; que se solventarán en las fórmulas matemáticas siguientes:

a. Promedio de tiempo de respuestas:

Es el valor en tiempo en que la técnica de minería de datos haciendo uso del indicador a través de la fórmula estadística y aplicando machine learning se obtuvo cuando se ejecutó la clasificación.

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Donde:

Tr: Es el tiempo de respuesta de la técnica de minería de datos

Tf_j: Es el tiempo final en que concluye el proceso de los datos

Tf_i: Es el tiempo inicial de respuesta en que se ejecuta el procesamiento de los datos

n: Es el total de pruebas

b. Grado de consumo de CPU:

Representó el valor numérico de consumo de recursos de CPU que la técnica de minería de datos utiliza del computador, para procesar los datos de la MYPE en cuanto a las necesidades de información relevante, para las operaciones de la misma. Este cálculo se realizó de la siguiente manera:

$$Uc = \sum_{j}^{n} uc_j \, / n$$

Donde:

Uc: uso del CPU cuando la técnica de minería se estuvo ejecutando

n: número de ejecuciones ocurridas durante el procesamiento de los datos

c. Grado de consumo de memoria (RAM):

Representó el valor numérico de consumo de memoria RAM que la técnica de minería de datos utiliza del computador, para procesar los datos de la MYPE en

cuanto a las necesidades de información relevante, para las operaciones de la misma. Este cálculo se realizó de la siguiente manera:

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Donde:

Cm: Grado de consumo de memoria en que se procesó los datos

Cm_j: Grado de consumo de memoria en la prueba *j* del proceso de los datos

n: Es el total de pruebas

La variable Información relevante de ventas de una MYPE está dimensionada por la exactitud y tiene como indicadores al error cuadrático medio, Precisión y Desviación absoluta media (MAD), que se resolverán con las fórmulas matemáticas siguientes:

a. Error cuadrático medio:

Esta medida absoluta de ajuste se usó para efectuar análisis comparativo acerca de la precisión entre métodos de pronósticos, que permitió seleccionar el que proporciona menor ECM.

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

Donde:

ECM = Error cuadrático medio.

Sumatoria de valores (F_t precisan las observaciones presentes o actuales de las series de tiempo - Y_t . Es decir, configuraron las series de tiempo pronosticadas) al cuadrado c x n = Cantidad de observaciones.

b. Precisión: Es la concordancia de los resultados entre si obtenidos por algoritmo, en esta se buscó encontrar la proporción de la predicción de ventas que se acercó a la realidad obtenida o realizada. La ecuación es la siguiente:

$$P = \frac{VP}{FP + VP}$$

Donde:

P : Es la Precisión

VP: Es verdadero positivo

FP: Es falso positivo

c. Desviación absoluta media (MAD)

Estableció la precisión de pronósticos realizados efectuando la media y luego la diferencia de la desviación con relación al pronóstico, para elegir el resultado de menor valor.

$$MAD = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

Donde

MAD = Desviación media absoluta

b. Sumatoria de n valores absolutos (F_t = número de observaciones presentes de las series de tiempo – Y_t que viene a ser las series de tiempo pronosticadas)

c x n = Cifra de observaciones

2.6. Criterios éticos.

Confidencialidad: Se garantizó que los datos e información de las personas y de la empresa se mantendrán en anonimato en el presente estudio.

Derechos de Autor: La investigación se desarrolló dentro del marco ético y respeto estricto de los derechos de autor, para lo cual se utilizará el sistema APA, reconociendo y citando a los autores de libros, artículos científicos publicadas en revistas, tesis y otras que contribuyeron a fortalecer la presente investigación.

Búsqueda del Bien: La investigación promueve el bien de la comunidad en general beneficiando a MYPE’s e Instituciones en general que son los promotores del empleo en porcentajes altos de la PEA de países emergentes., beneficiando a ciudadanos.

2.7. Criterios de Rigor Científico.

Objetividad: El presente estudio de investigativo, se solventó sobre criterios técnicos, legales y morales proyectando imparcialidad durante el desarrollo y materialización de la misma.

Confiabilidad: Las herramientas tecnológicas, metodologías científicas y el uso de instrumentos mecánico electrónico, para la recopilación de los datos y procesamiento de los mismos contribuye al rigor científico.

Validez: La presente investigación tiene validez lógica al tener sostenibilidad en modelos matemáticos; así como estándares establecidos en una ecuación

III. RESULTADOS.

Para realizar la comparación de técnicas de minería de datos y decidir cuál es la más eficiente que permita descubrir información relevante de ventas de una MYPE se aplicó en una computadora de escritorio con las características siguientes: Procesador Intel (R) Core(TM) i7-9700k CPU @ 3.60HGz 3.60 GHz, con RAM de 16.0 GB, Sistema operativo y procesador de 64 bits.

El total de registros que se han procesado es 5,522 a la que se le aplicó la técnica de minería de datos de clasificación destinando de la base de datos un 80% para el entrenamiento (training) y un 20% para pruebas (testing), haciendo uso de algoritmos respectivos y que son mencionados en el presente acápite, donde se plasma los valores por cada indicador como son: Promedio de tiempo de respuesta, consumo de memoria (RAM), Consumo de CPU, error cuadrático medio, precisión y desviación absoluta media (MAD) de los cuales se estima cual es más eficiente para descubrir información relevante de una MYPE, para la cual se desarrolló modelos como regresión logística, regresión lineal, super vector machine y árbol de decisiones.

Consumo de Tiempo de Respuesta

a. Promedio de Tiempo Respuesta regresión logística Regresión Logística.

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 0.0620 s

b. Promedio de Tiempo Respuesta Regresión Lineal

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 0.0067 s

c. Promedio de tiempo respuesta super vector machine

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 48.8991 s

d. Promedio de tiempo respuesta árbol de decisiones

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 0.1921 s

e. Promedio de tiempo respuesta Redes Neuronales 3 Capas

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 0.8907 s

f. Promedio de tiempo respuesta Bayes

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 0.0034 s

g. Promedio de tiempo respuesta Nearest Neighbors

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 0.0105 s

h. Promedio de tiempo respuesta K-means

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 0.0638 s

i. Promedio de tiempo respuesta Spectral Clustering

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

Tr = 1.8361 s

Grado de Consumo de CPU

a. Grado de Consumo de CPU Regresión Logística

$$Uc = \sum_{j}^{n} uc_j / n$$

Uc = 4.6 Gbs

b. Grado de Consumo de CPU de Regresión Lineal

$$Uc = \sum_{j}^{n} uc_j / n$$

Uc = 84.7

c. Grado de Consumo de CPU super vector machine

$$Uc = \sum_{j}^{n} uc_j / n$$

Uc = 13.8

d. Grado de Consumo de CPU árbol de decisiones

$$Uc = \sum_{j}^{n} uc_j \, / n$$

Uc = 3.9

e. Grado de Consumo de CPU Redes Neuronales 3 Capas

$$Uc = \sum_{j}^{n} uc_j \, / n$$

Uc = 98.8

f. Grado de Consumo de CPU Bayes

$$Uc = \sum_{j}^{n} uc_j \, / n$$

Uc = 4.2

g. Grado de Consumo de CPU Nearest Neighbors

$$Uc = \sum_{j}^{n} uc_j \, / n$$

Uc = 21.0

h. Grado de Consumo de CPU K-means

$$Uc = \sum_{j}^{n} uc_j \, / n$$

Uc = 84.5

i. Grado de Consumo de CPU Spectral Clustering

$$Uc = \sum_{j}^{n} uc_j \, /n$$

Uc = 59.3

Grado de consumo de memoria

a. Consumo de memoria Regresión Logística

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 6.13 s

b. Consumo de memoria de Regresión Lineal

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 6.3978

c. Consumo de memoria super vector machine

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 6.6089

d. Grado de consumo de memoria árbol de decisiones

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 6.5531

e. Grado de Consumo de memoria redes neuronales 3 capas

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 5.4714

f. Grado de consumo de memoria Bayes

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 7.1034

g. Grado de consumo de nearest neighbors

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 6.9165

h. Grado de consumo de K-means

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 6.4100

i. Grado de consumo de Spectral Clustering

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Cm = 6.2123

Error cuadrático medio

a. Error cuadrático medio Regresión Logística

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 0.00090

b. Error cuadrático medio de Regresión Lineal

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 6.3978

c. Error cuadrático medio super vector machine

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 0.00090

d. Error cuadrático medio árbol de decisiones

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 0.00090

e. Error cuadrático medio redes neuronales 3 capas

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 0.18778

f. Error cuadrático medio Bayes

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 0.17071

g. Error cuadrático Nearest Neighbors

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 0.22013

h. Error cuadrático K-means

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 0.45822

i. Error cuadrático Spectral Clustering

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

ECM = 0.60916

Precisión

a. Precisión Regresión Logística

$$P = \frac{VP}{FP + VP}$$

P = 99.93

b. Precisión de Regresión Lineal

$$P = \frac{VP}{FP + VP}$$

P = 97.60

c. Precisión super vector machine

$$P = \frac{VP}{FP + VP}$$

P = 99.91

d. Precisión árbol de decisiones

$$P = \frac{VP}{FP + VP}$$

P = 99.91

e. Precisión redes neuronales 3 capas

$$P = \frac{VP}{FP + VP}$$

P = 81.22

f. Precisión redes neuronales Bayes

$$P = \frac{VP}{FP + VP}$$

P = 98.29

g. Precisión redes neuronales Bayes

$$P = \frac{VP}{FP + VP}$$

P = 77.99

h. Precisión K-means

$$P = \frac{VP}{FP + VP}$$

P = 54.18

i. Precisión Spectral Clustering

$$P = \frac{VP}{FP + VP}$$

P = 39.08

Desviación Absoluta media

a. Desviación absoluta media (MAD) Regresión Logística

$$\text{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\text{n}}$$

MAD = 0.000898

b. Desviación absoluta media (MAD) Regresión Lineal

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

MAD = 0.000898

c. Desviación absoluta media (MAD) super vector machine

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

MAD = 0.000898

d. Desviación absoluta media (MAD) árbol de decisiones

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

MAD = 0.000898

e. Desviación absoluta media redes neuronales 3 capas

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

MAD = 7.53

f. Desviación absoluta media Bayes

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

MAD = 7.69

g. Desviación absoluta Nearest Neighbors

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

MAD = 8.74

h. Desviación K-means

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

MAD = 9.1447

g. Desviación Spectral Clustering

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} \left| F_t - Y_t \right|}{\mathrm{n}}$$

MAD = 1.3229

3.1. Resultados en Tablas y Figuras

Al entrenar el data set, se obtuvo medidas de rendimiento de la aplicación de la técnica de Clasificación, como técnica de minería, que proporcionó resultados como error cuadrático medio, precisión y desviación absoluta media que se detallan a continuación.

Tabla 2.

Resumen resultante de la técnica de minería de Clasificación.

			Consumo de recursos			Exactitud		
			Tr	Uc	Cm	ECM	P	MAD
Técnica de Minería de Datos	Clasificación	Regresión Logística	0.0620	4.6	6.13	0.00090	99.93%	0.000898
		Regresión Lineal	0.0067	85	6.3978	0.94014	97.60%	0.000898
		Super vector Machine	48.8991	14	6.609	0.00090	99.91%	0.000898
		Redes Neuronales (3 Capas)	0.8907	98.8	5.4714	0.18778	81.22%	7.53 E
		Bayes	0.0034	4.2	7.1034	0.017071	98.29	7.69 E
		Nearest Neighbors	0.0105	21.0	6.9165	0.22013	77.99	8.74 E
		Arboles de Decisión	0.1921	3.9	6.5531	0.00090	99.91%	0.000898
	Agrupamiento	Kmeans	0.0638	84.5	6.41	0.45822	54.18	9.14 E
		Spectral Clustering	1.8361	59.3	6.2123	0.60916	39.08	1.32 E

Nota: Fuente elaboración propia

A continuación, en la figura 5, se evidencia el nivel porcentual de precisión con relación a cada una de las técnicas que se utilizó, para encontrar la más eficiente, destacando que es la de regresión logística la más destacada.

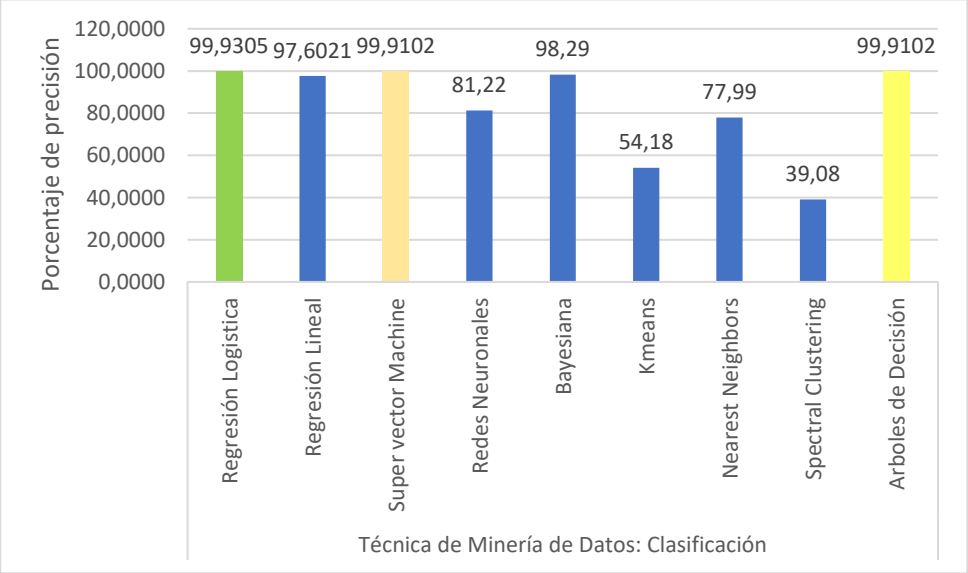


Figura 5. Representación del porcentaje de precisión. Fuente: Elaboración Propia

En la figura 6 se precisa que la técnica de regresión logística ofrece mejor tiempo de respuesta en comparación las demás que se precisan en la referida tabla.

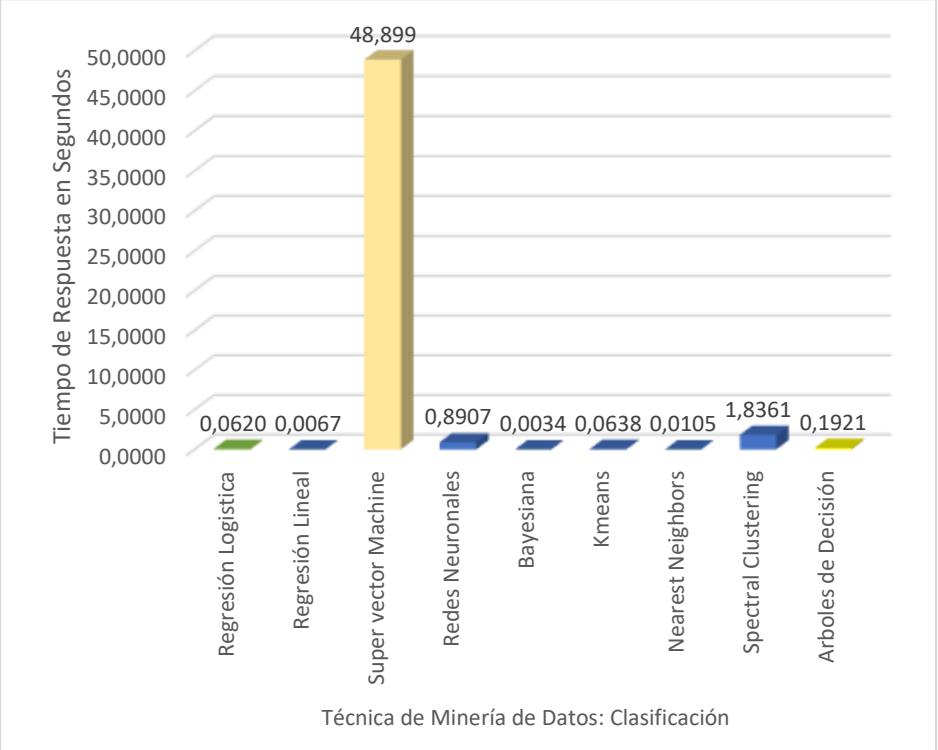


Figura 6. Representación del tiempo de respuesta en segundos. Fuente: Elaboración Propia.

3.2. Discusión de resultados.

En el estudio de comparación de técnicas de minería de datos, se evalúo su eficiencia que permita descubrir información relevante de ventas de una MYPE, evidenciando que de los resultados obtenidos aplicando la técnica de minería de datos de clasificación Regresión Logística es la más eficiente con relación a la predicción de ventas con las Boletas (normalizado con cero “0”) y facturas (normalizado con “1”), conforme se muestra en la tabla 6 cuyo resultado alcanzó el promedio de tiempo de respuesta de 0.0620 segundos y un buen nivel de precisión (P) que alcanzó el 0.9993%; en tanto que las demás no alcanzaron buen nivel; se evidencia además en regresión logística un error cuadrático (ECM) de 0.00090%, un grado de consumo de memoria (Cm) 6.13 Gigas, que considerando que las pymes disponen de escasos recursos para infraestructura tecnológica resulta la más viable, puesto que en

lo concerniente a desviación absoluta media son equiparables a todas los demás en esta técnica de minería de datos.

Los resultados del presente estudio, al ser contrastados con los obtenidos por Eti, S., Inel, M (2020), reveló que su metodología consistió en dividir aleatoriamente los datos de 50% para aprendizaje y 50% de prueba, proporcionó resultados en la que destaca las técnicas de redes neuronales y los árboles de decisión, donde ambas mostraron un nivel porcentual 94,286% en precisión.

Por lo tanto, se observó que la metodología en ambos casos es distinta en su aplicación; sin embargo, la técnica de árboles de decisión proporcionó, buen ratio de precisión porcentual en el presente estudio realizado, que no es el óptimo, prevaleciendo la regresión logística como el que nos proporciona mejor eficiencia por el consumo de recursos y nivel de exactitud.

3.3. Aporte práctico.

Proceso de selección de una MYPE comercial.

Para esta investigación fue necesario seleccionar una empresa como caso de estudio. Esta empresa debía ser una MYPE comercial teniendo en consideración que MYPE es aquella que tiene de 1 a 100 trabajadores con ventas ascendentes a un máximo de 1700 UIT (Unidades Impositivas Tributarias) conforme al art. 3ero. de la Ley 28015 que promueve e impulsa la formalización de las MYPE´s en el Perú. El rubro de negocio del caso de estudio genera información de ventas por ello es que la empresa seleccionada tiene actividades comerciales.

Para seleccionar la MYPE comercial, se tomó como referencia la cantidad de empresas a nivel nacional por región; las mismas que en total fueron previamente evaluadas, por el sistema financiero y han pasado a ser parte del Programa Reactiva Perú, cuya cantidad a nivel nacional ascienden a la suma de 501,298 y fueron empresas que se solventaron bajo criterios de elegibilidad, condiciones del préstamo garantizado y características de la garantía del gobierno que son las principales variables, por las que el sistema financiero peruano, se afianzó y a quienes el presente estudio estuvo dirigido, por cuanto son las que han requerido mayor impulso para promover el incremento de sus ventas y mejores tomas de decisiones, para el cumplimiento de sus pagos.

Tabla 3.

Empresas de tipo MYPE por regiones del Perú.

Nº	Regiones	Número de empresas
1	Amazonas	5,276
2	Áncash	22,045
3	Apurímac	5,153
4	Arequipa	31,450
5	Ayacucho	10,081
6	Cajamarca	27,663
7	Callao	8,619
8	Cusco	29,427
9	Huancavelica	1,728
10	Huánuco	8,131
11	Ica	9,549
12	Junín	23,252
13	La Libertad	25,833
14	Lambayeque	24,117
15	Lima	153,251
16	Loreto	4,134
17	Madre de Dios	3,424
18	Moquegua	4,432
19	Pasco	2,469
20	Piura	33,406
21	Puno	37,377
22	San Martín	10,050
23	Tacna	11,645
24	Tumbes	4,472
25	Ucayali	4,314
TOTAL EMPRESAS		**501,298**

Nota: Número de empresas por región que accedieron al crédito de reactiva Perú. Fuente: (MEF, 2020)

Considerando que Lambayeque es la zona de Influencia o donde la universidad ejerce su institucionalidad o rectoría y teniendo como soporte la Ley universitaria Nº 30220 establece que la universidad debe generar aporte a la universidad e investigación coincidiendo con la transversalidad de la responsabilidad social universitaria como núcleo y soporte, cuyo objeto es la satisfacción de los stakeholders y el logro de los ODS al 2030 con ámbito coordinado, consensuado y colaborativo se sostiene en la globalización, competitividad y empleabilidad. (Llerena, Silva, Quispe, & Ramos, 2020).

En Lambayeque lo conforman 24,117 empresas de los cuales en virtud al estudio se procedió a elegir el rubro de comercio que en Lambayeque lo conforman 13,128 empresas y es el rubro en que es más alto en cantidad de empresas.

Tabla 4.

Cantidad de empresas por rubro en la región Lambayeque.

RUBROS DE EMPRESAS	**CANTIDAD**
Comercio	13,128
Transporte, almacenamiento y comunicaciones	2,130
Industria manufacturera	2,024
Hoteles y restaurantes	781
Actividades inmobiliarias, empresariales, alquileres	1,502
Otras actividades de servicio comunitario	1,515
Agricultura, ganadería, caza y silvicultura	1,336
Organizaciones y órganos extraterritoriales	0
Construcción	1,017
Pesca	183
Electricidad, gas y agua	23
Servicios sociales y de salud	223
Enseñanza	206
Minería	29
Intermediación financiera	20
Administración pública y defensa	0
Hogares privados con servicio doméstico	0
TOTAL	**24,117**

Nota: Fuente: (MEF, 2020)

El criterio relevante del rubro al ser la de mayor cantidad de empresas, uso de TI, acceso de datos y procesos se identificó a una MYPE y se toma como caso de estudio a dicha empresa que es vinculada a unos de los rubros con más visibilidad y resaltante que está vinculada con la actividad económica de Comercio que accedió al crédito reactiva, rubro que representa la mayor cantidad de empresas en la Región Lambayeque.

Tabla 5.

Elección de la empresa como caso de estudio

NOMBRE DE EMPRESA	USO DE TI	ACCESO A DATOS
Empresa de transportes Emtrafesa	Alto	Negativo
Fametal	Medio	Negativo
Proyectos Ferretería Holgus E.I.R.L.	Medio	Positivo – Datos Inconsistentes
Comercial Damian EIRL	Medio	Positivo Algunos datos Inconsistentes

Nota: Selección de la empresa Proyectos Ferretería Damián E.I.R.L. por facilitar el acceso a sus datos y por contener data más consistente.

Fuente: Elaboración propia

Datos De La Empresa Generales de la Empresa

- **Razón Social :** COMERCIAL DAMIAN EIRL
- **Localización :** MZNA K LOTE 1-4 – URB SANTA MARIA – CHICLAYO – LAMBAYEQUE

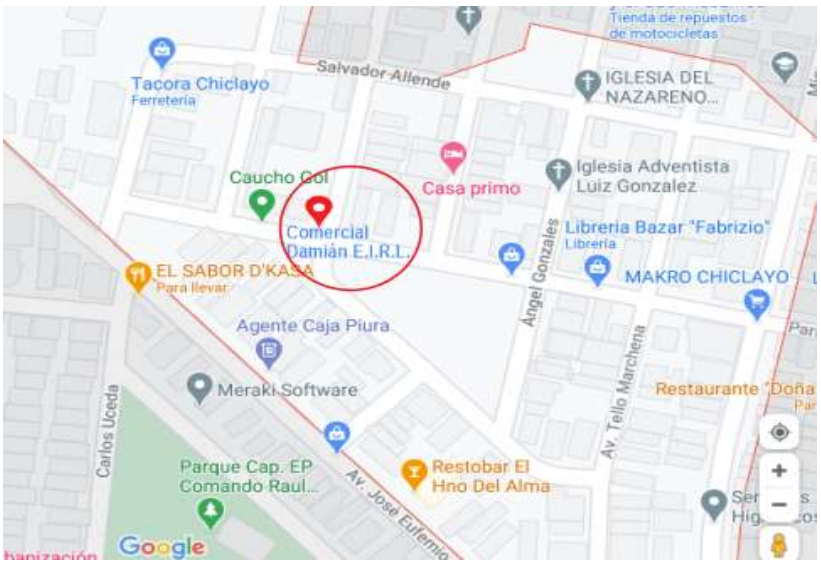


Figura 7. Plano de Ubicación de la Pyme seleccionada. Fuente: Google MAP

- **Visión**

  Contribuir con el desarrollo a través de la venta de productos y de Servicios de calidad, para generar bienestar y satisfacción del cliente

- **Misión**

  Posicionarnos como la mejor a nivel regional en tecnología, calidad en la venta y servicios, para distintos sectores comerciales, industria en general y para el hogar.

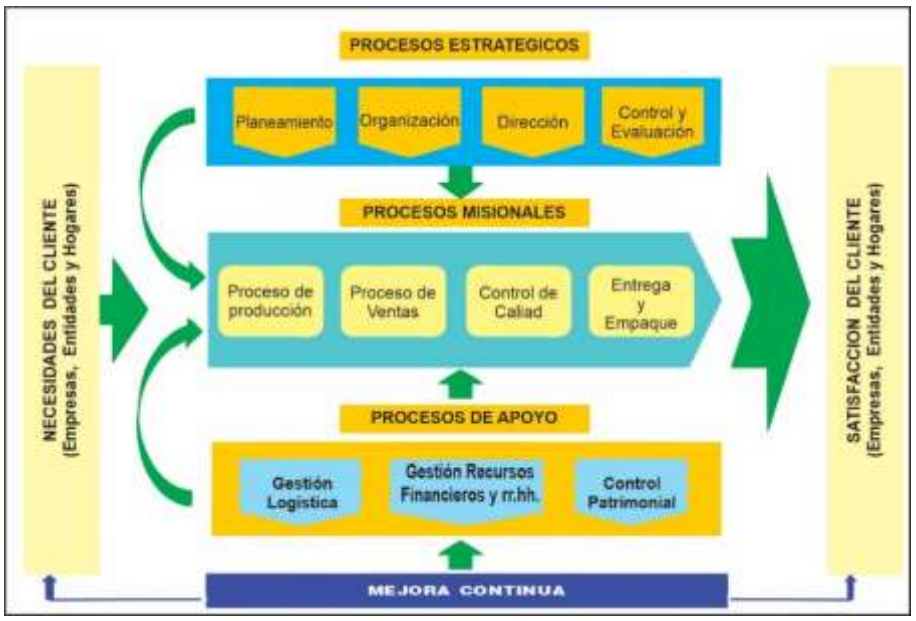


Figura 8. Mapa de Procesos de Comercial Damián E.I.R.L. Fuente: Elaboración propia

Preparar el conjunto de datos desde las bases de datos disponibles en el caso de estudio.

Esta etapa consistió en seleccionar los datos donde se procedió a obtener los datos de las bases de datos que son los datos estructurados y que conforman el historial de las transacciones de Comercial Damián EIRL. para la construcción de data sets que contiene información de ventas y de los clientes. Así mismo sumado a ello se obtuvo datos de hojas de cálculos que además conforman los datos en la cual la empresa la usa como punto de apoyo para

su sostenibilidad y el establecimiento de políticas y estrategias que empoderen la sostenibilidad en el negocio.

Como siguiente fase se desarrolló la integración de los datos, en el cual se apoya el presente estudio y se construyó la calidad de datos efectuándose el traslado de datos apropiados y necesarios a un contenedor ETL, al cual llegamos teniendo como origen la base de datos transaccional de la Empresa Damian EIRL construida en Postgres SQL en versión 2014; cuya estructura comprende productos, clientes, ventas, detalle de ventas, tipo de productos; las que son necesarias para nuestro estudio.

Se creó el modelo dimensional representando la base de datos origen en un Data Mart donde se asentaron las tablas preservando la estructura de como estaba compuesta cada tabla respecto a sus campos que la conforman con relación al tipo de datos, campos relevantes y necesarios para el estudio. Así mismo se creó la tabla de hechos la que se relacionó con las tablas prevista en el Data Mart, para a su vez aplicarle minería de datos a ese volumen de datos, para obtener el data set y posteriormente trasladarlo a formato de hoja de cálculo y finalmente a formato CSV, para el cumplimiento del objetivo propuesto.(Ver Anexo 5)

Esos datos de ventas históricas de los últimos 3 años son los que permitieron efectuar la estimación y la clasificación de clientes de un total de 5,522 que teniendo en cuenta que por la calidad de los datos se tomó como fundamento registros de a quienes se les vendió con boletas o facturas que representan una data significativa y cuya estructura lo conforma tipo, denominación, descripción, subtotal, IGV, total, year, mes. tipo de serie, conforme se precisa en el Anexo 6; considerando que para la PYME´s, lo relevante son los clientes y las ventas que es donde se sostiene el estudio. No obstante la calidad de los datos se presentó como una limitante al contener datos cualitativos que requieren ser unificados y recodificados haciendo uso de técnicas de observación y clasificación

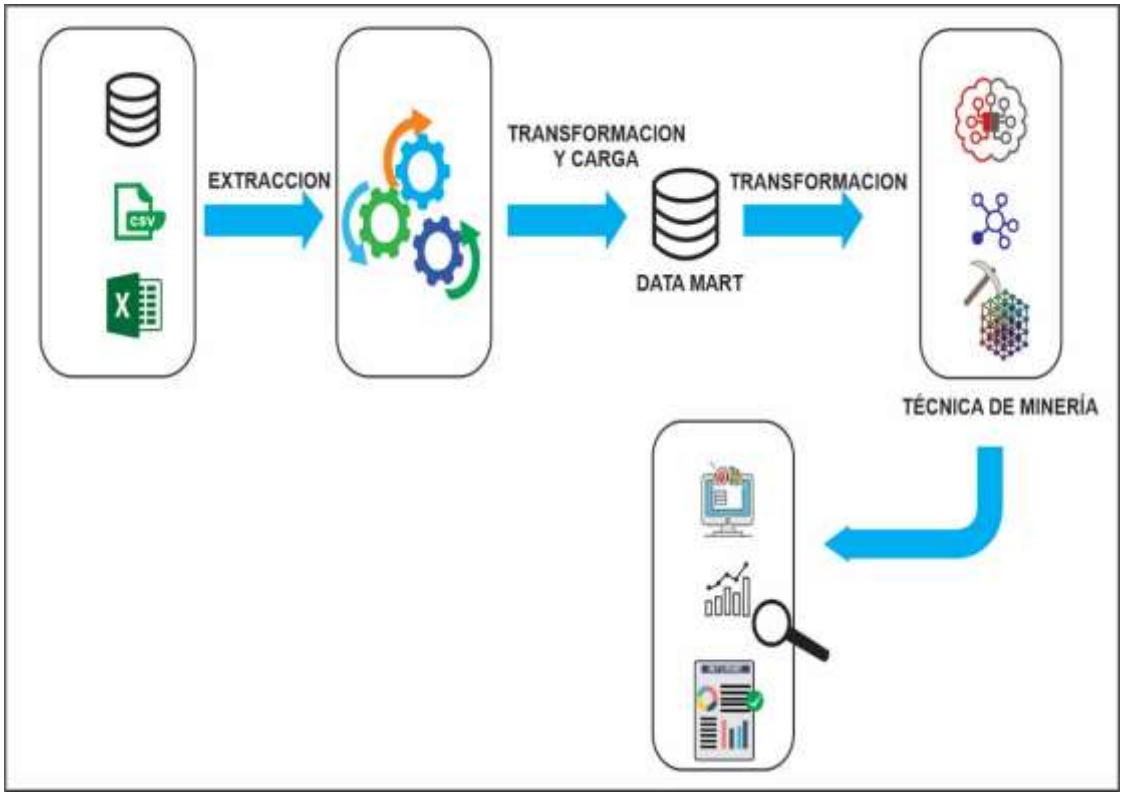


Figura 9. Tratamiento de los datos. Fuente: Elaboración propia

Las técnicas de minería se desarrollaron haciendo uso de la herramienta suite de Anaconda y Python que contiene aplicaciones y librerías abiertas, para el uso de la ciencia de datos. Así mismo; se hizo uso de Jupyter Notebook que es también un software open source y potente herramienta que además fue usado por su modelado estadístico y aprendizaje automático. Esta herramienta de software se aplicó a 5,522 registros con las técnicas de minería de datos seleccionadas, quedándonos con tan solo 1,113 registros. (Ver Anexo 7)

Seleccionar técnicas de data mining con mejor desempeño en el procesamiento de ventas.

Para la selección respecto al presente objetivo se efectuó una revisión de literatura científica de investigaciones realizadas y publicadas los últimos seis (6) años como artículos científicos y se encuentran sobre la base de datos como Ieeexplore, Scopus y Science Direct. Éstos artículos científicos, fueron extraídos teniendo en consideración los keyword “Machine Learning”, “data mining techniques” y operadores lógicos como “AND” “OR” para evaluar el desempeño de la técnica de minería de datos con su consecuente necesidad

que se ajuste a una MYPE logrando que de 15 Técnicas de minería de datos existentes las que se alinean a dicha necesidad es precisamente las técnicas de Classification, Clustering, Association Rule Mining, Temporal Mining y Neural Network, cuya descripción se detalla a continuación.

Un punto adicional que se tomó en cuenta es que los artículos en los cuales se basó la selección de la técnica de minería de datos es precisamente la descripción clara respecto a la técnica Data Mining con mayor precisión en sus resultados que ha coadyuvado a la construcción de la matriz elaborada donde se tuvo como preminencia el Año de investigación, técnica de minería y precisión en la técnica de sus resultados obtenidos.

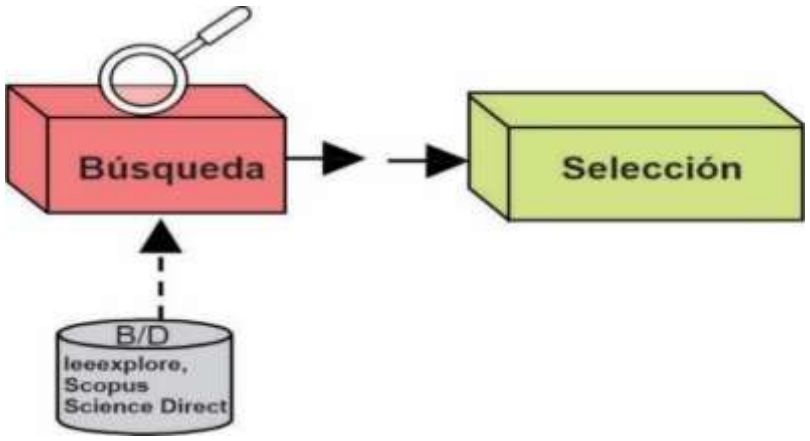


Figura 10. Representación de la Búsqueda de Artículos para la selección de la Técnica. *Fuente: Elaboración Propia*

Habiendo identificado las técnicas se puso en contexto su aplicación para virtualizar el objeto del presente estudio que es comparar cada uno de ellos en su eficiencia, su rendimiento, efectividad y la adaptación de la técnica.

Tabla 6.

Técnicas de Minería de Datos

N°	Técnica Data Mining	Precisión	Aplicación	Año
1	Clasificación	100%	Identificación de redes de interacción de gen significativos utilizando aprendizaje máquina y técnicas estadísticas. (Saranya & & Porkodi, 2019).	2019
2	Clustering	79.33%	Predicción de la lealtad del cliente en una empresa proveedora de servicios multimedia con segmentación de K-Means y algoritmo C4.5 (Moedjiono, Isak, & & Kusdaryono, 2016)	2016.
3	Neural Network	86.37%	Comparación de métodos de extracción de datos de clasificación para identificar a los funcionarios públicos en indonesio. (Sasmito & Ruldeviyani, 2020)	2020
4	Regression Analysis	98.74%	RAFFIA: Short-term Forest Fire Danger Rating Prediction via Multiclass Logistic Regression. (Lei Wang, Zhao, Wen, & & Qu, 2018)	2018
5	SVM	100%	SVM model for forest fire prediction. (Singh, Neethu, Madhurekaa, Harita, & & Mohan, 2021)	2021

Aplicar técnicas de minería en la base de datos del caso de estudio previamente seleccionado.

Regresión Logística

Se efectuó la prueba de la base de datos con la técnica de minería de datos de clasificación, haciendo uso del algoritmo de Regresión Logística, en la que se procedió a efectuar las mediciones de acuerdo a indicadores establecidos en operacionalización de variables pertinente a la dimensión de recursos y exactitud.

A un total de 1,113 registros que corresponde el 20% de la data set utilizado para testing y 80% para entrenamiento, previamente tratado o normalizado; se

logró detectar 386 registros con estado normal de verdaderos positivos o que son ventas con Boletas correctas y 0 falsos positivos o incorrectas. En tanto que 726 registros fueron facturas bien pronosticadas o verdaderos positivos y 1 de error o incorrectos.

```
ypred = model_1.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 11. Script para calcular la matriz de confusión de la Regresión Logística

Tabla 7.

Matriz de Confusión de Regresión Logística

Matriz de Confusión de la Clasificación			
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	386	0
	FACTURA (SI)	1	726
Regresión Logística			

Nota: Fuente Elaboración propia

Consumo de Recursos

```
# Tiempo de entrenamiento
inicio = time.time()
model_1.fit(x_train,y_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 12. Script para calcular el consumo de recursos.

a. Promedio de Tiempo Respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

b. Grado de Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j \, / n$$

c. Grado de Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
# muestra la presicion en el entrenamiento
accuracyTrain = model_1.score(x_train,y_train)
accuracyTest = model_1.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

Figura 13. Script para calcular la Exactitud

a. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n} (F_t - Y_t)^2}{n}$$

b. Precisión

$$P = \frac{VP}{FP + VP}$$

```
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

Figura 14. Script para el cálculo de la Desviación absoluta

c. Desviación absoluta media (MAD)

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} | F_t - Y_t |}{\mathrm{n}}$$

Regresión Lineal

Se efectuó la prueba de la base de datos con la técnica de minería de datos de clasificación, haciendo uso del algoritmo de Regresión Lineal, en la que se procedió a efectuar las mediciones de acuerdo a indicadores establecidos en operacionalización de variables, pertinente a la dimensión de recursos y exactitud.

Un total de 1,113 registros que corresponde el 20% del data set utilizado previamente tratado o normalizado; se logró detectar 386 registros con estado normal o verdaderos positivos o que son ventas con Boletas o correctas y 0 falsos positivos o incorrectas. En tanto que 726 registros fueron facturas bien pronosticadas o verdaderos positivos y 1 de error o incorrectos.

```
ypred = model_1.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 15. Script para calcular la matriz de confusión de la Regresión Lineal

Tabla 8.

Matriz de Confusión de regresión lineal

	Matriz de Confusión de la Clasificación		
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	386	0
	FACTURA (SI)	1	726
	Regresión Lineal		

Nota: Fuente Elaboración propia

Consumo de Recursos

```
# Tiempo de entrenamiento
inicio = time.time()
model_regresion.fit(x_train,y_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 16. Script para calcular la exactitud.

a. Promedio de Tiempo Respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

b. Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j \, / n$$

c. Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
# muestra la presicion en el entrenamiento
accuracyTrain = model_regresion.score(x_train,y_train)
accuracyTest = model_regresion.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

Figura 17. Script para calcular la exactitud

a. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

b. Precisión

$$P = \frac{VP}{FP + VP}$$

c. Desviación absoluta media (MAD)

$$\text{MAD} = \frac{\sum_{t=1}^{n} |F_t - Y_t|}{\text{n}}$$

Super Vector Machine

Se efectuó la prueba del data set con la técnica de minería de datos de clasificación, haciendo uso del algoritmo de Super vector machine, en la que se procedió a efectuar las mediciones de acuerdo a los indicadores establecidos en la operacionalización de variables en lo pertinente a la dimensión de recursos y exactitud.

A un total de 1,113 registros que corresponde el 20% del data set utilizado previamente tratado o normalizado; se logró detectar 386 registros con estado normal o verdaderos positivos o que son ventas con Boletas o correctas y 0 falsos positivos o incorrectas. En tanto que 726 registros fueron facturas bien pronosticadas o verdaderos positivos y 1 de error o incorrectos como falso negativo.

```
ypred = model_5.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 18. Script para calcular la matriz de confusión de la Super Vector Machine

Tabla 9.

Matriz de Confusión Super Vector Machine

	Matriz de Confusión de la Clasificación		
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	386	0
	FACTURA (SI)	1	726
	Super Vector Machine		

Nota: Fuente Elaboración propia

Consumo de Recursos

```
# Tiempo de entrenamiento
inicio = time.time()
model_5.fit(x_train,y_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 19. Script para calcular la exactitud.

a. Promedio de Tiempo Respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

b. Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j \, / n$$

c. Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
# muestra la presicion en el entrenamiento
accuracyTrain = model_5.score(x_train,y_train)
accuracyTest = model_5.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

Figura 20. Script para calcular la exactitud

a. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

b. Precisión

$$P = \frac{VP}{FP + VP}$$

c. Desviación absoluta media (MAD)

$$\text{MAD} = \frac{\sum_{t=1}^{n} |F_t - Y_t|}{\text{n}}$$

3.4. Árbol de decisiones

Se efectuó la prueba del data set con la técnica de minería de datos de clasificación, haciendo uso del algoritmo de árbol de decisiones, en la que se procedió a efectuar las mediciones de acuerdo a indicadores establecidos en

operacionalización variables correspondiente a la dimensión de recursos y exactitud.

A un total de 1,113 registros que corresponde el 20% del data set utilizado previamente tratado o normalizado; se logró detectar 386 registros con estado normal o verdaderos positivos o que son ventas con Boletas o correctas y 0 falsos positivos o incorrectas. En tanto que 726 registros fueron facturas bien pronosticadas o verdaderos positivos y 1 de error o incorrecto o Falso Negativos.

```
ypred = model_7.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 21. Script para calcular la matriz de confusión de árboles de decisiones.

Tabla 10.

Matriz de Confusión de árboles de decisiones

	Matriz de Confusión de la Clasificación		
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	386	0
	FACTURA (SI)	1	726
	Árboles de Decisión		

Nota: Fuente Elaboración propia

Consumo de Recursos

```
# Tiempo de entrenamiento
inicio = time.time()
model_7.fit(x_train,y_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 22. Script para calcular la exactitud.

a. Promedio de Tiempo Respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

b. Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j \, / n$$

c. Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
# muestra la presicion en el entrenamiento
accuracyTrain = model_7.score(x_train,y_train)
accuracyTest = model_7.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

Figura 23. Script para calcular la exactitud

a. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n} (F_t - Y_t)^2}{n}$$

b. Precisión

$$P = \frac{VP}{FP + VP}$$

c. Desviación absoluta media (MAD)

$$\text{MAD} = \frac{\sum_{t=1}^{n} |F_t - Y_t|}{\text{n}}$$

3.5. Redes Neuronales (3 Capas)

Se efectuó la prueba del data set con la técnica de minería de datos de clasificación, haciendo uso del algoritmo de redes neuronales, en la que se procedió a efectuar las mediciones de acuerdo a indicadores establecidos en operacionalización variables correspondiente a la dimensión de recursos y exactitud.

A un total de 1,113 registros que corresponde el 20% del data set utilizado previamente tratado o normalizado; se logró detectar 200 registros con estado normal o verdaderos positivos, que son ventas con Boletas o correctas y 23 falsos negativos o incorrectas. En tanto que 704 registros fueron facturas bien pronosticadas o verdaderos positivos y 186 de error o incorrectos o Falsos Positivos.

```
ypred = model_4.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 24. Script para calcular la matriz de confusión de redes neuronales

Tabla 11.

Matriz de Confusión de Redes Neuronales (3 capas)

Matriz de Confusión de la Clasificación			
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	200	186
	FACTURA (SI)	23	704
Redes Neuronales (3 capas)			

Nota: Fuente Elaboración propia

Consumo de Recursos

```
# Tiempo de entrenamiento
inicio = time.time()
model_4.fit(x_train,y_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 25. Script para calcular el consumo de recursos.

a. Promedio de Tiempo Respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

b. Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j \, / n$$

c. Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
# muestra la presicion en el entrenamiento
accuracyTrain = model_4.score(x_train,y_train)
accuracyTest = model_4.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

Figura 26. Script para calcular la exactitud

a. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

b. Precisión

$$P = \frac{VP}{FP + VP}$$

c. Desviación absoluta media (MAD)

$$\text{MAD} = \frac{\sum_{t=1}^{n} |F_t - Y_t|}{\text{n}}$$

3.6. Bayes

Se efectuó la prueba del data set con la técnica de minería de datos de clasificación, haciendo uso del algoritmo de Bayes, en la que se procedió a efectuar las mediciones de acuerdo a indicadores establecidos en operacionalización variables correspondiente a la dimensión de recursos y exactitud.

A un total de 1,113 registros que corresponde el 20% del data set utilizado previamente tratado o normalizado; se logró detectar 367 registros con estado normal o verdaderos positivos, que son ventas con Boletas o correctas y 0 falsos negativos o incorrectas. En tanto que 727 registros fueron facturas bien pronosticadas o verdaderos positivos y 0 de error o incorrectos o Falsos Negativos.

```
ypred = model_6.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 27. Script para calcular la matriz de confusión de Bayes

Tabla 12.

Matriz de Confusión de Bayes

	Matriz de Confusión de la Clasificación		
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	367	19
	FACTURA (SI)	0	727
		Bayes	

Nota: Fuente Elaboración propia

Consumo de Recursos

```
# Tiempo de entrenamiento
inicio = time.time()
model_6.fit(x_train,y_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 28. Script para calcular el consumo de recursos.

a. Promedio de Tiempo Respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

b. Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j \, / n$$

c. Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
# muestra la presicion en el entrenamiento
accuracyTrain = model_6.score(x_train,y_train)
accuracyTest = model_6.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

Figura 29. Script para calcular la exactitud.

a. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

b. Precisión

$$P = \frac{VP}{FP + VP}$$

c. Desviación absoluta media (MAD)

$$\text{MAD} = \frac{\sum_{t=1}^{n} |F_t - Y_t|}{\text{n}}$$

3.7. Nearest Neighbors

Se efectuó la prueba del data set con la técnica de minería de datos de clasificación, haciendo uso del algoritmo de Nearest Neighbors, en la que se procedió a efectuar las mediciones de acuerdo a indicadores establecidos en operacionalización variables correspondiente a la dimensión de recursos y exactitud.

A un total de 1,113 registros que corresponde el 20% del data set utilizado previamente tratado o normalizado; se logró detectar 170 registros con estado

normal o verdaderos positivos, que son ventas con Boletas o correctas y 29 falsos negativos o incorrectas. En tanto que 698 registros fueron facturas bien pronosticadas o verdaderos positivos y 216 de error o incorrectos o Falsos Positivos.

```
ypred = model_8.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 30. Script para calcular la matriz de confusión de Nearest Neighbors

Tabla 13.

Matriz de Confusión de Nearest Neighbors

	Matriz de Confusión de la Clasificación		
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	170	216
	FACTURA (SI)	29	698
	Nearest Neighbors		

Nota: Fuente Elaboración propia

Consumo de Recursos

```
# Tiempo de entrenamiento
inicio = time.time()
model_9.fit(x_train,y_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 31. Script para calcular el consumo de recursos.

a. Promedio de Tiempo Respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

b. Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j \, / n$$

c. Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
# muestra la presicion en el entrenamiento
accuracyTrain = model_9.score(x_train,y_train)
accuracyTest = model_9.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

Figura 32. Script para calcular la exactitud.

a. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n} (F_t - Y_t)^2}{n}$$

b. Precisión

$$P = \frac{VP}{FP + VP}$$

c. Desviación absoluta media (MAD)

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} | F_t - Y_t |}{\mathrm{n}}$$

3.8. K-means

Se efectuó la prueba del data set con la técnica de minería de datos de agrupamiento haciendo uso del algoritmo de K-means mediante el tipo de aprendizaje no supervisado en la que se procedió a efectuar las mediciones de acuerdo a indicadores establecidos en operacionalización variables correspondiente a la dimensión de recursos y dimensión de exactitud.

A un total de 1,113 registros que corresponde el 20% del data set utilizado previamente tratado o normalizado; se logró detectar 151 registros con estado normal o verdaderos positivos, que son ventas con Boletas o correctas y 284 falsos negativos o incorrectas. En tanto que 452 registros fueron facturas bien pronosticadas o verdaderos positivos y 226 de error o incorrectos o Falsos Positivos.

```
ypred = kmeans.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 33. Script para calcular la matriz de confusión K-means.

Tabla 14

Matriz de Confusión de Kmeans

	Matriz de Confusión de Agrupamiento		
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	151	226
	FACTURA (SI)	284	452
		Kmeans	

Nota: Fuente Elaboración propia

Consumo de Recursos

```
kmeans = KMeans(n_clusters=2, max_iter = 300)
inicio = time.time()
kmeans.fit(x_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 34. Script para calcular el consumo de recursos.

a. Promedio de tiempo respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

b. Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j \, / n$$

c. Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
# muestra la presicion en el entrenamiento
predictions = kmeans.predict(x_test)
print('accuracy_score(y_test, predictions))
```

a. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n} (F_t - Y_t)^2}{n}$$

b. Precisión

$$P = \frac{VP}{FP + VP}$$

c. Desviación absoluta media (MAD)

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} | F_t - Y_t |}{\mathrm{n}}$$

3.9. Spectral Clustering

Se efectuó la prueba del data set con la técnica de minería de datos de agrupamiento haciendo uso del algoritmo de Spectral Clustering mediante el tipo de aprendizaje no supervisado en la que se procedió a efectuar las mediciones de acuerdo a indicadores establecidos en operacionalización variables correspondiente a la dimensión de recursos y dimensión de exactitud.

A un total de 1,113 registros que corresponde el 20% del data set utilizado previamente tratado o normalizado; se logró detectar 97 registros con estado normal o verdaderos positivos, que son ventas con Boletas o correctas y 398 falsos negativos o incorrectas. En tanto que 280 registros fueron facturas bien pronosticadas o verdaderos positivos y 338 de error o incorrectos o Falsos Positivos.

```
ypred = spectral.fit.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

Figura 35. Script para calcular la matriz de confusión.

Tabla 15.

Matriz de Confusión de Spectral Clustering

Matriz de Confusión de Agrupamiento			
Realidad		BOLETA (NO)	FACTURA (SI)
	BOLETA (NO)	97	280
	FACTURA (SI)	398	338
Spectral Clusterings			

Nota: Fuente Elaboración propia

Consumo de Recursos

```
inicio = time.time()
spectral.fit(x_train)
fin = time.time()
print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3)
print('Grado de consumo de CPU        :',psutil.cpu_percent())
print('Consumo de memoria fisica      : {}'.format(memory))
```

Figura 36. Script para calcular el consumo de recursos.

d. Promedio de Tiempo Respuesta

$$Tr = \sum_{j}^{n} \frac{tf_j - tf_i}{n}$$

e. Consumo de CPU

$$Uc = \sum_{j}^{n} uc_j / n$$

f. Consumo de memoria

$$Cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Exactitud

```
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

Figura 37. Script para calcular error cuadrático medio y desviación absoluta media.

d. Error cuadrático medio

$$ECM = \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{n}$$

e. Precisión

$$P = \frac{VP}{FP + VP}$$

f. Desviación absoluta media (MAD)

$$\text{MAD} = \frac{\sum_{t=1}^{n} |F_t - Y_t|}{\text{n}}$$

Efectuada la pruebas de la técnica de minería de datos, destacando que es precisamente la técnica de minería de datos de regresión logística, la que nos proporciona mayor precisión se puede evidenciar por ejemplo que en la realidad de los hechos, de acuerdo a las transacciones realizadas por la empresa, se demostró ante el volumen de sus ventas, que revela la cantidad de clientes a quienes se les ha emitido Boleta o la cantidad de quienes han sido facturados, conforme se muestra en el Figura 10 y que se solventa en la matriz de confusión evidenciado en la tabla 7, lo que podría ser utilizada, para un cambio de políticas e incluso en el marketing o direccionamiento de las ventas ya sea apersonas naturales o jurídicas que de acuerdo a su la rentabilidad en sus ventas podría orientar incluso promociones.

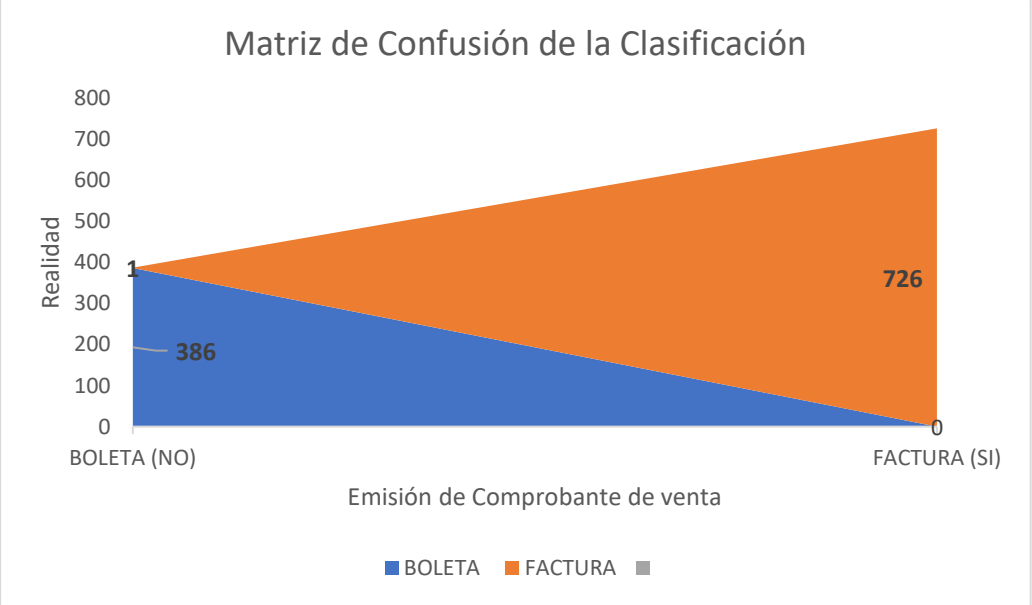


Figura 38. Representación de la Evidencia en cómo podría ayudar a la Empresa la información relevante para la toma de decisiones.

IV. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones.

a. Se determinó elegir a una MYPE comercial de la Región Lambayeque en virtud al área de influencia de la Universidad señor de Sipán y teniendo en consideración criterios establecidos por el gobierno al programa reactiva Perú creado mediante Decreto Legislativo Nº 1455 y modificado mediante Decreto Legislativo 1457por el impacto COVID-19.

b. Se preparó el conjunto de datos desde la base de datos de la Empresa Damian EIRL la misma que nos reveló un hallazgo resaltante en las MYPE´s de la Región Lambayeque que evidencia que para las mismas su prioridad y relevancia son las ventas y lo menos relevante el correcto registro de sus ventas que requiere de la preservación del celo y cuidado en la gestión de sus datos que redundan en inconsistencias.

c. Se seleccionaron la técnica de minería de datos predictiva de clasificación como regresión logística, regresión lineal simple, support vector machine, árboles de Decision, Neural Network, Nearest Neighbors, por su eficiencia y rendimiento como una de las técnicas de minería más usado y que proporciona mejores resultados. Así mismo se seleccionó la técnica descriptiva de agrupamiento o segmentación como K-mean, Spectral Clustering. Ambas Técnicas de minería de datos en virtud a lo analizado de los artículos científicos de las bases de datos Ieeexplore, Scopus y Science Direct, que son las más usadas por su eficiencia y rendimiento.

d. Se evaluó el desempeño de las técnicas de minería de datos y se evidenció que la más eficiente resulto ser la técnica de clasificación a través de su algoritmo de Regresión Logística debido a que el promedio de tiempo de respuesta fue de 0.0620 segundos, nivel de precisión (P) de 0.9993%; en tanto que en lo correspondiente a consumo de memoria, error cuadrático medio (ECM) y desviación absoluta media (MAD) sus niveles de eficiencia son iguales con otros algoritmos de la técnica de minería de datos de Clasificación.

4.2. Recomendaciones.

a. Se recomienda priorizar los procesos de ventas debido a que las MYPES guardan información de ventas en tanto que otras transacciones o gestiones no tienen la prioridad en sus rutinas de registros o gestiones

b. Se tome interés en la verificación sus datos, con la finalidad de detectar redundancia e inconsistencia de datos.

c. Se normalice los datos haciendo usos de técnicas de escalado de variables o de mínimos y máximos, para facilitar su procesamiento mediante técnicas data mining y evitar resultados erróneos.

d. Seleccionar las técnicas data mining, sobre la base de literatura de artículos científicos consultados en Scopus, Ieeexplore, y Science Direct que nos conducirá finalmente a la comparación de resultados donde se podrá evaluar y contrastar el mejor rendimiento o desempeño de cada una de ellas.

e. Aplicar las técnicas data mining de clasificación por ser una de las que contienen algoritmos determinantes que ofrecen resultados que pueden servir en un estudio del información de MYPE´s, conforme nos ha revelado el presente estudio en relación al nivel de precisión y el tiempo de respuesta; ya que siempre habrá más datos de lo que se espera, que requieren ser analizados y normalizados, para su procesamiento y posterior toma de decisiones o adopción de políticas y estrategias.

REFERENCIAS.

Abu-Bader, S. H. (2021). *Using statistical methods in social science research: With a complete SPSS guide.* New York, USA: Oxford University Press, USA.

Afifi, A. (2020). Demand Forecasting of Short Life Cycle Products Using Data Mining Techniques. *Springer, Cham, 583*, 151–162. doi:https://doi.org/10.1007/978-3-030-49161-1_14

Aggarwal, C. (2015). *Data Mining The Textbook.* New York: Springer.

Albrieu, R., Rapetti, M., López, C., & Larroulet, P. (2018). *Inteligencia artificial y crecimiento económico: oportunidades y desafíos para Perú.* Buenos Aires: CIPPEC.

Alessandro, L., Palesi, L. A., Nesi, P., & Pantaleo, G. (2022). Multi Clustering Recommendation System for Fashion Retail. *Multimedia Tools and Applications*, 1-28. Retrieved from https://link.springer.com/article/10.1007/s11042-021-11837-5#citeas

Anoopkumar, M. & Rahman, A. (2016). A Review on Data Mining Techniques and factors used in Educational Data Mining to predict student amelioration. *IEEE Xplore*, 1-25.

Award, M., & Khana, R. (2015). *Efficient learning machines - Theories, Concepts, and Aoplication for Engineers.* New York, USA: Apres Open.

Bell, J. (2015). *Machine Learning Hands-On for Developers and Technical Professionals.* Indianápolis, India: Wiley.

Bramer, M. (2016). *Principles of Data mining.* London: Springer.

CAPECE. (2021). *Reporte Oficial de industria Ecomerce en Perú.*

Carhuana, R., & Carhuana, M. (2019). *Aplicación de inteligencia de negocios para la toma de decisiones en el área.* Huancavelica: UNH.

del Moral, M., Chiclana, F., Tapia, J., Tapia, C., & Herrera, E. (2019). A comparison between Fuzzy Linguistic RFM Model and traditional RFM model applied to Campaign Management. Case study of retail business. *ScienceDirect*, 571-578.

Devi, S. G. (2014). A survey on distributed data mining and its trends.

Diaz, J., Deza, M., & Moreno, K. (2020). *Informe BID Desafios del desarrollo en el Post Covid – 19 (2020).* Santiago, Chile: BID.

Dine, N., & Nuñez, G. (2020). *Elementos para la innovación de las políticas dirigidas a las mipymes y para la defensa de la competencia a la luz de los desafíos impuestos por la pandemia y la recuperación económica.* Santiago, Chile: CEPAL y UNCTAD.

Dinov, I. (2018). *Data Science and Predictive Analytics.* Michigan, USA.: Springer.

Eti, S. , İnel, M. (2020). A research on the comparison of classification algorithm in finance | [Una investigación sobre la comparación del algoritmo de clasificación en las finanzas]. *Scopus*, 1-15.

Grández Márquez, M. A. (2017). *Aplicación de Minería de Datos para Determinar Patrones de Consumo Futuro en Clientes de una Distribuidora de Suplementos Nutricionales.* Lima: SIL.

Heredia, A. (2019). *Políticas e instrumentos para fomentar la incorporación de tecnologías digitales en el desarrollo empresarial de las MIPYMES de América Latina.* Santiago, Chile.

Heredia, A. (2020). *Políticas de fomento para la incorporación de las tecnologías digitales en las micro, pequeñas y medianas empresas de América Latina.* Santiago: CEPAL.

Hernández, S., Fernández, C., & Baptista, P. (2014). *Metodología de la Investigación* (Vol. 6ta. Edit.). México: Mc Graw Hill.

Igual, L., & Seguí,, S. (2017). *Introduction to Data Science - A Python Approach to Concepts, Techniques and Applications.* Barcelona, Spain: Springer.

INEI. (2021). *Informe Técnico Demografia Empresarial.* Lima, PErú: INEI.

Islam, M., Rafi, M. , Azad, A. & Ovi, J. . (n.d.).

Islam, M., Rafi, M. , Azad, A. & Ovi, J. (2021). *Weighted frequent sequential pattern mining. Applied Intelligence.* Bangladesh: Springer Link.

Jiachen, W., Yaling, W. & Yuea, W. (2020). Applying Clustering and Co-occurrence Methods to Identifying Key Events and Their Relations in Chinese Stock Market. 102-110.

Kopec, D. (2019). *Classic Computer Science Problems in Python.* New York: MANNING SHELTER ISLAND.

Kopec, D. (2019). *Classic Computer Science Problems in Python .* New York, USA: Manning - Shelter Island.

Lei Wang, L., Zhao, Q., Wen, Z., & & Qu, J. (2018). RAFFIA: Short-term Forest Fire Danger Rating Prediction via Multiclass Logistic Regression. *researchgate*, 1-16.

LI, H., Wu, Y. & Chen, Y. (2019). Time is money: Dynamic-model-based time series data-mining for correlation analysis of commodity sales. *Journal PreProf*, 1-25.

Llerena, S., Silva, B., Quispe, J., & Ramos, A. (2020). Responsabilidad Social Universitaria: Tranversalidad y Desarrollo Sostenible en Latinoamérica. *Journal of Business and Enterpreneurial Studie*, 328-340.

Mardiantien, C. R., Atastina, I., & Asror, I. (2020). Product segmentation based on sales transaction data using agglomerative hierarchical clustering and FMC model (Case Study: XYZ Company). *2020 3rd International Conference on Information and Communications Technology (ICOIACT)* (pp. 280-285). Yogyakarta, Indonesia : IEEE. doi:10.1109/ICOIACT50329.2020.9332023

Martinez, C., & Palencia, O. (2021). Modelo de minería de datos para el análisis de la productividad y crecimiento personal en las mujeres emprendedoras: el caso de la Asociación las Rosas. *Suma de Negocios*, 23-30.

Martínez, R., Carrasco, R., García, J., Gallego, C., & Herrera, E. (2019). A comparison between Fuzzy Linguistic RFM Model and traditional RFM model applied to Campaign Management. Case study of retail business. *ScienceDirect*, 281-289.

Masson, E., & Olesen, C. (2021). Acceso digital como reconstitución de archivos: muestreo algorítmico, visualización y producción de significado en grandes repositorios de imágenes en movimiento. *Journals*, 22.

MEF. (2020, 06 13). *Ministerio de Economía y Finanzas - Perú*. Retrieved from Lista_Empresas_Beneficiadas_Reactiva_Perú: https://www.mef.gob.pe

Mejía, H., Tuesta, V., Samillan, A., & Forero, M. (2021). *New Agile Enterprise Architecture Methodology for Small Latin American Organizations.* USA: Springer, Cham.

Ministerio de Economía y Finanzas. (2020). *Informe Marco Macroeconómico Multianual 2021-2024 (2020) probado mediante el D.U. Nº 075-2020.* Lima.

Ministerio de la Producción. (2020). *Las MIPYME en cifras 2018.* Lima, Perú: Ministerio de la Producción.

Moedjiono, S., Isak, Y., & & Kusdaryono, A. (2016). Customer loyalty prediction in multimedia Service Provider Company with K-Means segmentation and C4.5 algorithm," 2016 International Conference on Informatics and Computing . *IEE (ICIC)*, 210-215.

Morente, J. R. (2020). Using clustering methods to deal with high number of alternatives on Group Decision Making. *ELSEVIER*, 316-323. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050919320022

Musa, K., AlShehadeh, A., & & Alqerem, R. (2019). The Role of Data Mining Techniques in the Decision-Making Process in Jordanian Commercial Banks. *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 726-730.

Ñaupas, C. (2016). *Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias.* Lima - Perú: UNMSM.

Pabón, Torres & Bucheli. (2020). Un enfoque de Análisis Inteligente de Datos para Apoyar la Relación con los Clientes. *Revista Ibérica de Sistemas e Tecnologias de Informação*, 52-66.

Palakshappa, A. & Patil, M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University – Computer and Information Sciences*, 1-8.

Palma, J. & Marín,R. (2008). *Inteligencia Artificial - Técnicas, Métodos y Aplicaciones.* España: Mc Graw Hill.

Pan, H., & Zhou, H. (2020). Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce. *Electronic Commerce Research*, 297-320. doi:https://doi.org/10.1007/s10660-020-09409-0

Parrella, M., Albano, G., Perna, C., & La Rocca, M. (2021). (2021). Bootstrap joint prediction regions for sequences of missing values in spatio-temporal datasets. *Springer Link*, 1-22.

Presidencia del Concejo de Ministros. (2021). *Estrategia Nacional de Inteligencia Artificial en el Perú.* Lima - Perú: PCM.

Saranya, v., & & Porkodi, R. (2019). Identifying Significant Gene Interaction Networks Using Machine Learning and Statistical Techniques. *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT).*, pp. 1-7.

Sardo, F., Pinho, C. & Saur, I. (2021). Co-creation and consumer Experiences: A Systematic Literature Review. *Researchgate*, 1020-1024.

Sasmito, a., & Ruldeviyani, Y. (2020). Comparison of The Classification Data Mining Methods to Identify Civil Servants in Indonesian Social Insurance Company. *3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRIT)*, 111-116.

Serrano, L. (2020). *Exploring Machine Learning Basics.* New York - USA: Manning Author Picks.

Silvaa,J., Varela, N., Borrero,L. & Rojas, R. (2019). Association Rules Extraction for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm. *ScienceDirect*, 1207–1212.

Singh, K., Neethu, K., Madhurekaa, K., Harita, A., & & Mohan, P. (2021). Parallel SVM model for forest fire prediction. *SciencieDirect*, 1-12. doi:(https://www.sciencedirect.com/science/article/pii/S2666222121000046)

Varela, N.,Cabrera, H., Lopez,G., Vilora, A.,Gaitan, M. & Ardila, M. (2019). Methodology for the Reduction and Integration of Data in the Performance Measurement of Industries Cement Plants. *Springer*, 33-42.

Wisesa, O., Adriansyah, A., & Khalaf, O. I. (2020). Prediction analysis sales for corporate services telecommunications company using gradient boost algorithm. *International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP)* (pp. 101-106). Yogyakarta, Indonesia: IEEE. doi:10.1109/BCWSP50066.2020.9249397

Xie, Y. (2021). Construction of Japanese Interactive Analysis System Based on Perception Model and Pattern Mining. *IEEE Xplore*, 1584-1587.

Yadao,S.; Vinaya, A.; Janarthanan, M.; & Bhaumik, A. (2021). Web usage Mining: A Comparison of WUM Category Web Mining Algorithms. *Third International Conference on Intelligent Communication Technologies and Virtual* (pp. 1020-1024). Tirunelveli, India: IEEE.

ANEXOS.

Anexo 1. Resolución de aprobación del proyecto de investigación

USS | UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

RESOLUCIÓN N°0445-2021/FIAU-USS

Pimentel, 27 de mayo de 2021

VISTO:
El Acta de reunión N°1305-2021 del Comité de investigación de la Escuela profesional de INGENIERÍA DE SISTEMAS remitida mediante oficio N°0227-2021/FIAU-IS-USS de fecha 19 de mayo de 2021, y;

CONSIDERANDO:
Que, de conformidad con la Ley Universitaria N° 30220 en su artículo 48° que a letra dice: "La investigación constituye una función esencial y obligatoria de la universidad, que la fomenta y realiza, respondiendo a través de la producción de conocimiento y desarrollo de tecnologías a las necesidades de la sociedad, con especial énfasis en la realidad nacional. Los docentes, estudiantes y graduados participan en la actividad investigadora en su propia institución o en redes de investigación nacional o internacional, creadas por las instituciones universitarias públicas o privadas.";

Que, de conformidad con el Reglamento de grados y títulos en su artículo 21° señala: "Los temas de trabajo de investigación, trabajo académico y tesis son aprobados por el Comité de Investigación y derivados a la Facultad o Escuela de Posgrado, según corresponda, para la emisión de la resolución respectiva. El periodo de vigencia de los mismos será de dos años, a partir de su aprobación. En caso un tema perdiera vigencia, el Comité de Investigación evaluará la ampliación de la misma.

Que, de conformidad con el Reglamento de grados y títulos en su artículo 24° señala: La tesis es un estudio que debe denotar rigurosidad metodológica, originalidad, relevancia social, utilidad teórica y/o práctica en el ámbito de la escuela profesional. Para el grado de doctor se requiere una tesis de máxima rigurosidad académica y de carácter original. Es individual para la obtención de un grado; es individual o en pares para obtener un título profesional. Asimismo, en su artículo 25° señala: "El tema debe responder a alguna de las líneas de investigación institucionales de la USS S.A.C.".

Que, según documentos de Vistos el Comité de investigación de la Escuela profesional de INGENIERÍA DE SISTEMAS acuerdan aprobar los temas de las Tesis a cargo de los estudiantes del curso de Investigación I que se detallan en el anexo de la presente Resolución.

Estando a lo expuesto, y en uso de las atribuciones conferidas y de conformidad con las normas y reglamentos vigentes;

SE RESUELVE:

ARTÍCULO 1°: APROBAR, el tema de la Tesis perteneciente a la línea de investigación de INFRAESTRUCTURA, TECNOLOGÍA Y MEDIO AMBIENTE, a cargo de los estudiantes del Programa de estudios de INGENIERÍA DE SISTEMAS según se detalla en el anexo de la presente Resolución.

ARTÍCULO 2°: ESTABLECER, que la inscripción del Tema de la Tesis se realice a partir de emitida la presente resolución y tendrá una vigencia de dos (02) años.

ARTÍCULO 3°: DEJAR SIN EFECTO, toda Resolución emitida por la Facultad que se oponga a la presente Resolución.

REGÍSTRESE, COMUNÍQUESE Y ARCHÍVESE

Cc: Interesado, Archivo

UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

RESOLUCIÓN N°0445-2021/FIAU-USS

Pimentel, 27 de mayo de 2021

ANEXO

N°	AUTOR (ES)	TEMA DE TESIS
1	RIMARACHIN ESCRIBANO NERI RUT NIÑO MORENO NAJHELY YAMILETT	EVALUACIÓN DE TÉCNICAS DE CIFRADO PARA EL INTERCAMBIO DE DATOS DE INTERNET DE LAS COSAS EN EL ÁMBITO DE LA SALUD
2	GUEVARA CHAMBERGO JHON DENNIS BOBADILLA CAMPOS ROLANDO MARTIN	DESARROLLO DE UNA METODOLOGÍA DE GESTIÓN DE RIESGOS AD HOC BASADA EN MARCOS INTERNACIONALES Y BUENAS PRÁCTICAS PARA UNA EMPRESA MANUFACTURERA PERUANA
3	CIEZA CELIS JESUS ABELARDO OJEDA ROMERO ANTHONNY JHONATAN	EVALUACIÓN DEL DESEMPEÑO DE LOS ESQUEMAS DE SEGURIDAD DE RED PARA COMBATIR VULNERABILIDADES EN REDES INALÁMBRICAS BASADAS EN EL PROTOCOLO WPA2
4	MENDOZA FERRÉ ESPERANZA NATALY CABRERA SANCHEZ KEVIN ALONSO	COMPARACIÓN DEL RENDIMIENTO DE TECNOLOGÍAS DE VIRTUALIZACIÓN PARA EL DESPLIEGUE DE APLICACIONES CON ARQUITECTURA DE MICROSERVICIOS
5	TEMOCHE GOMEZ LENNIN BILLEY	DESARROLLO DE UN MÉTODO PARA DETECTAR CON EFICIENCIA LAS VULNERABILIDADES INFORMÁTICAS DE ATAQUE CROSS-SITE SCRIPTING UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO
6	CASTRO MEDINA MIGUEL ANGEL	IMPLEMENTACIÓN DE UNA METODOLOGÍA AD HOC DE GESTIÓN DE LA SEGURIDAD DE LA INFORMACIÓN PARA UNA EMPRESA EDITORA DE DIARIO REGIONAL PERUANO
7	MURO ESPINOZA JUAN JOSE	DESARROLLO DE UNA METODOLOGÍA AD HOC DE GESTIÓN DE LA SEGURIDAD DE LA INFORMACIÓN PARA UN INSTITUTO SUPERIOR PEDAGÓGICO PERUANO
8	DIAZ ZAVALA ROXANA KARINA FRIAS VASQUEZ LADY	DESARROLLO DE UNA METODOLOGÍA AD HOC DE GESTIÓN DE LA SEGURIDAD DE LA INFORMACIÓN PARA UNA UNIDAD DE GESTIÓN EDUCATIVA PERUANA
9	CARRASCO BORDA APARICIO	DESARROLLO DE UN MODELO DE PROCESOS AD HOC PARA EL DESARROLLO DE SOFTWARE POR LICENCIA PARA UNA MYPE DE SERVICIOS DE TI BASADO EN ISO/IEC 29110
10	OTERO MORALES JAVIER LIZARDO AQUINO SOSA NOELIA STEPHANY	DESARROLLO DE UN MODELO DE PROCESOS BASADO EN NORMAS DE PEQUEÑAS ORGANIZACIONES PARA MEJORAR LA CONSTRUCCIÓN DE SOFTWARE EN UN ÁREA DE DESARROLLO DE GOBIERNO MUNICIPAL
11	CALDERON YNOÑAN PAMELA DEL CARMEN PRIETO NEIRA FRANCK ALBERSON	DESARROLLO DE UN MÉTODO BAJO EL ENFOQUE ÁGIL EN ENTORNOS DE EXPERIENCIA DE USUARIO UI/UX PARA ASEGURAR LA USABILIDAD WEB
12	FLORES TINEO HUGO GALVANI DOLORIER POMA RONY RAUL	EVALUACIÓN DE LA USABILIDAD EN ENTORNOS VIRTUALES DE APRENDIZAJE PARA USUARIOS DE LAS ZONAS RURALES DEL PERÚ UTILIZANDO LA NORMA ISO/IEC 25010
13	CHANCAFE CASTRO JULIO JOEL	DESARROLLO DE UN MODELO DE PROCESOS AD HOC PARA EL DESARROLLO DE SOFTWARE PARA UNA MUNICIPALIDAD BASADO EN ISO/IEC 29110
14	SALAZAR DAVILA GIANFRANCO STEVEN	COMPARACIÓN DE TÉCNICAS DE VALIDACIÓN DE REQUISITOS DE SOFTWARE PARA MEDIR LA INFLUENCIA EN EL ÉXITO DE LOS PROYECTOS DE DESARROLLO EN PEQUEÑAS EMPRESAS PERUANAS
15	RIOJA MESIA CHARLES SEGUNDO FERNANDEZ RIOJA JUAN NICANOR	IMPLEMENTACIÓN DE UN MODELO DE GESTIÓN DE INCIDENCIAS BASADO EN ITIL PARA MEJORAR EL SERVICIO DE TI EN UNA MUNICIPALIDAD DISTRITAL DE LA REGIÓN LAMBAYEQUE
16	ALFARO PAJARES JUAN PEDRO	EVALUACIÓN DEL DESEMPEÑO DE PROCESOS DE NEGOCIO GESTIONADOS POR BPM EN UNA EMPRESA CONSTRUCTORA PERUANA
17	MONSALVE FERNANDEZ LENIN ESTALIN	IMPLEMENTACIÓN DE UN MODELO DE GESTIÓN DE SERVICIOS DE TI BASADO EN ITIL PARA MEJORAR LA GESTIÓN DE LOS SERVICIOS DE LA DIRECCIÓN DE TECNOLOGÍA DE UN GOBIERNO REGIONAL PERUANO
18	PEREZ CAMPOS DE QUIROZ BETTY MAGALY	EVALUACIÓN DEL DESEMPEÑO DE PROCESOS DE NEGOCIO GESTIONADOS POR BPM EN UNA MICRO EMPRESA PERUANA DESARROLLADORA DE SOFTWARE
19	MONTJOY PITA BRUNO	DESARROLLO DE UN SISTEMA DE RECOMENDACIÓN AUTOMÁTICA PARA EL TRATAMIENTO DE LAS PLAGAS EN CULTIVOS DE ARROZ DE LAS VARIEDADES QUE SE PRODUCEN EN LA REGIÓN LAMBAYEQUE
20	CRUZ FLORES JOSE ANTONIO CHAVEZ ANGULO GERMAN NEPTALI	IMPLEMENTACIÓN DE ARQUITECTURA EMPRESARIAL BASADO EN METODOLOGÍA ÁGIL PARA ALINEAR LAS TECNOLOGÍAS DE INFORMACIÓN CON LOS OBJETIVOS DE NEGOCIO DE UN ESTABLECIMIENTO PERUANO DE SALUD BUCAL

UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

RESOLUCIÓN N°0445-2021/FIAU-USS

Pimentel, 27 de mayo de 2021

N°	AUTOR (ES)	TEMA DE TESIS
21	PISFIL CORONADO JOSE LUIS FELIPE	IMPLEMENTACIÓN DE ARQUITECTURA EMPRESARIAL BASADA EN METODOLOGÍA ÁGIL PARA ALINEAR TI CON LOS PROCESOS DE NEGOCIO EN UNA EMPRESA CONSTRUCTORA PERUANA DE OBRAS CIVILES
22	ABAD HERRERA JOHNNY RENSO TEPE ESPINOZA LUIS RAMON	IMPLEMENTACIÓN DE ITIL V4 PARA MEJORAR LOS SERVICIOS DE TI EN EL CENTRO DE SISTEMAS DE INFORMACIÓN DE UNA UNIDAD DE GESTIÓN EDUCATIVA LOCAL PERUANO
23	URRUTIA VASQUEZ MIGUEL JULCA ROJAS ALEX ROGELIO	DESARROLLO DE UN MÉTODO DE IDENTIFICACIÓN AUTOMÁTICA DE ATAQUES SPOOFING DE ENVENENAMIENTO ARP EN LA SUPLANTACIÓN DE IDENTIDAD EN REDES LAN
24	SANCHEZ CELADA ERLIN FERNANDEZ ROMAN ISMAEL	COMPARACIÓN DE ARQUITECTURAS DE IDS HÍBRIDO PARA LA IDENTIFICACIÓN DE ATAQUES DE DOS EN LOS SERVIDORES WEB DE UNA MUNICIPALIDAD PROVINCIAL PERUANA
25	PERALES CHAVEZ JEFFERSON ADRIAN	IMPLEMENTACIÓN DE UN MODELO DE ARQUITECTURA DE INDUSTRIA 4.0 PARA MEJORAR LA INTEROPERABILIDAD ENTRE SISTEMAS DE UNA EMPRESA PERUANA
26	MAGALLANES CARBAJAL KENSER	EVALUACIÓN DE LA EFICIENCIA DE LOS ALGORITMOS DE CRIPTOGRAFÍA PARA CUMPLIR CON LOS NIVELES DE SEGURIDAD DE DATOS DE UNA EMPRESA FINANCIERA PERUANA
27	RACCHUMI LECCA JESÚS MANUEL	DESARROLLO DE UN MIDDLEWARE PARA MEJORAR LA COMUNICACIÓN ENTRE DOS INTERFACES DE LMS Y CRM EN EL PROCESO DE REGISTRO Y EMISIÓN DE CREDENCIALES DE USUARIOS
28	CASTRO QUESQUEN JAIME ELTON	COMPARACIÓN DE ALGORITMOS DE CIFRADO DE DATOS EN EL ASEGURAMIENTO DE VIDEO LLAMADA SOBRE REDES IP
29	PEREZ DIAZ NEILER WILTER CHINCHAY MALDONADO JORGE OBED	IMPLEMENTACIÓN DE TECNOLOGÍA SANDBOX PARA PROTEGER DE ATAQUES RANSOMWARE EN UNA RED INFORMÁTICA LOCAL DE UNA ENTIDAD FINANCIERA
30	MOSCOSO PAREDES ANIBAL	DISEÑO DE UN MODELO DE ARQUITECTURA DE SEGURIDAD DE BAJO COSTO PARA REFORZAR LA SEGURIDAD DE LA RED DEL HOGAR ANTE ATAQUES INFORMÁTICOS
31	MARTINEZ CUMPA JORGE JOSE	EVALUACIÓN DE FACTIBILIDAD DE USO DE TECNOLOGÍA WIRELESS 5GHZ PARA PROPORCIONAR SERVICIOS DE COMUNICACIÓN INALÁMBRICA EN LOS CENTROS POBLADOS RURALES DE LA REGIÓN LAMBAYEQUE
32	CAMPOS BARRERA SANDRO PAUL PASTOR OLIVA CESAR AUGUSTO	IMPLEMENTACIÓN DE UN MÉTODO DE CLASIFICACIÓN PARA DETECTAR LA DESERCIÓN DE ESTUDIANTES DE LA CARRERA DE INGENIERÍA DE INDUSTRIAS ALIMENTARIAS DE UNA UNIVERSIDAD NACIONAL PERUANA BASADO EN APRENDIZAJE DE MAQUINA
33	PICON VASQUEZ ANGEL GABRIEL CESPEDES SALAZAR JUAN CARLOS	DESARROLLO DE UN MÉTODO DE CLASIFICACIÓN AUTOMÁTICA BASADA EN TÉCNICAS ESTADÍSTICAS Y DE MACHINE LEARNING PARA CLASIFICAR A LOS POSTULANTES DE ACUERDO AL PERFIL DE TRABAJO DE UN CALL CENTER
34	MIÑANO SANCHEZ CARLOS JOHNY	COMPARACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA DESCUBRIR INFORMACIÓN RELEVANTE DE VENTAS DE UNA MYPE COMERCIAL
35	MARTOS PAREDES JOEL HAROLD VILLAZON SOSA JAIR AUGUSTO	IMPLEMENTACIÓN DE UN MODELO DE PROCESOS DE SEGURIDAD DE LA INFORMACIÓN PARA UNA PYME PERUANA BASADO EN LA NORMA ISO/IEC 27005 Y LA METODOLOGÍA OCTAVE-S
36	QUISPE PUEMAPE LUIS ALONSO	IMPLEMENTACIÓN DE UN SISTEMA DE GESTIÓN DE SEGURIDAD DE LA INFORMACIÓN APLICANDO LA NORMA ISO/IEC 27001:2014 EN UNA EMPRESA PERUANA DE TELECOMUNICACIONES
37	CHUCO AGUILAR GERSON RAUL	IMPLEMENTACIÓN DE UN SISTEMA DE GESTIÓN DE SEGURIDAD DE LA INFORMACIÓN BASADA EN ISO/IEC 27001 PARA MEJORAR EL NIVEL DE SEGURIDAD DE LOS ACTIVOS DE INFORMACIÓN EN UNA EMPRESA CONSTRUCTORA DE OBRAS CIVILES
38	CAJUSOL ROJAS JOSE DEL CARMEN	IMPLEMENTACIÓN DE UNA PLATAFORMA WEB PARA LA PLANIFICACIÓN Y MONITOREO DE RUTAS DE RECOJO DE RESIDUOS SÓLIDOS DE UN MUNICIPIO DE LA REGIÓN LAMBAYEQUE
39	VALLEJOS RAMOS FERNANDO RAFAEL	DESARROLLO DE UN MÉTODO DE OPTIMIZACIÓN DE USO DE TELA EN EL PROCESO DE ELABORACIÓN DE PRENDAS TEXTILES DE MICROEMPRESAS PERUANAS
40	REQUEJO NAVARRO JERSONS EXFRANSHER	EVALUACIÓN DE ALGORITMOS CRIPTOGRÁFICOS PARA MEJORAR SEGURIDAD EN UNA RED PRIVADA VIRTUAL

Facultad de Ingeniería, Arquitectura y Urbanismo

UNIVERSIDAD SEÑOR DE SIPÁN S.A.C.

Anexo 2. Carta de aceptación de la institución para la recolección de datos.

COMERCIAL DAMIAN E.I.R.L
MZA. "K" LOTE. 1 - 4 - URB. SANTA MARIA - CHICLAYO - TELF: 074 - 320413
RUC. N°20487443973

Chiclayo, 17 de Junio de 2021

Dr. Mario Fernando Ramos Moscol
Decano De La Facultad De Ingeniería, Arquitectura y Urbanismo
De la Universidad Particular Señor de Sipán

De mi consideración:

Es sumamente grato expresarle mi más cordial saludo a nombre propio y a nombre de quienes conformamos el Empresa Comercial Damián EIRL. Valga la ocasión, para expresarle que en virtud de la carta de presentación hemos recibido al Estudiante: Miñano Sánchez Carlos Johny de código Universitario 2130816458 y DNI Nº 18110843. A quien estamos autorizando para que recoja la información pertinente para el desarrollo de la INVESTIGACION del tema Comparación de Técnicas de Minería de Datos para descubrir Información Relevante de Ventas de una Mype Comercial

Agradecido por su gentil atención, me suscribo de usted, exteriorizándole las muestras de mi consideración más distinguida.

Atentamente,

COMERCIAL DAMIAN E.I.R.L.
Roberto Damian Bances
GERENTE GENERAL

Anexo 3. Instrumentos de recolección de datos, con su respectiva validación de los instrumentos.

Formato para informe de consumo de CPU.

Consumo de CPU	
Ítem	**Valor**
Uso	
Velocidad	
Procesos	
Subprocesos	
Tiempo	

. Formato de informe de consumo de memoria.

Consumo de Memoria	
Ítem	**Valor**
Uso	
Disponibilidad	
Confirmada	
En caché	
Tiempo	

Formato para informe de promedio de tiempo de respuesta.

Promedio de Tiempo de respuesta	
Ítem	**Valor**
Velocidad	
Tiempo	
CPU	
Memoria	
Disco	

Formato para el registro de matriz de confusión.

		Resultados Clasificación	
		Vta. Boletas	Vta Facturas
Resultados Reales	Vta.s Boletas		
	Vtas Factura		

Ítem	**Valor**
Verdadero positivo (VP)	
Falso Positivo (FP)	
Verdadero Negativo (VN)	
Falso Negativo (FN)	

Ítem	**Valor**
Prom. Tiempo rpta.	
Grado Consumo CPU	
Grado Consumo de memoria	
Error cuadrático	
Precisión	
Desviación Absoluta media	

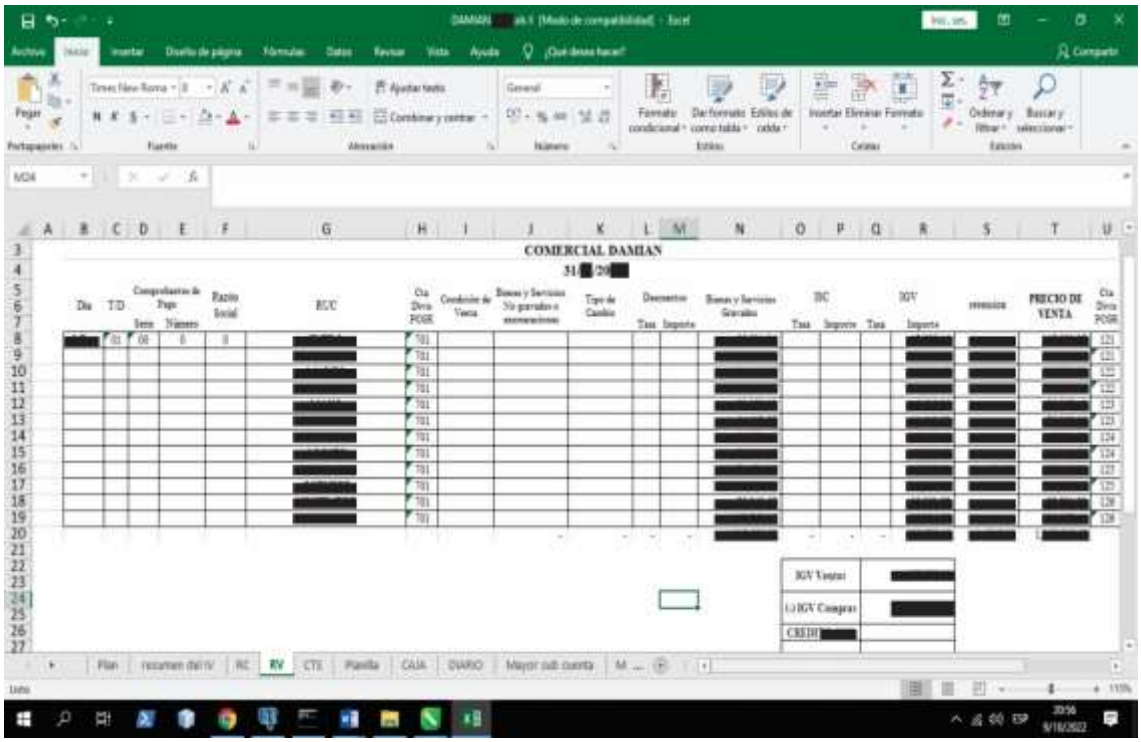
Verdadero Positivo (VP)		**Falso Positivo (FP)**	
Ítem	**Valor**	**Ítem**	**Valor**
Realidad		Realidad	
Predicción del modelo AA		Predicción del modelo AA	
Numero de resultados		Numero de resultados	
Falso Negativo (FN)		**Verdadero Negativo (FP)**	
Ítem	**Valor**	**Ítem**	**Valor**
Realidad		Realidad	
Predicción del modelo AA		Predicción del modelo AA	
Numero de resultados		Numero de resultados	

Anexo 4.: Listado de las técnicas de minería de datos

Ítem	Técnica	Descripción	Necesidades de información de pymes	
			Proyección de ventas	Selección de productos
1	EDM Funda-mental	Técnica que desarrolla métodos para la exploración de datos en el ámbito escolar que estudiar cuestiones educativas. Anoopkumar, M. & Rahman, A. (2016)		
2	Classifica-tion	Técnica que consiste en dividir los objetos y asignarles categorías exhaustivas y exclusivas conocidas por las clases; es decir asignarse cada objeto a una clase no a más de una. Bramer, M. (2016).	X	X
3	Cluste-ring	Técnica para organizar un grupo de datos con características similares, a finde hallar la estructura dentro de un conjunto de datos. Dinov, I. (2018).	X	X
4	Associa-tion Rule Mining	Técnica que se emplea con frecuencia para examinar transacciones. Comportamientos que permiten a obtener a un resultado correlacionado. Bell, J. (2015).	X	X
5	Sequentia l Mining	Técnica que se emplea en la minería de patrones secuenciales frecuentes de bases de datos de secuencias. Islam, M., Rafi, M. , Azad, A. & Ovi, J. (2021)		
6	Text Mining	Combina el componente bibliométrico con minería de textos y análisis de contenido. Permiten tener visión general de aportes científicos actuales, relaciones y tendencia. Sardo, F., Pinho, C. & Saur, I. (2021)		
7	Interactive Mining	Modelo de percepción y minería de patrones bajo el soporte máximo y mínimo de múltiples segmentos de acuerdo a las condiciones establecidas. Xie, Y. (2021)		

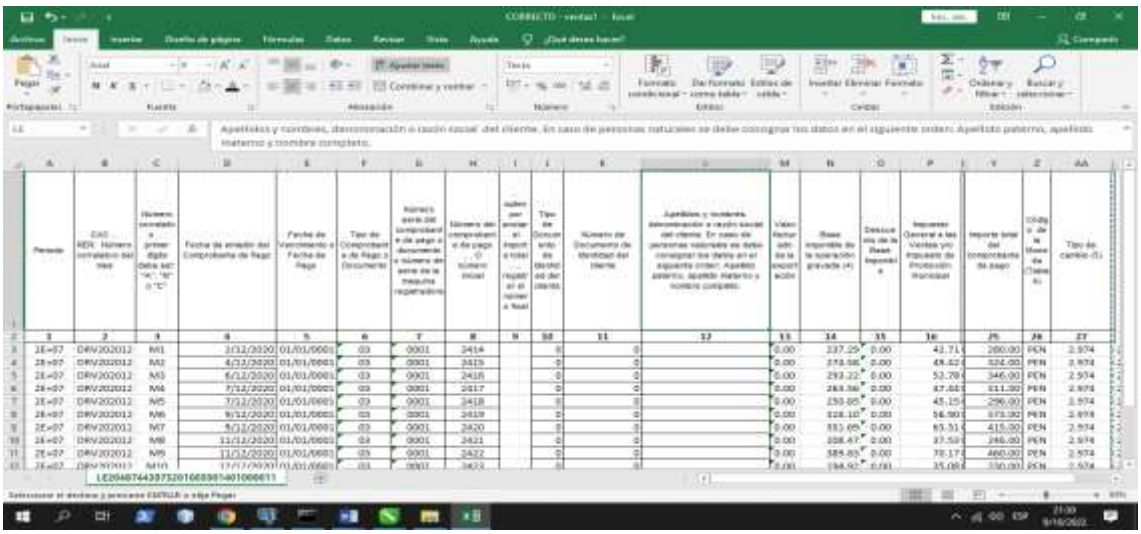
Item	Técnica	Descripción	Necesidades de información de pymes	
			Proyección de ventas	Selección de productos
8	**Temporal Mining**	Técnica que procesa ingente volumen de datos espacio-temporales de dominios como ciencias climáticas, ciencias sociales, economía, neurociencia, epidemiología, transporte, salud móvil, calidad del aire y ciencias. Parrella, M., Albano, G., Perna, C., & La Rocca, M. (2021)	**X**	**X**
9	**Neural Network**	Técnica que permite clasificar y estimar valores basados en neuronas artificiales. procesa información en estado dinámico a imputs externas. Bell, J. (2015).	**X**	**X**
10	**Distributed Data Mining**	Técnica Minería de datos para minar sobre fuentes de datos descentralizadas su objetivo es consolidar q el conocimiento local se fusiona para descubrir el conocimiento global. Devi, S. G. (2014)		
11	**Web Mining**	Ocupa de la minería de información en Internet usualmente útil en los dominios del comercio electrónico. Yadao, S., Babu, A., Janarthanan, M. & Bhaumik, A. (2021)		
12	**Regression Analysis**	Predice el valor de una variable categórica en relación a las variables independientes con base a la extracción de datos e identificando patrones, tendencias, etc. Award, M. & Khana, R. (2015).	**X**	**X**
13	**Correlation Analysis**	Técnica que permite entender relación de variables aleatorias X e Y cuyo coeficiente de correlación en encuentra entre -1 y +1 y si es cercano a 0 es débilmente correlacionado. Aggarwal (2015).		
14	**Statistical Methods**	Técnica que aplica los métodos estadísticos avanzados y hace uso de la minería de datos para descubrir modelos y patrones. Abu-Bader, S. (2021)		
15	**Visualisation Analysis Methods**	Técnicas de procesamiento digital de imágenes extraer información de imágenes, pero también para administrar y visualizar grandes conjuntos de datos. Masson, E., & Olesen, C. (2021)		

Anexo 5: Preparar conjunto de datos



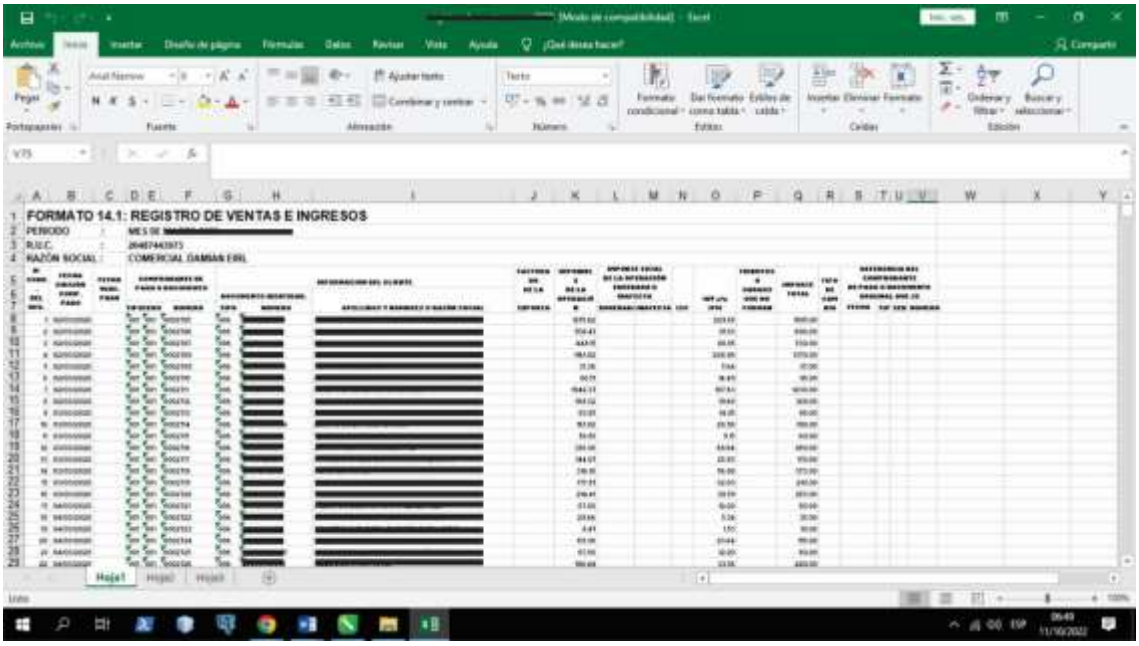
Registro de transacciones de ventas de Comercial Damián E.I.R.L.

Fuente: Area de Contabilidad Comercial Damian



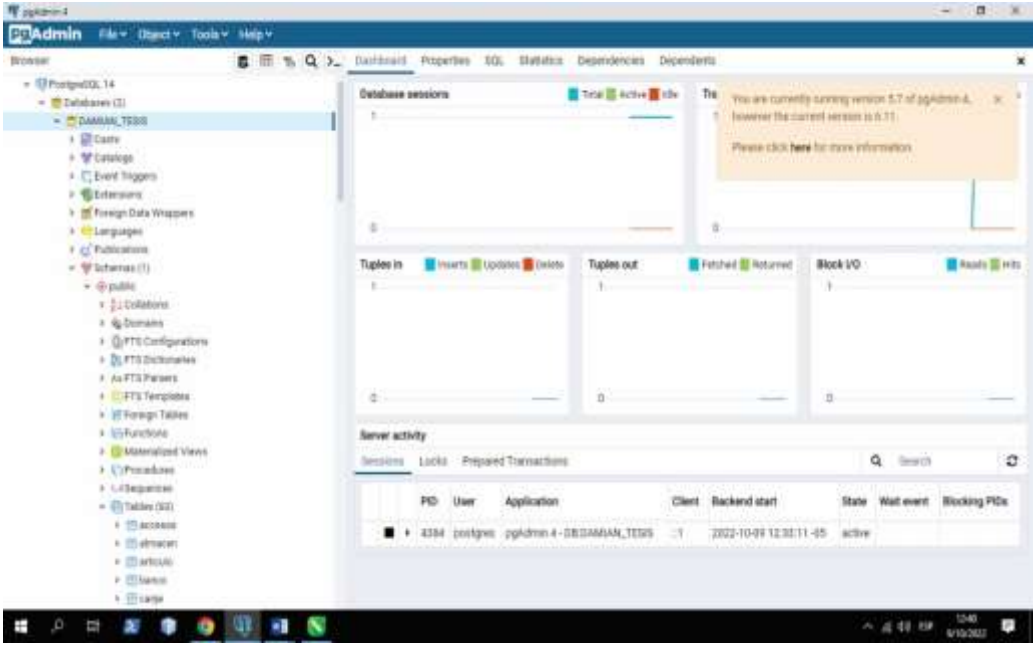
Archivos de registros de ventas

Fuente; Área de contabilidad (Ventas históricas)

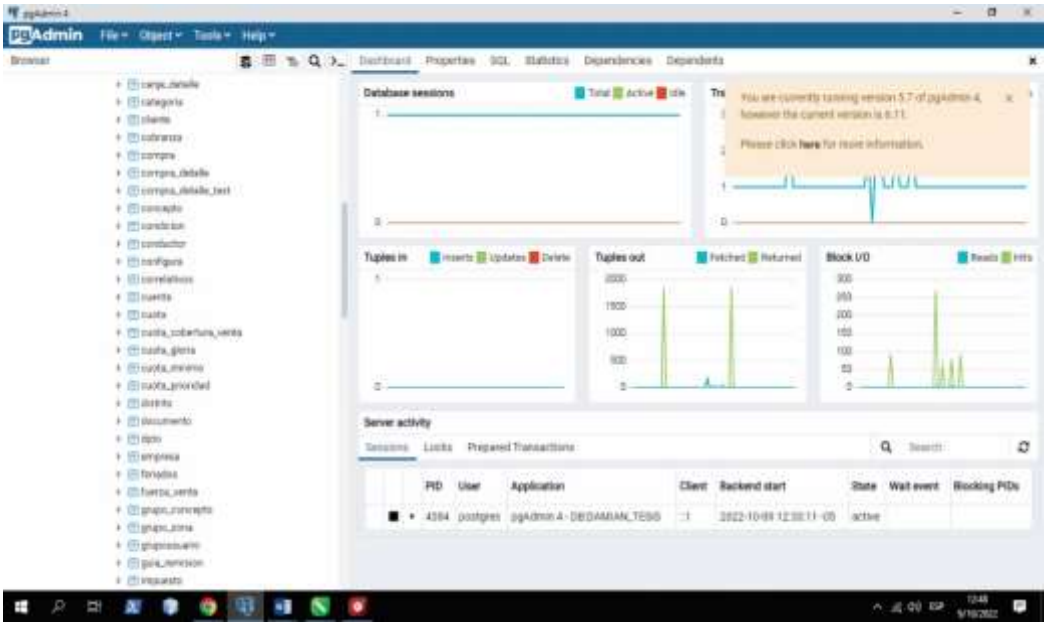


Registro de ventas

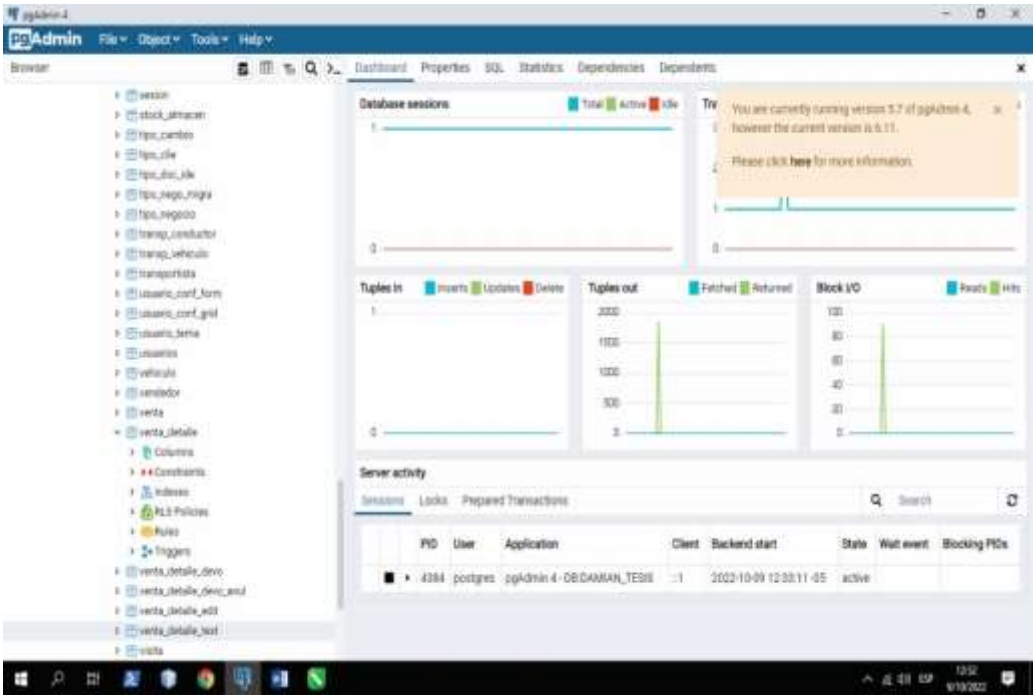
Fuente : Área de Contabilidad



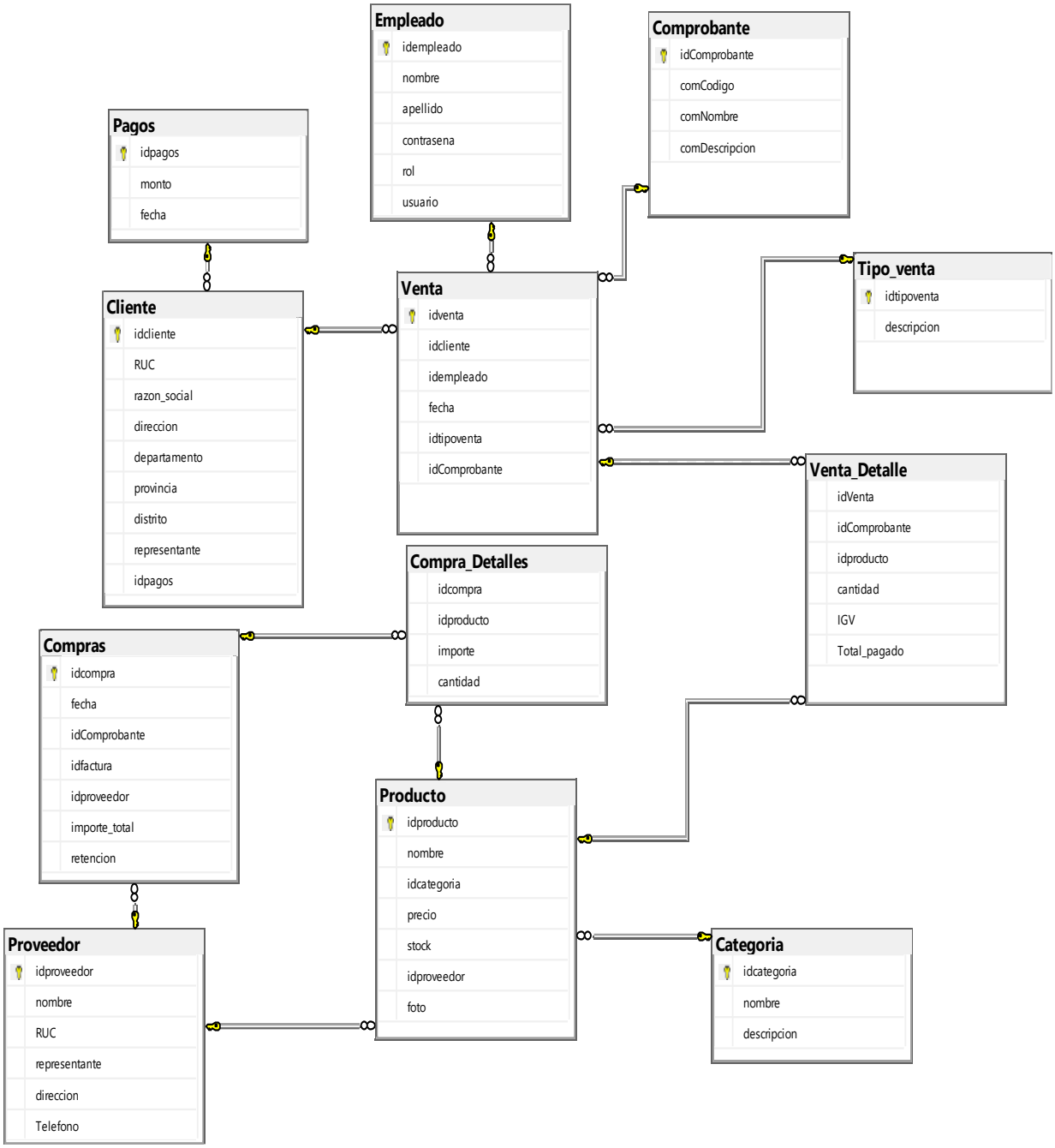
Base de Datos de la empresa Damián EIRL (renombrada para el estudio) y sus tablas.



Base de Datos de la empresa Damián EIRL (Tablas)

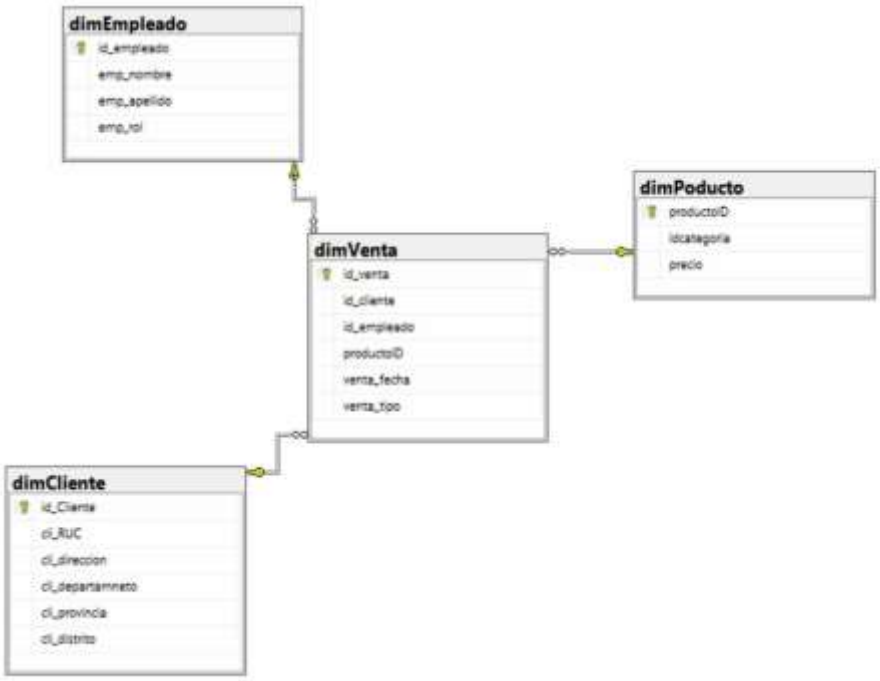


Base de Datos de la empresa Damián EIRL (Tablas)



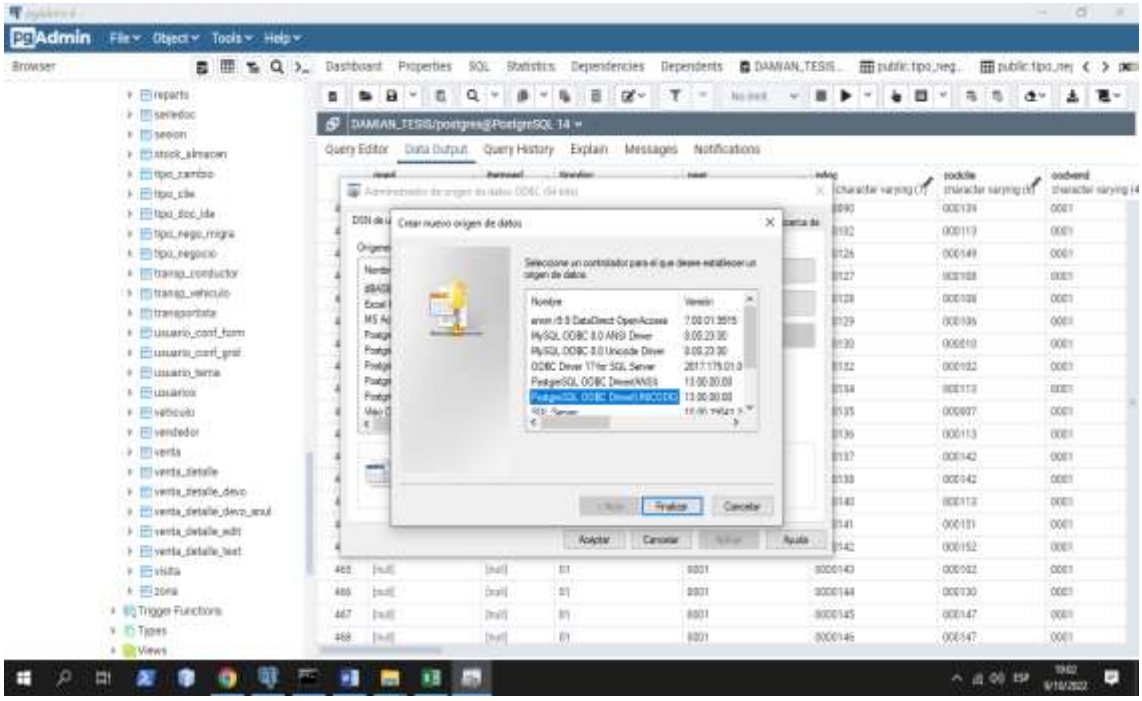
Esquema de la base De datos

Fuente: Elaboración propia

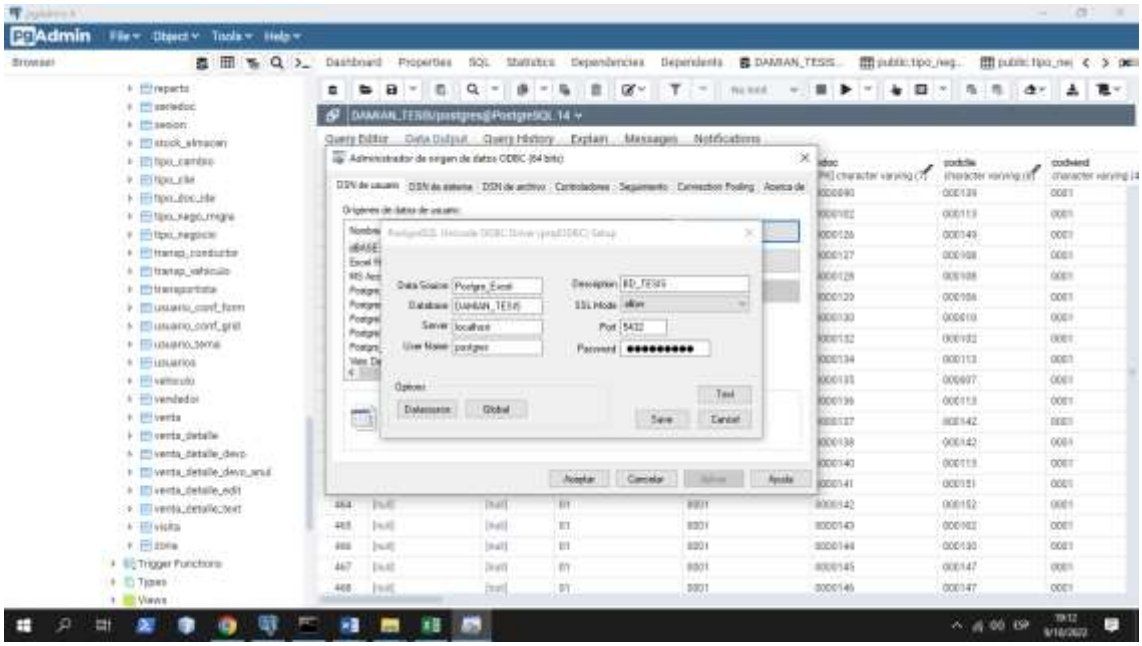


Esquema ETL

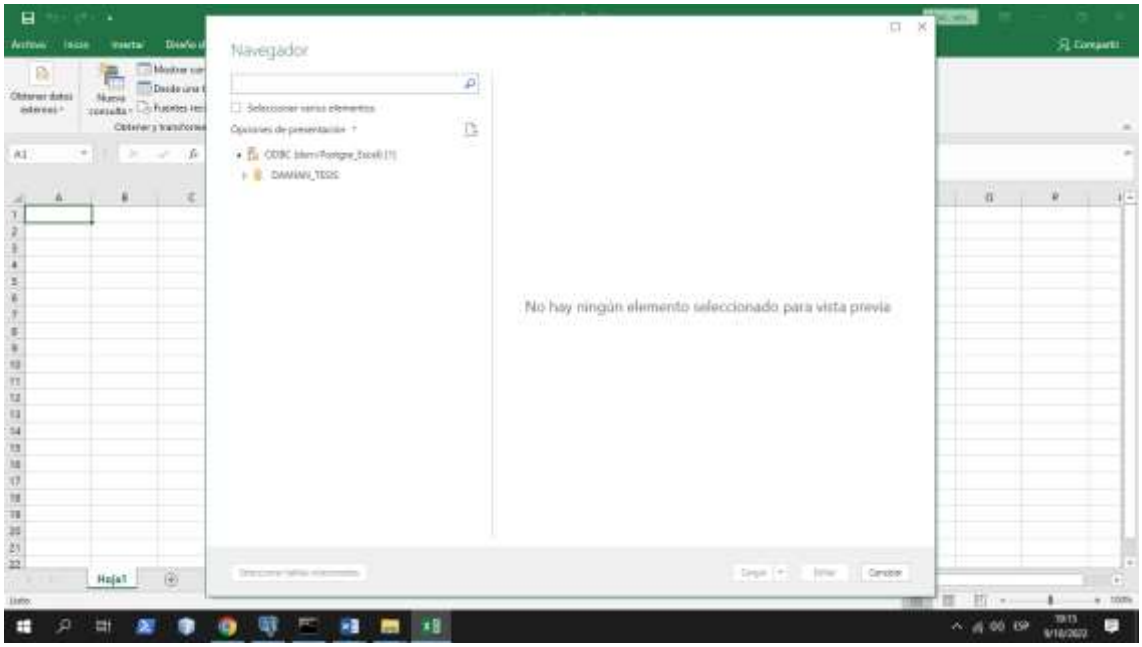
Fuente: Elaboración propia



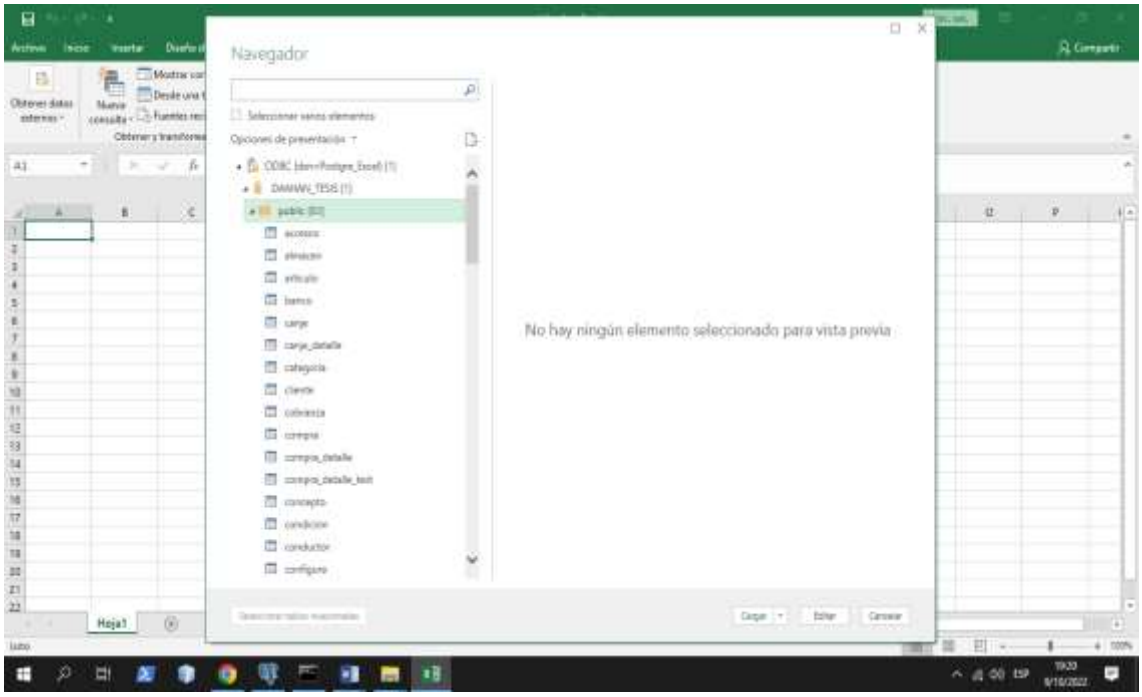
Exportando a Excel la tabla requerida



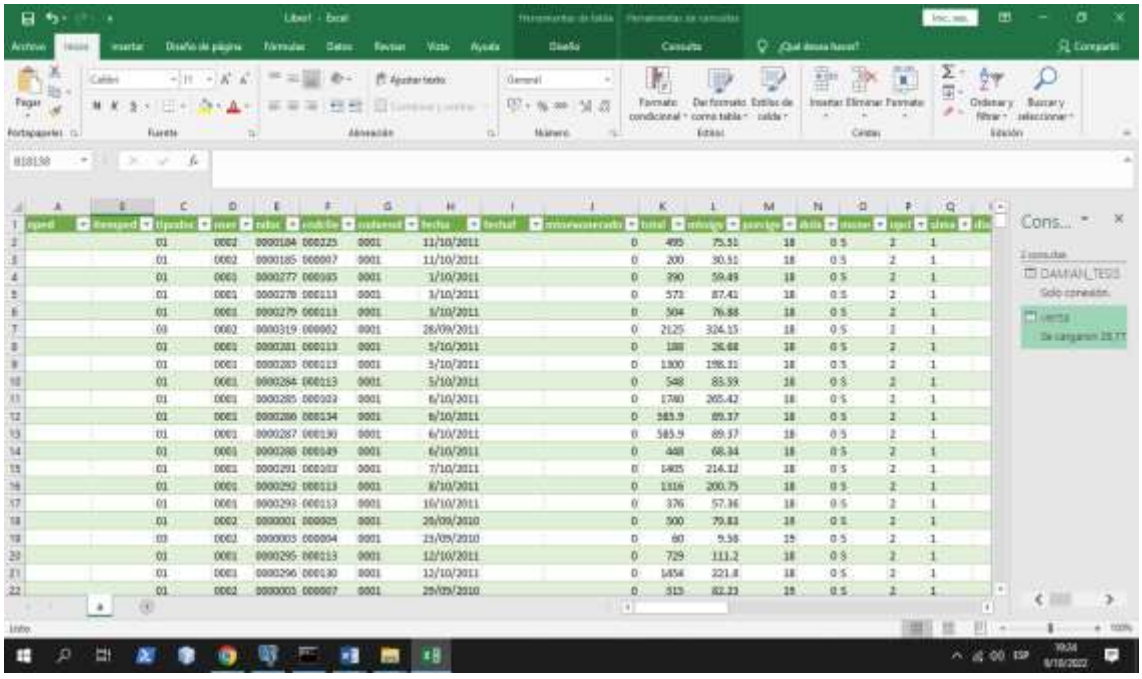
Probamos la conexión



Aparece esta ventana El nombre de origen de datos



Muestra las tablas contenidas en la base de dato Damian_Tesis



Cargamos las tablas que nos interesa para el estudio

Aquí se hace un análisis de los datos y se descarta datos irrelevantes no necesarios para el estudio teniendo en cuenta los archivos fuentes del área de contabilidad, comparando y descartando algunas columnas.

Anexo 6.: Estructura del data set generado.

	TIPO	DENOMINACION	DESCRIPCION	SUBTOTAL	IGV	TOTAL	Year	Month	Day
TIPO	1.000000	0.317673	0.083223	0.249854	0.237745	0.248695	-0.033000	0.056173	0.100103
DENOMINACION	0.317673	1.000000	0.011669	0.068295	0.064502	0.067906	0.023931	0.045090	0.009227
DESCRIPCION	0.083223	0.011669	1.000000	0.090275	0.088337	0.090224	0.171376	0.027141	-0.001145
SUBTOTAL	0.249854	0.068295	0.090275	1.000000	0.979222	0.999533	-0.004684	0.000817	-0.008254
IGV	0.237745	0.064502	0.088337	0.979222	1.000000	0.984962	0.008376	-0.001994	-0.004806
TOTAL	0.248695	0.067906	0.090224	0.999533	0.984962	1.000000	-0.002728	0.000396	-0.007756
Year	-0.033000	0.023931	0.171376	-0.004684	0.008376	-0.002728	1.000000	-0.168146	-0.097084
Month	0.056173	0.045090	0.027141	0.000817	-0.001994	0.000396	-0.168146	1.000000	0.039805
Day	0.100103	0.009227	-0.001145	-0.008254	-0.004806	-0.007756	-0.097084	0.039805	1.000000
TIPO_SERIE	-0.931464	-0.325531	-0.094550	-0.263949	-0.250843	-0.262677	0.064450	-0.067986	-0.100638

Anexo 7: Código Python

```
# from  sklearn.linear_model  import LogisticRegression, SGDClassifier
from sklearn.ensemble import RandomForestClassifier from sklearn.svm import
SVC
from sklearn.naive_bayes import GaussianNB, MultinomialNB

from sklearn.tree import DecisionTreeClassifier from sklearn.neighbors import
KNeighborsClassifier from sklearn.neural_network import MLPClassifier from
sklearn.linear_model import LinearRegression from sklearn.cluster import KMeans
from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report from sklearn.metrics import
confusion_matrix
from sklearn.metrics import mean_squared_error as MSE

from sklearn.metrics import mean_absolute_percentage_error as MAPE

from sklearn.metrics import accuracy_score, f1_score, roc_auc_score


from sklearn.model_selection import train_test_split from sklearn.model_selection
import cross_val_score

from sklearn.pipeline import Pipeline from sklearn.decomposition import PCA
from mpl_toolkits.mplot3d import Axes3D

import sklearn.metrics as sm import pandas as pd
import numpy as np import seaborn as sb
import matplotlib.pyplot as plt import time
import psutil


In [2]:

%matplotlib inline plt.rcParams['figure.figsize'] = 7,4
In [3]:

df = pd.read_csv('data.csv') del(df['Unnamed: 0']) del(df['VALOR_UNITARIO'])
del(df['PRECIO_UNITARIO']) del(df['CANTIDAD'])
df

Out[3]:

5565 rows × 10 columns

In [5]:

data1, data2 = train_test_split(df, test_size=0.2, random_state=6)
```

In [6]:

```
x_train = np.array(data1.drop(['TIPO_SERIE'], 1))
y_train = np.array(data1['TIPO_SERIE'])

x_test = np.array(data2.drop(['TIPO_SERIE'], 1))
y_test = np.array(data2['TIPO_SERIE'])
```

REGRESIó N LOGÍSTICA

In [7]:

```
# REGRESIóN LOGÍSTICA
model_1 = LogisticRegression( solver="liblinear", max_iter=1000, penalty="l1",
)
```

In [8]:

```
# Tiempo de entrenamiento inicio = time.time() model_1.fit(x_train,y_train) fin =
time.time() print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3) print('Grado de consumo de
CPU        :',psutil.cpu_percent()) print('Consumo de memoria fisica       :
{}'.format(memory))
```

```
Tiempo:  0.05784320831298828
Grado de consumo de CPU        : 28.5
Consumo de memoria fisica      : 6.920780181884766
```

In [9]:

```
predictions = model_1.predict(x_train)
print(predictions)
```

```
[1 0 1 ... 1 0 1]
```

In [10]:

```
# muestra la presicion en el entrenamiento accuracyTrain =
model_1.score(x_train,y_train) accuracyTest = model_1.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

```
accuracyTrain:  0.9957322551662174  accuracyTest:  0.9991015274034142
```

In [11]:

```
predictions = model_1.predict(x_test)
print(accuracy_score(y_test, predictions))
```

```
0.9991015274034142
```

In [13]:

```
ypred = model_1.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

```
[[386   0]
[  1 726]]
```

In [14]:

```
# Tiempo de prediccion
#inicio = time.time()
#print(model_1.predict([[225,11,10,2011,495.0,75.51,19.49]]))
#fin = time.time()
#print(fin-inicio)
```

In [15]:

```
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

```
MSE:  0.0008984725965858042
MAPE:  0.0008984725965858042
```

REGRESION LINEAL

In [16]:

```
model_regresion = LinearRegression()
model_regresion
```

Out[16]:

```
LinearRegression()
```

In [17]:

```
# Tiempo de entrenamiento inicio = time.time() model_regresion.fit(x_train,y_train)
fin = time.time()
print('Tiempo: ', fin-inicio)

memory = psutil.virtual_memory().used / (1024.0 ** 3) print('Grado de consumo de
CPU        :',psutil.cpu_percent()) print('Consumo de memoria fisica       :
{}'.format(memory))
```

```
Tiempo:  0.0019958019256591797

Grado de consumo de CPU        : 85.3

Consumo de memoria fisica      : 6.923931121826172
```

In [18]:

```
accuracyTrain = model_regresion.score(x_train,y_train)

accuracyTest = model_regresion.score(x_test,y_test)

print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

```
accuracyTrain:  0.8533106778050024  accuracyTest:  0.9401372221091934
```

In [19]:

```
ypred = model_1.predict(x_test)

cm = confusion_matrix(y_test, ypred)

print(cm)

[[386   0]

[  1 726]]
```

In [20]:

```
print('MSE: ', MSE(y_test, ypred))

print('MAPE: ', MAPE(y_test, ypred))
```

```
MSE:  0.0008984725965858042

MAPE:  0.0008984725965858042
```

REDES NEURONALES 1 capa

In [21]:

```
# REDES NEURONALES 1 capa

  model_2 = MLPClassifier( activation="relu", hidden_layer_sizes=(240),
max_iter=1000, verbose=False, solver="adam", learning_rate_init=0.008,
alpha=0.05,
  beta_1=0.7, tol=0.02, learning_rate="constant",
)
```

In [22]:

```
# Tiempo de entrenamiento inicio = time.time() model_2.fit(x_train,y_train) fin =
time.time() print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3) print('Grado de consumo de
CPU        :',psutil.cpu_percent()) print('Consumo de memoria fisica       :
{}'.format(memory))
```

```
Tiempo:  0.9943404197692871
Grado de consumo de CPU        : 100.0
Consumo de memoria fisica      : 6.922210693359375
```

In [23]:

```
# muestra la presicion en el entrenamiento accuracyTrain =
model_2.score(x_train,y_train) accuracyTest = model_2.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

```
accuracyTrain:  0.9764150943396226  accuracyTest:  0.9730458221024259
```

In [24]:

```
predictions = model_2.predict(x_test)
print(accuracy_score(y_test, predictions))
```

```
0.9730458221024259
```

In [25]:

```
# muestra un reporte de la clasificacion segun cada etiqueta
print(classification_report(y_test, predictions))
```

In [26]:

```
ypred = model_2.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

```
[[358  28]
[  2 725]]
```

In [27]:

```
# Tiempo de prediccion
#inicio = time.time()
#print(model_1.predict([[225,11,10,2011,495.0,75.51,19.49]]))
#fin = time.time()
#print(fin-inicio)
```

In [28]:

```
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

```
MSE:  0.026954177897574125
MAPE:  113298103833220.03
```

In [ ]:

REDES NEURONALES 3 capas

In [37]:

```
# REDES NEURONALES 3 capa
model_4 = MLPClassifier( activation="tanh", hidden_layer_sizes=(50, 20, 60),
max_iter=1000,
verbose=False, solver="adam", learning_rate_init=0.0008, alpha=0.003,
beta_1=0.07, tol=0.002, learning_rate="invscaling",
)
```

In [38]:

```
# Tiempo de entrenamiento inicio = time.time() model_4.fit(x_train,y_train) fin =
time.time() print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3) print('Grado de consumo de
CPU        :',psutil.cpu_percent()) print('Consumo de memoria fisica       :
{}'.format(memory))
```

```
Tiempo:  1.137993574142456
Grado de consumo de CPU        : 100.0
Consumo de memoria fisica      : 6.918354034423828
```

In [39]:

```
# muestra la presicion en el entrenamiento accuracyTrain =
model_4.score(x_train,y_train) accuracyTest = model_4.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

```
accuracyTrain:  0.8110961365678346  accuracyTest:  0.7879604672057502
```

In [40]:

```
predictions = model_4.predict(x_test)
print(accuracy_score(y_test, predictions))
```

```
0.7879604672057502
```

In [41]:

```
# muestra un reporte de la clasificacion segun cada etiqueta
print(classification_report(y_test, predictions))
```

In [42]:

```
ypred = model_4.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

```
[[243 143]
[ 93 634]]
```

In [43]:

```
# Tiempo de prediccion
#inicio = time.time()
#print(model_1.predict([[225,11,10,2011,495.0,75.51,19.49]]))
#fin = time.time()
#print(fin-inicio)
```

In [44]:

```
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

```
MSE:  0.21203953279424978
MAPE:  578629601719659.4
```

SUPER VECTOR MANCHINE (SVM)

In [45]:

```
# SUPER VECTOR MANCHINE (SVM)
model_5 = SVC(kernel="linear", probability=True)
```

In [46]:

```
# Tiempo de entrenamiento inicio = time.time() model_5.fit(x_train,y_train) fin =
time.time() print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3) print('Grado de consumo de
CPU        :',psutil.cpu_percent()) print('Consumo de memoria fisica       :
{}'.format(memory))
```

```
Tiempo:  73.42755484580994
Grado de consumo de CPU        : 25.1
Consumo de memoria fisica      : 6.917575836181641
```

In [47]:

```
# muestra la presicion en el entrenamiento accuracyTrain =
model_5.score(x_train,y_train) accuracyTest = model_5.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

```
accuracyTrain:  0.9952830188679245  accuracyTest:  0.9991015274034142
```

In [48]:

```
predictions = model_5.predict(x_test)
print(accuracy_score(y_test, predictions))
```

```
0.9991015274034142
```

In [49]:

```
# muestra un reporte de la clasificacion segun cada etiqueta
print(classification_report(y_test, predictions))
```

In [50]:

```
ypred = model_5.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

```
[[386   0]
[  1 726]]
```

In [51]:

```
# Tiempo de prediccion
#inicio = time.time()
#print(model_1.predict([[225,11,10,2011,495.0,75.51,19.49]]))
#fin = time.time()
#print(fin-inicio)
```

In [52]:

```
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

```
MSE:  0.0008984725965858042
MAPE:  0.0008984725965858042
```

BAYESIANO

In [53]:

```
# BAYESIANO
model_6 = GaussianNB()
```

In [54]:

```
# Tiempo de entrenamiento inicio = time.time() model_6.fit(x_train,y_train) fin =
time.time() print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3) print('Grado de consumo de
CPU        :',psutil.cpu_percent()) print('Consumo de memoria fisica       :
{}'.format(memory))
```

```
Tiempo:  0.0019943714141845703
Grado de consumo de CPU        : 19.4
Consumo de memoria fisica      : 6.9173736572265625
```

In [55]:

```
# muestra la presicion en el entrenamiento accuracyTrain =
model_6.score(x_train,y_train) accuracyTest = model_6.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

```
accuracyTrain:  0.9827044025157232  accuracyTest:  0.9829290206648698
```

In [56]:

```
predictions = model_6.predict(x_test)
print(accuracy_score(y_test, predictions))
```

```
0.9829290206648698
```

In [57]:

```
# muestra un reporte de la clasificacion segun cada etiqueta
print(classification_report(y_test, predictions))
```

In [58]:

```
ypred = model_6.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

```
[[367  19]
[  0 727]]
```

In [59]:

```
# Tiempo de prediccion
#inicio = time.time()
#print(model_1.predict([[225,11,10,2011,495.0,75.51,19.49]]))
#fin = time.time()
#print(fin-inicio)
```

In [60]:

```
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

```
MSE:  0.017070979335130278
MAPE:  76880856172542.16
```

ARBOL DE DECISIONES

In [61]:

```
# ARBOL DE DECISIONES
model_7 = RandomForestClassifier(
n_estimators=100, random_state=2016, min_samples_leaf=8
)
```

In [62]:

```
# Tiempo de entrenamiento inicio = time.time() model_7.fit(x_train,y_train) fin =
time.time() print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3) print('Grado de consumo de
CPU        :',psutil.cpu_percent()) print('Consumo de memoria fisica       :
{}'.format(memory))
```

```
Tiempo:  0.21543097496032715
Grado de consumo de CPU        : 11.4
Consumo de memoria fisica      : 6.91650390625
```

In [63]:

```
# muestra la presicion en el entrenamiento accuracyTrain =
model_7.score(x_train,y_train) accuracyTest = model_7.score(x_test,y_test)
```

```
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

```
accuracyTrain:  0.9970799640610961  accuracyTest:  0.9991015274034142
```

In [64]:

```
predictions = model_7.predict(x_test)
print(accuracy_score(y_test, predictions))
```

```
0.9991015274034142
```

In [65]:

```
# muestra un reporte de la clasificacion segun cada etiqueta
print(classification_report(y_test, predictions))
```

In [66]:

```
ypred = model_7.predict(x_test)
cm = confusion_matrix(y_test, ypred)
print(cm)
```

```
[[386   0]
[  1 726]]
```

In [67]:

```
# Tiempo de prediccion
#inicio = time.time()
#print(model_1.predict([[225,11,10,2011,495.0,75.51,19.49]]))
#fin = time.time()
#print(fin-inicio)
```

In [68]:

```
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

```
MSE:  0.0008984725965858042
MAPE:  0.0008984725965858042
```

NEIGHBORS

In [69]:

```
# NEIGHBORS
model_8 = KNeighborsClassifier(
n_neighbors=120, p=2, weights="uniform", algorithm="auto"
)
```

In [70]:

```
# Tiempo de entrenamiento inicio = time.time() model_8.fit(x_train,y_train) fin =
time.time() print('Tiempo: ', fin-inicio)
memory = psutil.virtual_memory().used / (1024.0 ** 3) print('Grado de consumo de
CPU        :',psutil.cpu_percent()) print('Consumo de memoria fisica       :
{}'.format(memory))
```

```
Tiempo:  0.008973360061645508
Grado de consumo de CPU        : 21.0
Consumo de memoria fisica      : 6.916492462158203
```

In [71]:

```
# muestra la presicion en el entrenamiento accuracyTrain =
model_8.score(x_train,y_train) accuracyTest = model_8.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , '' accuracyTest: '', accuracyTest)
```

```
accuracyTrain:  0.7942497753818508  accuracyTest:  0.779874213836478
```

In [72]:

```
predictions = model_8.predict(x_test)
print(accuracy_score(y_test, predictions))
```

```
0.779874213836478
```

In [73]:

```
# muestra un reporte de la clasificacion segun cada etiqueta
print(classification_report(y_test, predictions))
```

In [74]:

```
ypred = model_8.predict(x_test)
```

```
cm = confusion_matrix(y_test, ypred)
print(cm)
```

```
[[170 216]
[ 29 698]]
```

```
In [75]:
# Tiempo de prediccion
#inicio = time.time()
#print(model_1.predict([[225,11,10,2011,495.0,75.51,19.49]]))
#fin = time.time()
#print(fin-inicio)
```

```
In [76]:
print('MSE: ', MSE(y_test, ypred))
print('MAPE: ', MAPE(y_test, ypred))
```

```
MSE:   0.22012578616352202
MAPE:   874013943856268.8
```