



**FACULTAD DE INGENIERÍA, ARQUITECTURA Y
URBANISMO**

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

**ANÁLISIS COMPARATIVO DE TÉCNICAS DE
MINERÍA DE DATOS APLICADAS A BUSINESS
INTELLIGENCE**

**PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO
DE SISTEMAS**

Autor(a) (es):

Bach. Alvarez Gonzaga Braulio Ricardo

ORCID: <https://orcid.org/0000-0002-3902-788X>

Asesor(a):

Mg. Bravo Ruiz Jaime Arturo

ORCID: <https://orcid.org/0000-0003-1929-3969>

Línea de Investigación:

Infraestructura, Tecnología y Medio Ambiente

Pimentel – Perú 2021

**ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS
APLICADAS A BUSINESS INTELLIGENCE**

Bach. Alvarez Gonzaga Braulio Ricardo
Autor

Mg. Bravo Ruiz Jaime Arturo
Asesor

Dr. Vásquez Leyva Oliver
Presidente de Jurado

Mg. Díaz Vidarte Miguel Orlando
Secretario de Jurado

Mg. Bances Saavedra David Enrique
Vocal de Jurado

Dedicatorias

Al iniciar un nuevo proyecto, nos llenamos de ilusión y optimismo. La idea de ser capaces de alcanzar lo que nos proponemos es cautivadora y nos emociona el alma. Sin embargo, en el camino se presentan obstáculos que deben ser superados. A lo largo de este camino, me acompañaron personas que me motivaron a seguir adelante, perseguir mis metas y llenaron de confianza.

Mi familia fue el mayor aliento desde que decidí estudiar ingeniería de sistemas. Le dedico este trabajo a mi esposa, Diana Valdivia Dorregaray y a mis padres, Ricardo Alvarez García y Lupe Gonzaga Espino, por el ánimo permanente durante todo este tiempo, en el que con su amor y confianza se volvieron parte de este desafío y que hoy abre nuevos horizontes para seguir soñando y creciendo profesionalmente.

Agradecimientos

Mi agradecimiento a los profesionales que directa o indirectamente contribuyeron durante todo este proceso de investigación. A mis maestros de la escuela de Ingeniería de Sistemas, quienes pusieron a disposición su experiencia profesional y académica para formarnos con exigencia. Asimismo, agradezco a mis compañeros de estudio, con quienes tuve la oportunidad de aprender y compartir conocimiento. El apoyo solidario, optimismo y confianza fueron elementos que siempre estuvieron presentes y nos ayudaron a salir adelante.

Resumen

La toma de decisiones constituye un proceso de vital importancia para las universidades, siendo uno de los indicadores más importantes en sus sistemas business intelligence el rendimiento académico. No obstante, el crecimiento de los sistemas de información genera un reto para la gestión y procesamiento de grandes volúmenes de datos de los cuales se desea obtener información relevante. En ese sentido, la minería de datos ofrece una serie de técnicas que permite realizar este descubrimiento con un alto nivel de precisión. El presente trabajo titulado “ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS APLICADAS A BUSINESS INTELLIGENCE” tiene como objetivo general analizar comparativamente el rendimiento de técnicas de minería de datos aplicadas a soluciones business intelligence. El método propuesto inició con la selección de dos técnicas de minería de datos bajo el método no probabilístico con base a las técnicas de minería de datos disponibles y documentadas en diversas investigaciones. Posteriormente, se diseñó un método de aplicación conformado por cinco etapas: análisis y comprensión de las fuentes de datos, implementación de la base de datos en SQL Server, proceso ETL, implementación de los algoritmos de minería de datos a partir de los datos de entrada obtenidos del proceso business intelligence y procesamiento de datos. Los resultados evidenciaron que el modelo propuesto, el cual utilizó datos de entrada obtenidos de un proceso business intelligence obtuvo un rendimiento en cuanto a su precisión superior al 90% en ambas técnicas de minería de datos. Árbol de decisiones obtuvo 93.69% y Naive Bayes 93.67%. Asimismo, en cuanto al análisis de error, Naive Bayes fue la que mejor resultado obtuvo, obteniendo un error porcentual absoluto medio (MAPE) de 6.2%. La investigación concluye que las técnicas de minería aplicadas a datos obtenidos de un proceso business intelligence tienen muy buena precisión para la predicción del rendimiento académico y podrían ser utilizada en el análisis de otras variables académicas como la morosidad y la deserción, siendo la de mejor rendimiento Naive Bayes.

Palabras clave: Técnicas de minería de datos, Árbol de decisión, Naive Bayes, inteligencia de negocios, algoritmos, base de datos.

Abstract

Decision-making is a process of vital importance for universities, with academic performance being one of the most important indicators in their business intelligence systems. However, the growth of information systems creates a challenge for the management and processing of large volumes of data from which it is wanted to get relevant information. On this matter, data mining offers a series of techniques that allow this discovery to be made with a high level of accuracy. This research work entitled "COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES APPLIED TO BUSINESS INTELLIGENCE" has as its general objective to conduct a comparative analysis the performance of data mining techniques applied to business intelligence solutions. The proposed method began with the selection of two data mining techniques under the non-probabilistic method based on the available data mining techniques and documented in several researches. Afterwards, an application method formed by five stages was designed: analysis and understanding of data sources, implementation of the database in SQL Server, ETL process, implementation of data mining algorithms from input data obtained from the business intelligence process and data processing. The results showed that the proposed model, which used input data obtained from a business intelligence process, got a performance in terms of its greater precision than 90% in both data mining techniques. Decision tree obtained 93.69% and Naive Bayes 93.67%., regarding the error analysis, the best result was obtained by Naive Bayes, obtaining a mean absolute percentage error (MAPE) of 6.2%. In this way, the research concludes that the mining techniques applied to obtained data from a business intelligence process have very good accuracy for the prediction of academic performance and could be used in the analysis of other academic variables such as defaulting and dropping out, being the best performing Naive Bayes.

Keywords: Data mining techniques, Decision Tree, Naive Bayes, business intelligence, algorithms, database.

Índice

I. INTRODUCCIÓN	8
1.1. Realidad Problemática.	8
1.2. Trabajos previos.	10
1.3. Teorías relacionadas al tema.	16
1.4. Formulación del Problema.	47
1.5. Justificación e importancia del estudio.	47
1.6. Hipótesis.	48
1.7. Objetivos.	48
1.7.1. Objetivo general.	48
1.7.2. Objetivos específicos.	49
II. MATERIAL Y MÉTODO	50
2.1. Tipo y Diseño de Investigación.	50
2.2. Población y muestra.	51
2.3. Variables, Operacionalización.	52
2.4. Técnicas e instrumentos de recolección de datos, validez y confiabilidad.	55
2.5. Procedimiento de análisis de datos.	55
2.6. Criterios éticos.	56
2.7. Criterios de Rigor Científico.	57
III. RESULTADOS.	58
3.1. Resultados en Tablas y Figuras.	58
3.2. Discusión de resultados.	63
3.3. Aporte práctico.	65
IV. CONCLUSIONES Y RECOMENDACIONES	94
4.1. Conclusiones.	94
4.2. Recomendaciones.	96
REFERENCIAS.....	97
ANEXOS.	101

I. INTRODUCCIÓN

1.1. Realidad Problemática.

Desde inicios del siglo XXI, con la expansión progresiva del Internet, las organizaciones iniciaron un vertiginoso proceso de crecimiento. A partir de la expansión del internet e incremento de la información, es muy difícil encontrar pequeñas y medianas empresas que no cuenten con algún sistema informático, ya sea para automatizar todos los procesos de la empresa o solo aquellos que pertenecen a un departamento específico (Bustamante, Galvis-Lista, & Gómez, 2012). En ese sentido, todo se incrementó: clientes, productos, servicios, adquisiciones, etc. A nivel estratégico, las bases de datos (BD) son el insumo fundamental para alcanzar los objetivos de toda organización.

El crecimiento de las organizaciones a nivel mundial trae consigo el incremento de los datos que se generan producto de los procesos y las interacciones de sus distintos usuarios. Un estudio realizado por SEAGATE, empresa líder en almacenamiento de datos, mostraba que para el año 2022 las empresas generarán el 60% de la información en el mundo. Asimismo, la organización Global Data Protection Index reveló que entre los años 2016 y 2018 se incrementó en más del 500% los datos que gestionan las organizaciones. Claramente, se evidencia la necesidad de contar con las técnicas necesarias para analizar y procesar los datos, lo cual conlleva a la obtención de información significativa para sus usuarios. Esto contribuye en la toma de decisiones.

Para el usuario final, esta información es presentada en reportes, gráficos o tablas, lo cual le permite tomar decisiones estratégicas en función de los conocidos key performance indicator (KPI). Por este motivo, la inteligencia de negocios también denominada business intelligence (BI) es crucial para la gestión de toda empresa, ya que busca integrar los datos de diversas fuentes distribuidas. Asimismo, el mercado de desarrolladores de soluciones de inteligencia de negocios incrementó debido a la demanda por parte de las organizaciones (Chen, Baoran, & Yang, 2016).

Los desarrolladores de soluciones de business intelligence, cuentan con opciones metodológicas limitadas. Básicamente, el mercado es dominado por dos metodologías Kimball e Inmon. Siendo la primera la que posee mayor acogida, pues está diseñada para proyectos basados en data warehouse (Anastasios, Panos, & Alkis, 2013). Durante el proceso de modelado, se seleccionan las técnicas de minería de datos a utilizar, con el objetivo de realizar un análisis significativo de los datos. El uso de estas técnicas permite generar predicciones a partir del análisis de los datos (Gallego, Navarro, & Castillo, 2015). De acuerdo con ello, una decisión crucial consiste en la elección correcta de dicha técnica, la cual debe ser seleccionada con base a sus indicadores de rendimiento.

La minería de datos ayuda en los procesos de reingeniería y mejora de las metodologías para los procesos business intelligence, las cuales tienen impacto directo en la administración de la información para los procesos de decisión en las organizaciones. El correcto conocimiento de sus aplicaciones conlleva a una mejor gestión de los departamentos que componen una empresa y los procesos que se desprenden de ellos (Khanbabaei, Mahmood, & Radfar, 2019). Como consecuencia de esto, se pudo identificar la forma en cómo fluye la información al interior de los procesos.

De acuerdo con lo descrito anteriormente, existe la necesidad de conocer el rendimiento de las técnicas de minería de datos para ser implementadas en soluciones de business intelligence. El mercado de programación posee diversas técnicas, las cuales son utilizadas casi de forma automática en el ciclo de desarrollo de soluciones BI, siendo las más utilizadas Árbol de decisión, regresión logística, Naive Bayes y bosques aleatorios (Pérez-Gutiérrez, 2020). No obstante, los estudios muestran aplicaciones a casos específicos para la gestión de volúmenes de datos, obteniendo resultados satisfactorios, pero con limitada documentación en relación a la comparación de su rendimiento.

En síntesis, se observa un crecimiento sostenido de las organizaciones desde hace varios años, con una necesidad de proporcionar información para

optimizar la toma de decisiones. En este sentido, el análisis y procesamiento de los datos es determinante, incrementando la aplicación de técnicas de minería de datos en la construcción de soluciones business intelligence. Su correcta aplicación permite obtener métricas e indicadores con base a predicciones (Pérez-Gutiérrez, 2020). Por este motivo, se requiere profundizar en el análisis y comparación de los principales indicadores de rendimiento de las técnicas de minería de datos más importantes.

1.2. Trabajos previos.

Los investigadores Schuh, Prote, & Hünnekes, (2020) realizaron la investigación titulada “Data mining methods for macro level process planning”, en el Laboratorio de Máquinas y Herramientas de Ingeniería de la Universidad de Aschen en Alemania. Identificaron la necesidad de contar con información de los talleres por medio de un método que permita realizar procesos de planificación a nivel macro, lo cual generaba dificultades para la gestión de sus datos. Por este motivo, propusieron un modelo basado en minería de datos usando bosques aleatorios, el cual sería utilizado para extraer la información relevante que sería adaptable a los procesos de planificación ya definidos; las técnicas trabajaron sobre una base de datos que había acumulado información durante 18 meses y 103,001 operaciones obtenidas de 186 fuentes distintas. Los resultados de su estudio evidenciaron que el modelo de Bosque Aleatorio o Random Forest (RF), por medio del método de validación cruzada, obtuvo una precisión de clasificación del 90.6% para 66 fuentes diferentes y 73.9% de precisión de clasificación para el total; a partir del conjunto de datos, el modelo de predicción corresponde en gran medida con la asignación real. Los investigadores concluyeron que el método Random Forest (RF) puede ser utilizado con éxito para resolver el problema de asignación de recursos.

Los investigadores Adnan, Saqib, Alyas y Rehman, (2020) realizaron la investigación titulada “Effective demand forecasting model using business intelligence empowered with machine learning”, en la Universidad Garrison en Pakistán. Identificaron la importancia del business intelligence en la toma de

decisiones empresariales y que se requiere el análisis de los datos durante todo el proceso del negocio con el objetivo de conocer y predecir demandas futuras basadas en la recopilación de datos de diversas fuentes. Por este motivo, propusieron una solución business intelligence basada en algoritmos de aprendizaje máquina para la obtención de pronósticos de demanda, los cuales serían comparados con datos reales para determinar el porcentaje de error; en este modelo se utilizaron los algoritmos de redes neuronales recurrentes (NNR) y el algoritmo Deep AR. Los resultados obtenidos mostraron que después de aplicar la solución propuesta en los datos de la organización en tiempo real, se obtuvo una precisión de hasta el 92.38% para la tienda en términos de pronóstico de demanda inteligente. Los investigadores concluyeron que el uso del modelo aplicado a business intelligence incrementa la productividad operativa de la organización y reduce significativamente las pérdidas.

Los investigadores Anwar & Addin, (2019) realizaron la investigación titulada "Using data mining techniques to guide academic programs design and assessment", en la Facultad de Ciencias de la Computación de la Universidad de Najran en Arabia Saudita. Identificaron la necesidad de contar con una aplicación para el análisis de los datos y guiar el diseño y evaluación de programas académicos de forma más eficiente. Por este motivo, propusieron la aplicación de técnicas de minería de datos de reglas de asociación para descubrir el conjunto de reglas que dirigen la relación entre los componentes centrales de un programa académico (objetivos y resultados). El conjunto de datos fue procesado previamente y transformados en una representación adecuada para aplicar las técnicas por medio de un sistema de etiquetas. Los resultados de su investigación mostraron un promedio de confianza de 62% en el décimo programa y 82% en el quinto programa, siendo los valores extremos. Los investigadores concluyeron que las reglas descubiertas son de relevante importancia para guiar el diseño y evaluación de los programas académicos de ingeniería; asimismo, las reglas descubiertas revelan una serie de correlaciones.

Los investigadores Ghazzawi & Alharbi, (2019) realizaron la investigación titulada “Analysis of customer complaints data using data mining techniques”, en la universidad de Taif en Arabia Saudita. Identificaron la necesidad de contar con un sistema para procesar los datos de la red de transporte público para la región de Nueva York, la cual contaba con más de 15.3 millones de personas. Por lo anterior, propusieron una aplicación basada en técnicas de minería de datos utilizando el conjunto de datos de los comentarios de los clientes; se probaron los modelos de clasificación Naive Bayes (NB), K Nearest Neighbor (KNN) y Random Trees (RT). Los resultados de su estudio evidenciaron que el método basado en NB obtuvo un promedio de precisión de 86.5%; el modelo K-NN, un nivel de precisión de 81.5% y RT un nivel promedio de precisión de 85.6%. Los investigadores concluyeron que los modelos propuestos tienen una alta tasa de precisión y pueden ser utilizados en el análisis las causas de insatisfacción de los clientes para optimizar el servicio público.

Los investigadores Siyuan, Xingsen, Renhu, & Shouzhen, (2019) realizaron la investigación “Extension data mining method for improving product manufacturing quality” en la Universidad Tecnológica Guangzhoy en China. Identificaron la necesidad de proveer información a una empresa productiva para la elaboración de productos de alta calidad, reducir la tasa de productos no conformes y por ende incrementar la satisfacción de los clientes. Por esta razón, propusieron un método basado en minería de datos mediante Árboles de Decisión (DT) para identificar los factores que impactan en la tasa de productos calificados, utilizando los datos recopilados en cada proceso de la línea de producción; de esta forma se obtienen reglas de calidad las cuales se incorporan posteriormente al proceso. Los resultados evidenciaron que el método basado en DT obtuvo una tasa calificada del producto de 95.7%. Los investigadores concluyeron que el método de minería de datos proporciona una nueva idea para que las empresas mejoren la calidad de fabricación del producto y la tasa de calificación del producto, lo cual es útil para que las empresas logren un sistema de gestión de calidad más eficiente.

Los investigadores Vilorio, Rodríguez, Payares y Vargas, (2019) realizaron la investigación “Determinating student interactions in a virtual learning environment using data mining”, en la Universidad de la Costa en Colombia. Identificaron la necesidad de determinar la interacción de los alumnos matriculados en la modalidad virtual de la Universidad de Mumbai en la India. Por este motivo, propusieron un modelo basado en métodos de minería de datos, el cual analizaría los datos de los estudiantes desde el año 2015 a 2018. Los resultados obtenidos mostraron que los algoritmos con mejor rendimiento respecto a la precisión en clasificación fueron JRip 94.41%, KNN 98.7% y Árbol de Decisión (DT) 92.90%. Los investigadores concluyeron que el modelo de minería de datos permitió determinar que las interacciones de los alumnos en el curso virtual de inglés se encuentran en el nivel promedio con un porcentaje de 69%, y los factores que más influyeron en el modelo fueron las interacciones en los exámenes, tareas, recursos del estudiante, estado civil y situación laboral; la información obtenida es de utilidad para tomar decisiones en la institución.

El investigador Parama, (2018) realizó la investigación “Business intelligence model to analyze social media information”, en el Departamento de Ciencias de la Computación de la Universidad de Bina Nusantara en Indonesia. Identificaron la necesidad de contar con una solución business intelligence para analizar el contenido de las redes sociales y ayudar en la toma de decisión de la organización. Por este motivo propusieron analizar las técnicas de clasificación de texto usando los algoritmos Naive Bayes (NB), Support Vector Machine (SVM) y Árbol de Decisión (DT) con el objetivo de determinar cuál genera mayor precisión para ser implementada en el desarrollo de una solución business intelligence. Los resultados que obtuvieron evidenciaron que el algoritmo de mejor rendimiento fue SVM con una precisión de 78.9%; el algoritmo Naive Bayes obtuvo una precisión de 74.6% y Árbol de Decisión alcanzó una precisión de 57.6%. Los investigadores concluyeron que la mejor técnica para ser implementada en una solución de business intelligence es la que utiliza el algoritmo SVM, la cual permite automatizar el análisis de datos con alta precisión y contribuir al proceso de toma de decisión.

Los investigadores Harley & Liu, (2017) realizaron la investigación titulada “Towards industry 4.0 utilizing data-mining techniques: a case study on quality improvement”, en el Departamento de Ingeniería de la Universidad de Cardiff en Reino Unido. Identificaron la necesidad de una empresa de manufactura de contar con un método de análisis de datos de sus procesos de producción para optimizar la calidad de sus productos durante la fabricación. Por este motivo, diseñaron un método aplicando técnicas de minería de datos y utilizando software de código abierto, el cual fue ejecutado sobre un conjunto de datos de 5000 instancias. Los resultados mostraron que el algoritmo JRip tuvo una precisión de 95.4%, mientras que el algoritmo PART tuvo una precisión de 96.9%. Los investigadores concluyeron que los principios de la minería de datos pueden ser utilizados en los procesos de fabricación y toma de decisión con el objetivo de realizar una gestión más eficiente que ayude a una mejor calidad de los productos.

Los investigadores Reuter, Brambring, Weirich, & Kleines, (2016) realizaron la investigación titulada “Improving data consistency in production control by adaptation of data mining algorithms”, en la Universidad de Aachen en Alemania. Identificaron la necesidad en algunas empresas de Alemania de mejorar la calidad de análisis de los volúmenes de datos de los procesos de producción para obtener conocimiento útil que permita implementar mejoras en el proceso de producción. Por este motivo, propusieron un modelo para incrementar la calidad de los datos relevantes para los procesos de producción mediante la adaptación de algoritmos de minería de datos Decision Tree (DT), y K Nearest Neighbor (KNN) y de esta forma estimar los valores probables para la inconsistencia de datos transaccionales. Los algoritmos fueron evaluados en casos reales de empresas alemanas con el objetivo de determinar su eficiencia. Los resultados mostraron que el método utilizando DT obtuvo un rendimiento de 66.5% con una eficiencia del proceso de 578 segundos; por otro lado, método utilizando la técnica KNN obtuvo un rendimiento de 67.9% con una eficiencia del tiempo de 329 segundos. Los investigadores concluyeron que los algoritmos de DM evidenciaron ventajas frente a los enfoques existentes y que los métodos basados en minería de datos permiten aumentar la coherencia de los datos en

el control de la producción, siendo un factor importante para la toma de decisiones.

Los investigadores Ozyirmidokuz, Kumru, & Mustafa, (2015) realizaron la investigación denominada “A data mining based approach to a firm’s marketing channel”, en el Departamento de Tecnologías Informáticas de la Universidad de Erciyes en Turquía. Identificaron la necesidad de las empresas para analizar los grandes volúmenes de datos de marketing y conseguir información sobre la cual soportar la toma de decisión. Por esta razón, propusieron un modelo para la obtención de conocimiento de los datos ingresados al canal de marketing y así mejorar la eficiencia del sistema. Utilizaron un Cross Industry Standard Process for Data Mining (CRISP-DM) para el análisis de datos, los cuales se agruparon aplicando el mapa de auto organización de Kohonen (SOM) para la reducción de atributos. Asimismo, se aplicó el análisis de detección de anomalías por medio de Árbol de Decisiones (DT). Los resultados evidenciaron una tasa de precisión del modelo de 92.67%, lo cuales corresponden a un total de 278 registros procesados correctamente y 7.33% de registros incorrectos de un total de 300. Los investigadores concluyeron que el modelo DT aplicado puede ser utilizado para la predicción de comportamientos futuros de las empresas de marketing y de ayuda en el proceso de planificación.

El investigador Suharjito, (2015), realizó la investigación titulada “Data mining of automatically promotion tweet for products and services using Naïve Bayes algorithm to increase twitter engagement followers atPT. Bobobobo”, en la Universidad Bina Nusantara de Indonesia. Identificó la necesidad de maximizar la promoción de los productos y servicios para obtener más seguidores; para ello, se requería la extracción de datos para descubrir la información de tendencias de los seguidores y así generar tweets de forma automática. Por este motivo, propuso un modelo aplicando minería de datos utilizando el algoritmo Naive Bayes (NB) para clasificar a los seguidores en razón de los productos y servicios y luego analizar las palabras de tendencias e incluirlas en los tweets de promoción. Los resultados mostraron una precisión de clasificación del algoritmo NB de 90.31% de los datos de prueba de tweets de

productos usados y 80.91% de precisión en los datos de prueba de tweets de servicios utilizados; realizando una combinación de los datos de prueba de productos y servicios se obtuvo una precisión de 83.5%. El investigador concluyó que el modelo utilizado tiene una alta precisión para determinar las tendencias dentro de las publicaciones de tweets a nivel de productos y servicios, lo cual permite incrementar el número de seguidores.

1.3. Teorías relacionadas al tema.

1.3.1. Business intelligence (BI)

La inteligencia de negocios, también conocida como business intelligence (BI) se basa en un proceso de análisis de información estructurada de una organización, la cual es denominada data warehouse. De esta información, se pueden descubrir tendencias y extraer conclusiones para tomar decisiones (Vercellis, 2009). La finalidad de business intelligence es contribuir de forma constante a las empresas para que mantengan su nivel de competencia, facilitando la información requerida en el tiempo preciso.

Por otro lado, la gestión de los datos es determinante para que las empresas tomen decisiones acertadas y mantengan su posición competitiva en el mercado. En ese sentido, business intelligence cumple un rol determinante en la supervivencia de las organizaciones, en las cuales la información crece de forma permanente, Asimismo, es necesario comprender que la toma de decisiones puede ser un factor crítico en determinados momentos; de ello, dependerá que la organización sostenga su competitividad al largo plazo basada en el análisis y uso eficiente de sus datos (Phanikanth & Sudarsan, 2017).

Respecto a su arquitectura, una solución business intelligence integra un conjunto de tecnologías que ayudan a recopilar, almacenar, consultar, y analizar grandes volúmenes de datos y proporcionan acceso a los datos necesarios que requieren las empresas para la toma de decisiones mediante la obtención de análisis e informes (Denis-Cătălin , Ioana-Gilia, & Miruna, 2019). En la

actualidad, la inteligencia de negocios se encuentra ampliamente vinculada a las técnicas de minería de datos.

1.3.1.1. Componente de business intelligence

La inteligencia de negocios tiene los siguientes componentes: En primer lugar, se encuentran las fuentes de información, con las cuales se inicia el proceso para alimentar la información de la data warehouse. En segundo lugar, se encuentra el proceso ETL, el cual engloba los procesos de extracción, transformación y carga de datos en la data warehouse. Los datos obtenidos deben ser transformados. Usualmente, la información obtenida directamente de los sistemas transaccionales no se encuentra preparada para tomar decisiones. En tercer lugar, se encuentra la data warehouse; con la metadata se busca almacenar los datos de una forma que incremente su facilidad de acceso. En cuarto lugar, se ubica el motor OLAP, el cual debe promover capacidad de cálculo y consultas para el análisis de escenarios de volúmenes de datos.

Gracias al avance de los sistemas de gestión de base de datos, existen otras alternativas tecnológicas al OLAP, que se encuentran integradas a los sistemas de gestión de bases de datos (SGBD); Por último, se encuentra la herramienta de visualización que permita el análisis y navegación de la información por medio de una aplicación web o de escritorio (Vercellis, 2009).

En la siguiente imagen, se muestran los componentes de la solución business intelligence, en las cuales se aprecia que las fuentes de información pueden ser variadas, por ejemplo, sistemas de transacciones, fuentes de información externas y sistemas que se encuentran al interior de los departamentos o áreas de la empresa.

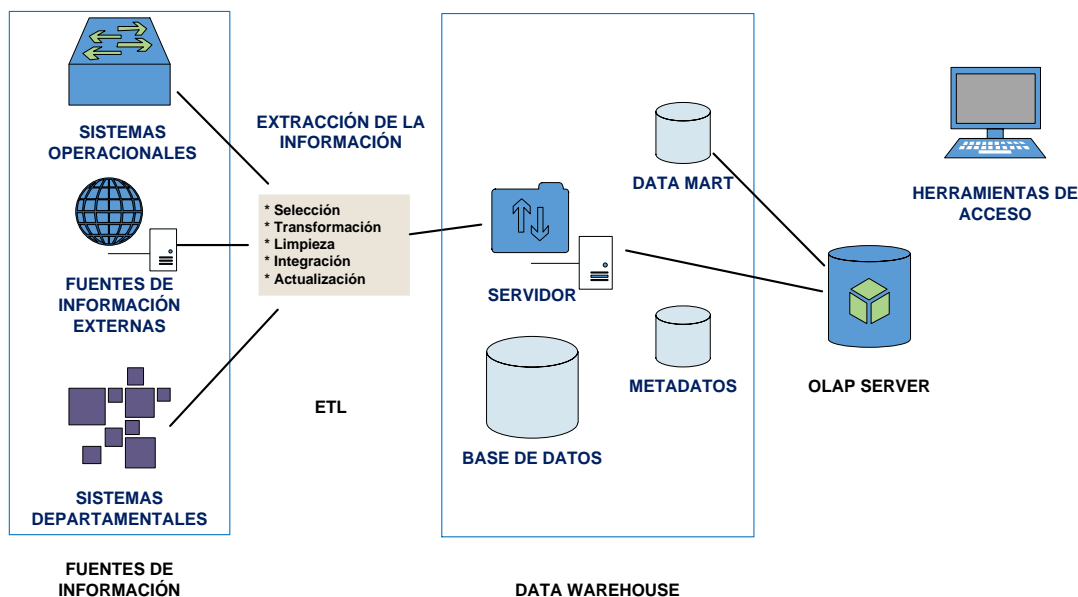


Figura 1. Componentes business intelligence. Fuente: (Cano, 2007)

1.3.1.2. Fuentes de información

Una fuente de información es aquella que provee los datos. Las fuentes a las que se puede acceder son principalmente la de sistemas transaccionales u operacionales, las cuales incluyen softwares creados a medida como ERP¹, CRM² o SCM³. Asimismo, se puede contar con sistemas de áreas específicas como, ventas, presupuestos u hojas de cálculo utilizadas para registrar información.

Existen factores que contribuyen durante el proceso complejo de carga de datos. Una de ellas es la cantidad de fuentes de información diferentes de las que se obtienen los datos. Este número es variable para cada empresa; existen

¹ ERP hace referencia a las iniciales de Enterprise Resource Planning, utilizado para planificación de recursos empresariales.

² CRM son las iniciales de Customer Relationship Management, sistema empleado para relaciones con los clientes.

³ CSM refiere a Supply Chain Management, el cual es un sistema especializado para la gestión de la cadena de suministro.

empresas que cuentan con una fuente de datos, generalmente empresas pequeñas, y empresas que obtienen información de docenas de fuentes, ya que la magnitud de sus operaciones le exige tener procesos descentralizados. En grandes corporaciones, el promedio es alrededor de 8 bases de datos y pueden llegar a 50 (Cano, 2007). En la siguiente imagen, se observa el flujo de la información desde las Fuentes de información hasta el usuario final de business intelligence.

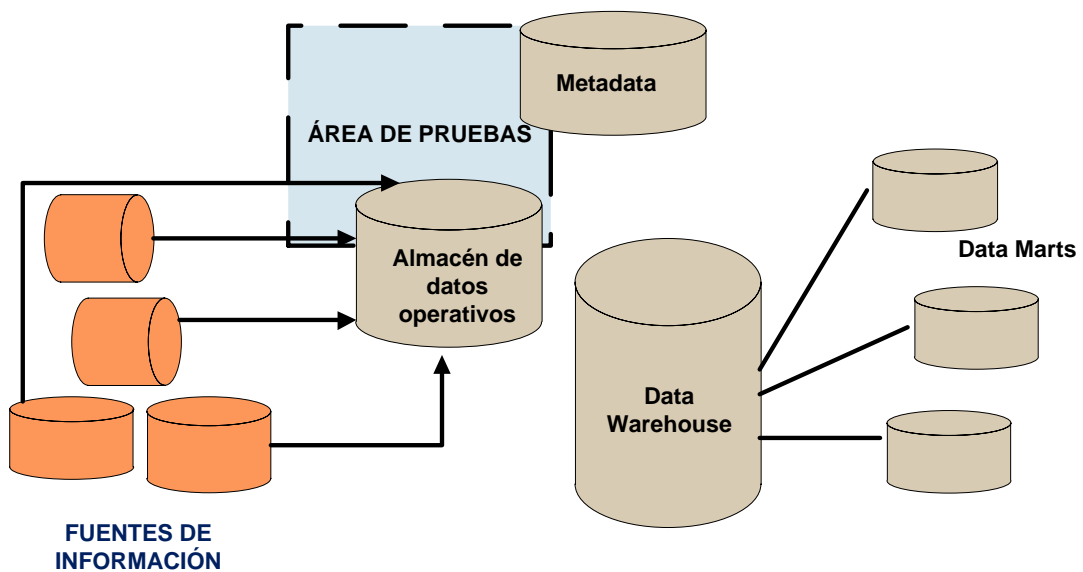


Figura 2. Proceso de integración de datos. Fuente: (Cano, 2007)

La responsabilidad respecto a la calidad de los datos no es una tarea exclusiva del área de tecnologías de información. Es un trabajo que debe realizarse en conjunto con los propietarios de los procesos y sistemas de información que le brindan soporte. Desde la planificación del proyecto se debe poner especial interés en la calidad de los datos, ya que, si no es la óptima, no se podrán conseguir los resultados esperados del proyecto. Asimismo, se debe comprender que el problema de la calidad de datos es un asunto estratégico al que se debe asignar recursos de forma prioritaria. (Cano, 2007). Por lo anterior, se debe entender que una tarea fundamental del proceso de integración de datos es velar por su calidad.

1.3.1.3. Data warehouse (DW)

La definición de data warehouse, es la solución a la necesidad de los usuarios que requieren información histórica con una alta consistencia para ser analizada de manera previa a la toma de decisiones (Cano, 2007). Se puede entender la data warehouse como un conjunto de información implementada como soporte a los sistemas de toma de decisiones (Kimball & Ross, 2013). A continuación, se muestra una figura en la que se aprecia la ubicación de la data warehouse dentro del esquema de inteligencia de negocios.

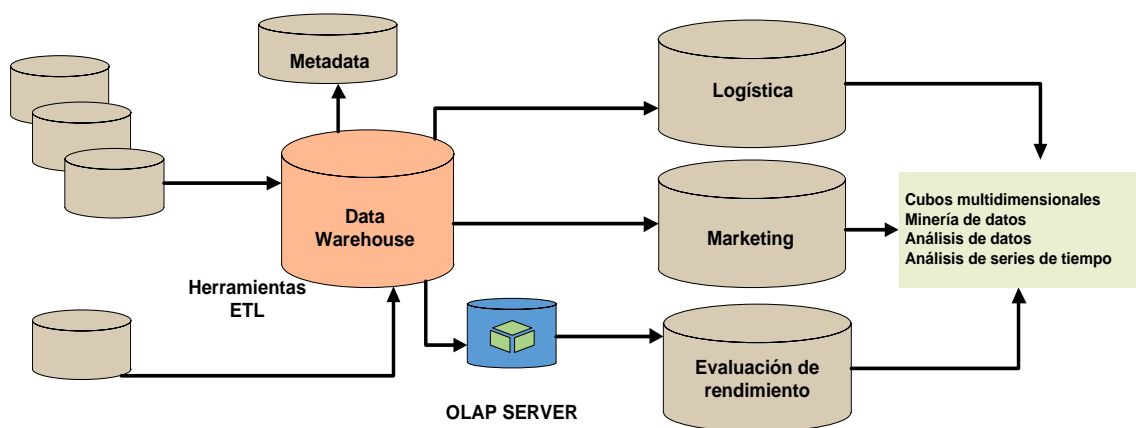


Figura 3. Arquitectura de una data warehouse. Fuente: (Kimball & Ross, 2013)

El entorno de una data warehouse permite el acceso a información organizada para efectuar consultas, las cuales permiten a los usuarios obtener información significativa. En algunas oportunidades, el valor de la información se configura en el proceso continuo de consultas y análisis de los datos. Normalmente, las consultas iniciales entregan bloques de información que posteriormente es reprocesada mediante nuevas consultas. No todas las técnicas de análisis elegidos son adecuadas, modificando nuevamente las consultas en función de las necesidades (Cano, 2007).

1.3.1.4. Motor OLAP

En la actualidad, el mercado cuenta con herramientas tecnológicas para analizar la información almacenada en una data warehouse, pero la más utilizada entre

los desarrolladores de soluciones business intelligence es OLAP, el cual es el acrónimo de On-Line Analytical Processing. Los usuarios requieren analizar información en los diferentes niveles y deben trabajar sobre varias dimensiones. Por ejemplo, ventas de productos que deben ser clasificados por zona, cliente, proveedor, fecha, entre otros datos que sean de utilidad para la empresa y tomar decisiones respecto a los productos. Los usuarios realizan este análisis en los niveles más altos de agregación o detalle. OLAP otorga dichas funciones con la flexibilidad requerida para determinar las relaciones y tendencias que no ofrecen otras herramientas con flexibilidad limitada (Cano, 2007). A este tipo de análisis se les denomina “análisis multidimensional”, ya que facilitan el análisis de un hecho desde diversas perspectivas.

En la siguiente imagen, se muestra la forma natural aplicada para analizar la información por parte de los usuarios responsables de la toma de decisión, ya que los modelos de negocio en su mayoría son de tipo multidimensional. La representación generalizada del OLAP es un cubo.

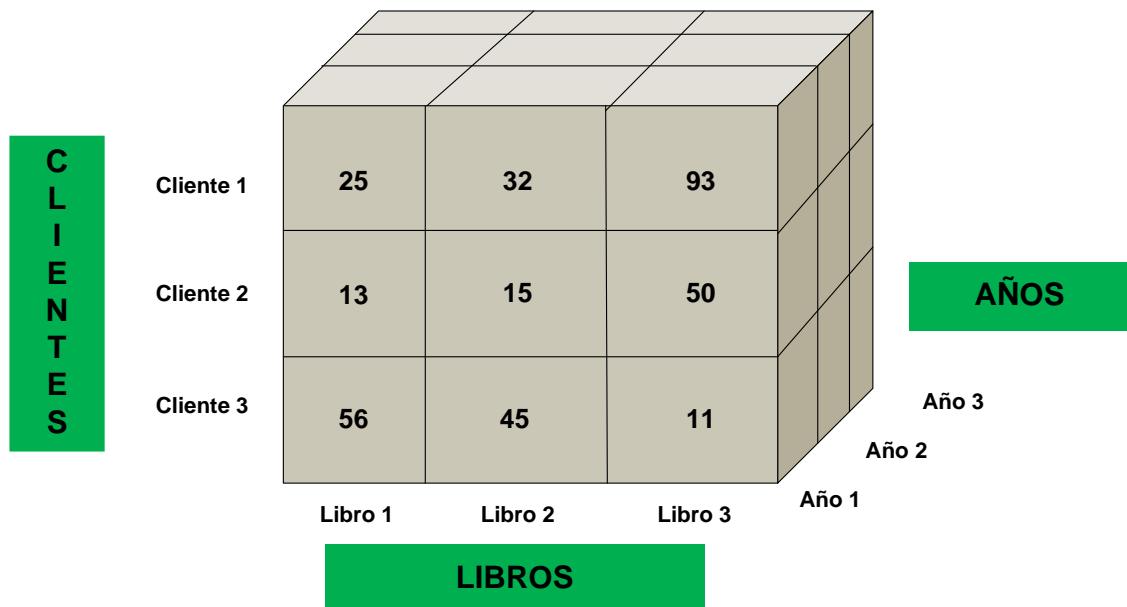


Figura 4. Representación gráfica del OLAP. Fuente: (Kimball & Ross, 2013)

Las herramientas OLAP permiten la rotación del cubo; en otras palabras, permite cambiar el orden de las dimensiones. Esto se realiza con el objetivo de

realizar un mejor análisis de los datos desde diferentes perspectivas de acuerdo a las necesidades de la empresa o de áreas específicas.

1.3.2. Base de datos (BD)

Una base de datos (BD) es definida como un conjunto de datos interrelacionados que poseen significado (Silberschatz, Korth, & Sudarshan, 2002). Asimismo, es un almacenamiento de datos estructurado y controlado de forma central para proveer de información a múltiples sistemas.

En la actualidad, las bases de datos son utilizadas por las organizaciones. Algunas de las áreas de aplicación son las siguientes: empresas de rubro financiero, instituciones educativas, hospitales, comercio, procesos de producción, gestión de recursos humanos, entre otras (Silberschatz, Korth, & Sudarshan, 2002). La gestión de los datos involucra también el proceso de análisis mediante diferentes métodos.

1.3.2.1. Abstracción de Datos

Todo sistema de información de calidad se caracteriza por la recuperación de los datos de forma eficiente. Lo anterior es una preocupación constante de los desarrollados, lo cual conllevó al diseño y construcción de estructuras que ayuden a representar los datos. Las personas que usan los sistemas de bases de datos no conocen ampliamente su funcionalidad. Por ello, los programadores ocultan la capa de complejidad bajo niveles que simplifiquen su comprensión, a los cuales se les conoce como niveles de abstracción (Silberschatz, Korth, & Sudarshan, 2002). A continuación, se realiza una explicación de cada nivel:

a) Nivel físico

Es el más básico de abstracción, en el cual se describe la forma de almacenamiento de los datos. Asimismo, se explican las estructuras de datos complejos de bajo nivel (Silberschatz, Korth, & Sudarshan, 2002).

b) Nivel lógico

En este nivel de abstracción se describen cuáles son los datos que serán almacenados en la base de datos y cuáles son las relaciones existentes. Una base de datos completa se refiere a una estructura de características simples que permite organizar los datos. Si bien las estructuras del nivel lógico pueden incluir parte de complejidad del nivel físico, quienes utilizan la base de datos a este nivel no se deben preocupar por ello (Silberschatz, Korth, & Sudarshan, 2002). El nivel lógico de abstracción es utilizado por los administradores de base de datos, pues ellos deciden la información que se mantendrá.

c) Nivel de vista

Es el máximo nivel de abstracción, en el que se describe parte de la base de datos. A pesar de que previamente hayan utilizado estructuras simples en el nivel anterior, se encuentra complejidad. Esto se debe a la diversidad de información alojada en una base de datos.

Gran parte de los usuarios de los sistemas de base de datos no requieren la totalidad de información. En reemplazo, se requiere acceder únicamente a una sección de la base de datos (Silberschatz, Korth, & Sudarshan, 2002). Para simplificar sus interacciones con los sistemas de información, se utiliza la abstracción del nivel vista. Una misma base de datos puede proporcionar múltiples vistas.

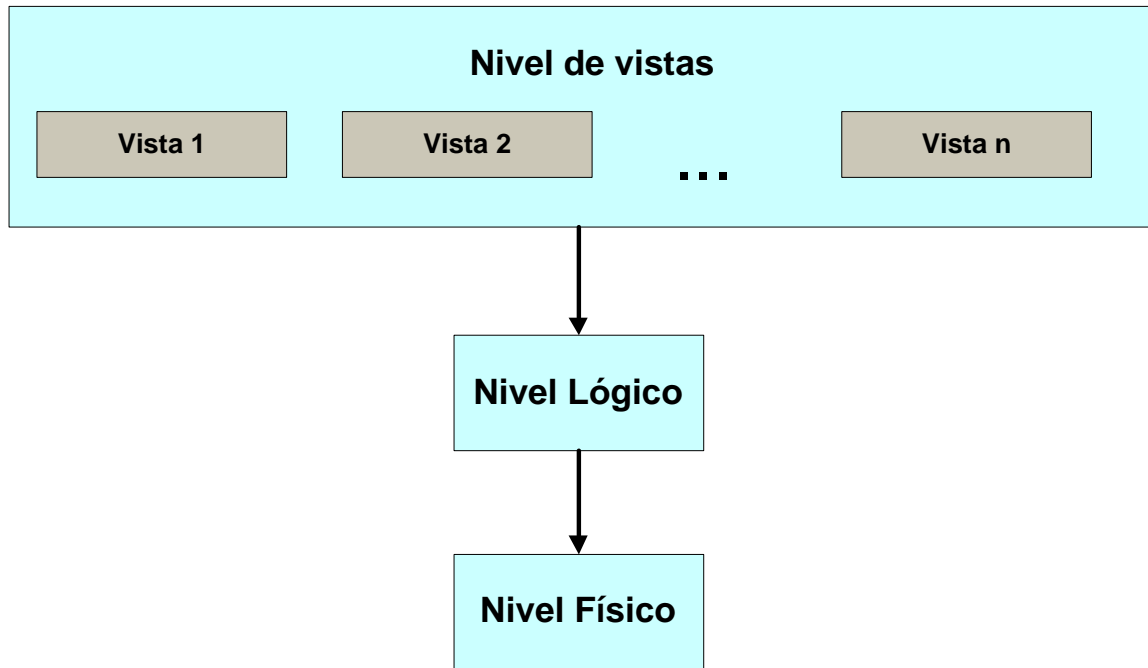


Figura 5. Niveles de abstracción de la base de datos. Fuente: (Silberschatz, Korth, & Sudarshan, 2002)

1.3.2.2. Modelos de bases de datos

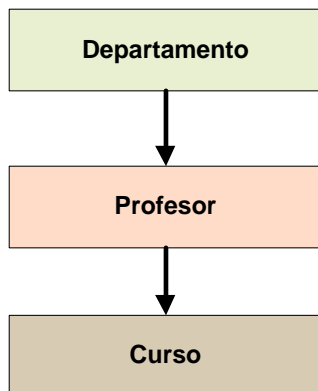
Desde la aparición del concepto de software, las bases de datos tuvieron una evolución significativa en cuanto a su diseño. De acuerdo con ello, los modelos más conocidos son 5: Modelo jerárquico, Modelo en red, Modelo Relacional, Modelo Multidimensional, Modelo de Objetos (Vélez, 2018). A continuación, se realiza una ampliación de cada modelo.

a) Modelo de base de datos Jerárquico

La estructura jerárquica fue empleada en las primeras bases de datos. Las vinculaciones entre sus datos configuran una estructura de tipo árbol. En la actualidad, las bases de datos jerárquicas más empleadas por las empresas son: IBM, IMS 4 y Windows de Microsoft.

⁴ IMS son las iniciales de Information Management System en cual es un gestor de base de datos transaccionales de alta capacidad.

ESTRUCTURA LÓGICA



EJEMPLO DE BASE DE DATOS

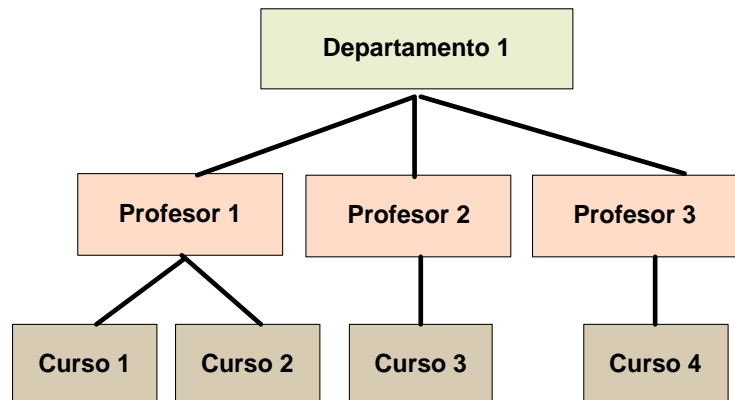


Figura 6. Modelo de base de datos jerárquico. Fuente: (Vélez, 2018)

b) Modelo de base de datos en Red

La estructura en red posee una complejidad en sus relaciones superior a la estructura jerárquica. Este tipo de modelo, permite establecer relaciones entre sus registros con otros que se puedan relacionar mediante diferentes rutas (Vélez, 2018).

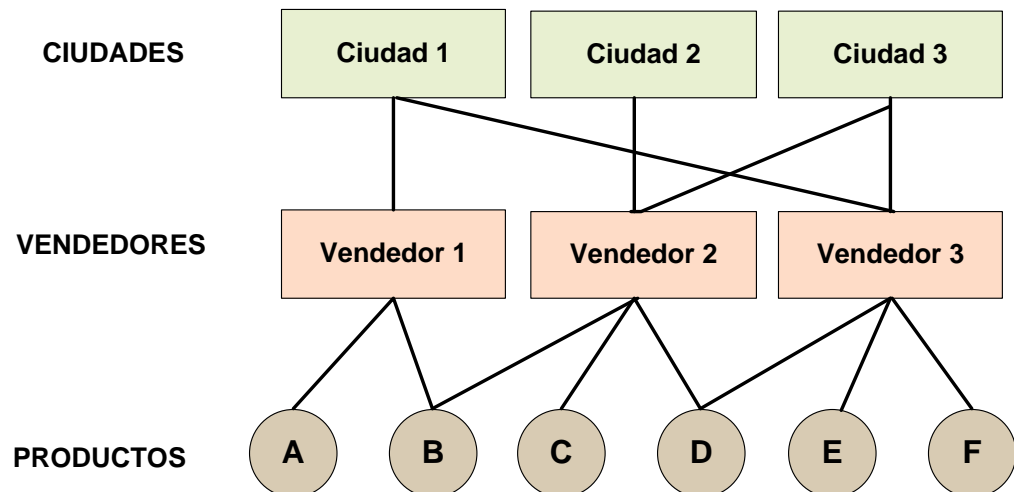


Figura 7. Modelo de base de datos en red. Fuente: adaptado de (Vélez, 2018)

c) Modelo de base de datos con estructura relacional

Este tipo de BD es la más empleada en la actualidad. Esta estructura realiza un almacenamiento de los datos en filas y columnas, las cuales contienen los atributos. Las tablas serán relacionadas mediante claves genéricas (Vélez, 2018).

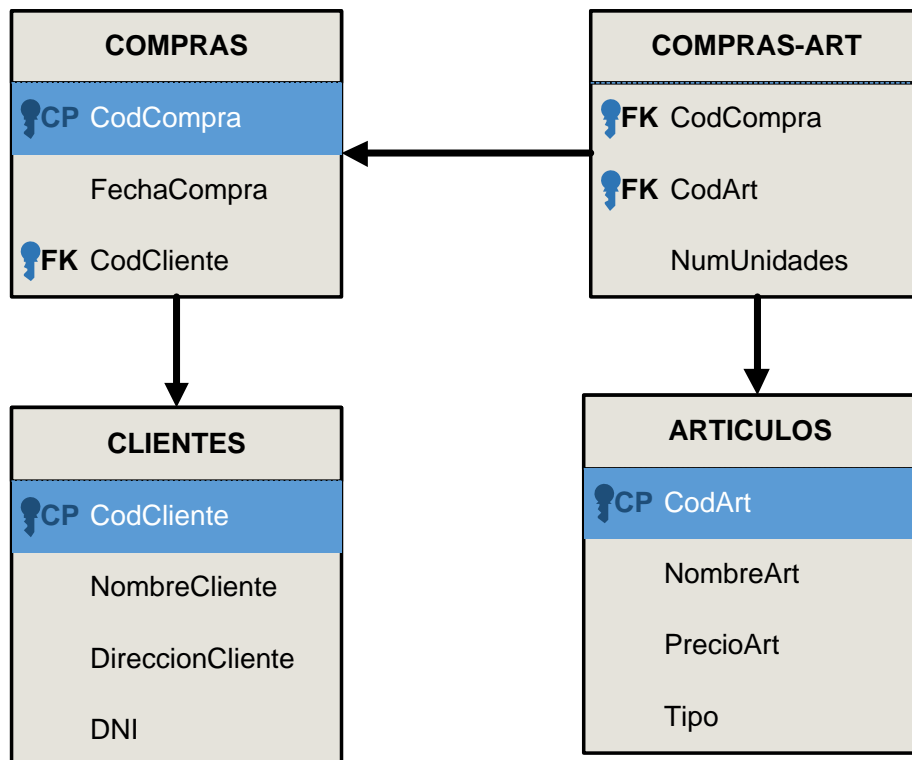


Figura 8. Modelo de base de datos relacional. Fuente: (Vélez, 2018)

d) Modelo de base de datos con estructura multidimensional

La estructura multidimensional presenta similitudes con el modelo relacional. Sin embargo, en lugar de contar con dimensiones (filas y columnas), tiene un número mayor de dimensiones: "n" dimensiones. Su estructura permite la representación gráfica de un cubo (Vélez, 2018).

	Abril	Mayo	Junio
Producto 1	212	534	254
Producto 2	21	46	33
Producto 3	310	321	200
Producto 4	120	234	131
Producto 5	43	78	55
Producto 6	12	32	21
	Chile	Brasil	Perú

Figura 9. Modelo de base de datos multidimensional. Fuente: (Vélez, 2018)

e) Modelo de base de datos con estructura orientada a objetos

Este tipo de estructura fue elaborada bajo el modelo de programación orientada a objetos (POO) como JAVA y C++. De esta manera, soporta datos del tipo imagen, audio y texto de forma natural. La estructura fue ampliamente difundida en los sistemas web que gestionan contenidos multimedia (Vélez, 2018).

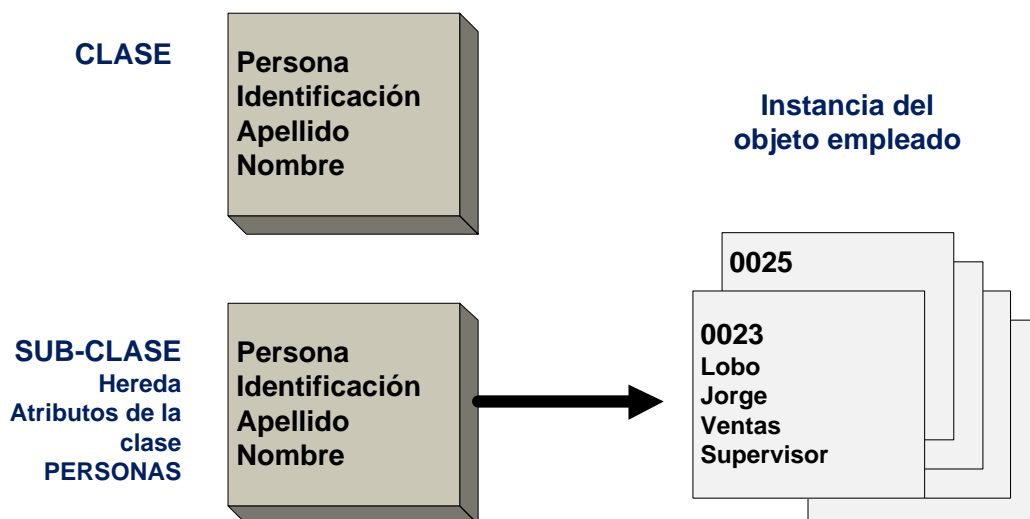


Figura 10. Modelo de base de datos orientada a objetos. Fuente: (Vélez, 2018)

1.3.2.3. Sistemas de gestión de base de datos (SGBD)

Los SGBD se definen como el conjunto de programas que facilitan el acceso a un conjunto de datos, los cuales se encuentran relacionados. La información a la que se accede es de gran importancia para la empresa (Silberschatz, Korth, & Sudarshan, 2002). El objetivo fundamental de un SGBD es brindar un método de almacenamiento y recuperación de datos de una base de datos de forma eficiente.

Los sistemas de bases de datos fueron construidos para gestionar enormes volúmenes de datos. Los cuales son obtenidos de diversas fuentes. Para poder gestionar dichos datos, se requiere determinar previamente estructuras para el alojamiento de la información. Además, es de vital importancia que los SGBD suministren información confiable, aunque se presenten problemas en los sistemas de información (Silberschatz, Korth, & Sudarshan, 2002).

Si los datos serán compartidos con un grupo de usuarios, el sistema debe ser capaz de evitar resultados inconsistentes. Para las empresas, la información es realmente importante para la toma de decisiones. Por este motivo, los ingenieros de informática y sistemas orientaron gran parte de sus esfuerzos al desarrollo de técnicas para el análisis de datos (Silberschatz, Korth, & Sudarshan, 2002).

1.3.3. Metodologías de desarrollo business intelligence (BI)

En la actualidad, existen diversas metodologías de desarrollo de soluciones business intelligence (BI), siendo las más reconocidas la metodología de Ralph Kimball y Bill Inmon (Shaker & Abdeltawab, 2011). A continuación, se realiza una breve explicación de cada una.

1.3.3.1. Metodología Inmon

Bill Inmon observó que los datos operativos suelen estar orientados a las aplicaciones y, en consecuencia, no están integrados y que existe la necesidad que los datos de almacén de datos estén integrados. El entorno operacional de la construcción de la data warehouse es compatible con el ciclo de vida tradicional de un sistema informático. No obstante, el desarrollo de la Data Warehouse trabaja bajo un ciclo de vida diferente llamado CLDS.

El CLDS es casi exactamente el reverso al SDLC, System Development Life Cycle, ya que inicia con datos. Una vez que los datos están a la mano, se integran y luego se prueban para ver que sesgo hay en los datos (Inmon, 2002). A continuación, se aprecia una figura de la concepción de desarrollo:

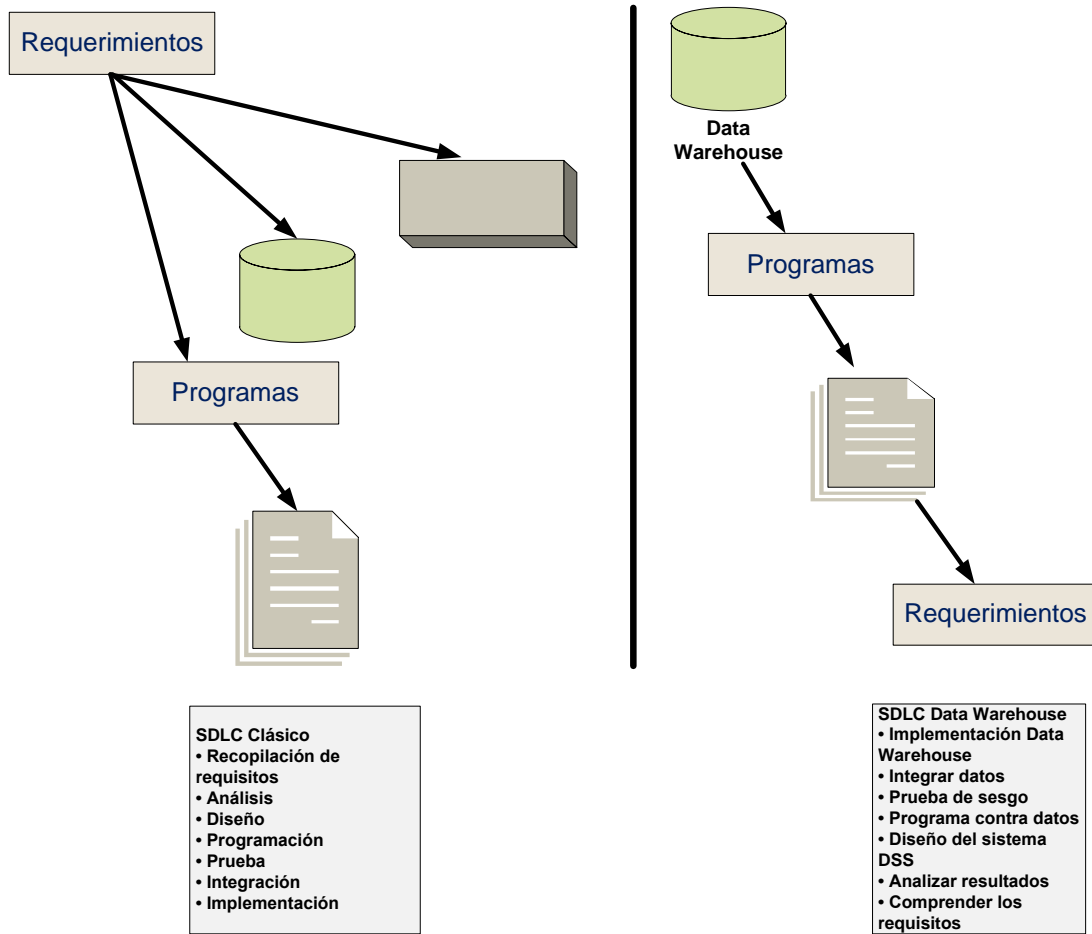


Figura 11. Ciclo de vida inverso del desarrollo del Data Warehouse. Fuente: (Inmon, 2002)

Bill Inmon es valorado por muchos expertos como el pionero de la data warehouse. Inmon manifiesta que una DW contar con las siguientes características: a) Debe estar dirigida a un área específica de la empresa. b) Debe ser capaz de integrar diversas fuentes de información. c) Debe ser variable en el tiempo, ya que los datos son variables a lo largo del tiempo en las organizaciones d) No debe ser volátil, es decir, los datos no deben ser eliminados (Gutiérrez, 2012). La propuesta metodológica de Inmon es interactiva y sigue una estructura contraria al modelo clásico, tal como se mencionó. De esta forma, la metodología principalmente consiste en lo siguiente:

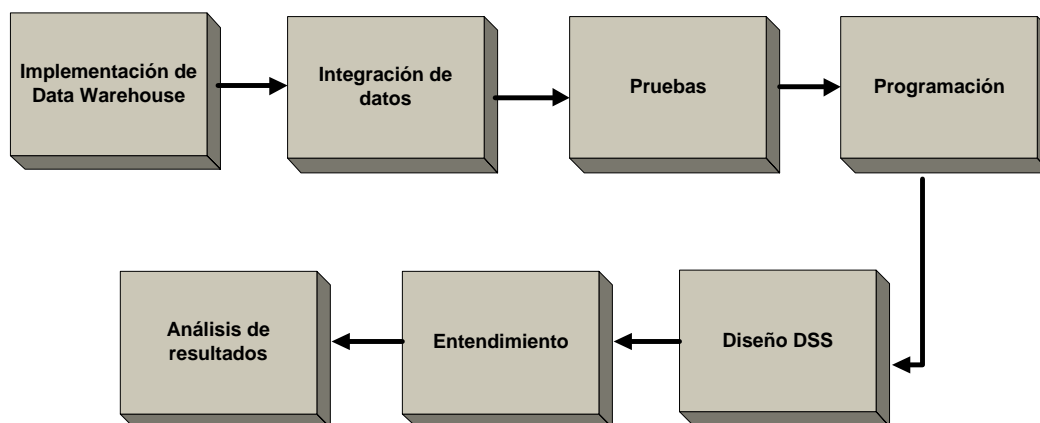


Figura 12. Metodología Inmon. Fuente: (Gutiérrez, 2012)

Esta metodología establece que el desarrollo de un proyecto DW exige una alta inversión de tiempo, pues durante el proceso se involucran las necesidades generales de la organización. Estas necesidades y requerimientos de información van evolucionando con el tiempo y la data warehouse debe ser capaz de cubrir las necesidades para un mayor número de usuarios, sin que ello implique una disminución de su rendimiento (Gutiérrez, 2012).

Por lo anterior, al alcanzar el máximo pico de rendimiento se construyen segmentos del DW, los cuales se alimentan de los datos obtenidos mediante OLTP (procesamiento de transacciones en línea) y que permite tener la información alojada de forma que sea distribuida a los distintos departamentos. Con este procedimiento, se conduce a una reducción de la demanda de la data warehouse. En la siguiente figura, se observa dicha estructura:

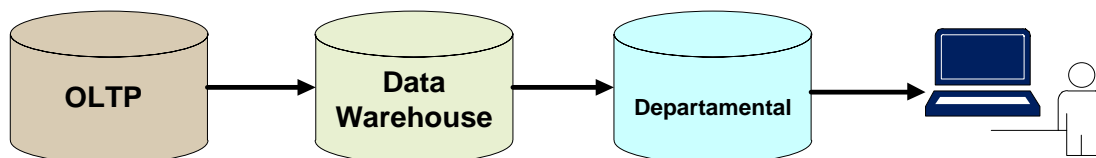


Figura 13. Data warehouse. Fuente: (Gutiérrez, 2012)

1.3.3.1.1. Implementación de data warehouse

Siguiendo la metodología propuesta por Inmon, para implementar el DW se requiere una secuencia de pasos: El primer paso consiste en determinar un estándar y en función de ello se deben identificar las fuentes de datos. El segundo paso consiste en interiorizar el flujo que sigue la información; por este motivo, se debe recurrir a un modelo de proceso. Dicho modelo posee información, la cual es representada, principalmente, en diagramas, por ejemplo: diagramas de flujo de datos y diagramas de transacciones. El último paso consiste en trabajar el modelo de datos, tomando en cuenta el elemento tiempo para que luego se establezcan las relaciones (Gutiérrez, 2012).

1.3.3.2. Metodología Kimball

Kimball desarrolló una metodología cuyo enfoque se basa en el diseño y construcción de la data warehouse. Este enfoque metodológico establece que el objetivo que dirige los proyectos business intelligence es el negocio. Por este motivo, una de las fases iniciales de todo proyecto BI es determinar las necesidades de la organización, ya que de ellas se desprenderán los objetivos y métricas que guiarán el proyecto (Gutiérrez, 2012). En la figura 14, se observan los componentes de la metodología.

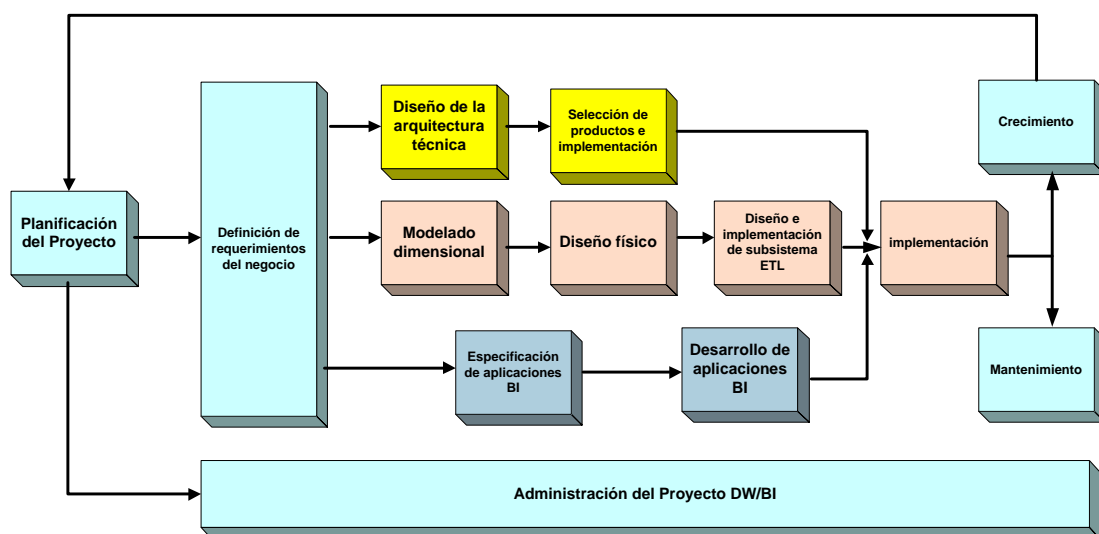


Figura 14. Metodología Kimball. Fuente: (Kimball & Ross, 2013)

La arquitectura de esta metodología está conceptualizada bajo el principio que los sistemas fuente mantienen pocos datos históricos. Un buen almacén de datos puede aliviar los sistemas fuente de gran parte de la responsabilidad de representar el pasado. En muchos de los casos, los sistemas de origen son aplicaciones para fines especiales sin ningún compromiso para compartir datos comunes como producto, cliente, geografía o calendario con otros sistemas operacionales en la organización. Por supuesto, una aplicación cruzada ampliamente utilizada por las organizaciones es el sistema de planificación de recursos empresariales, denominado comúnmente como ERP. La gestión de datos maestros operacionales del sistema podría ayudar a abordar estas deficiencias (Kimball & Ross, 2013). A continuación, se muestra gráficamente la arquitectura de la metodología:

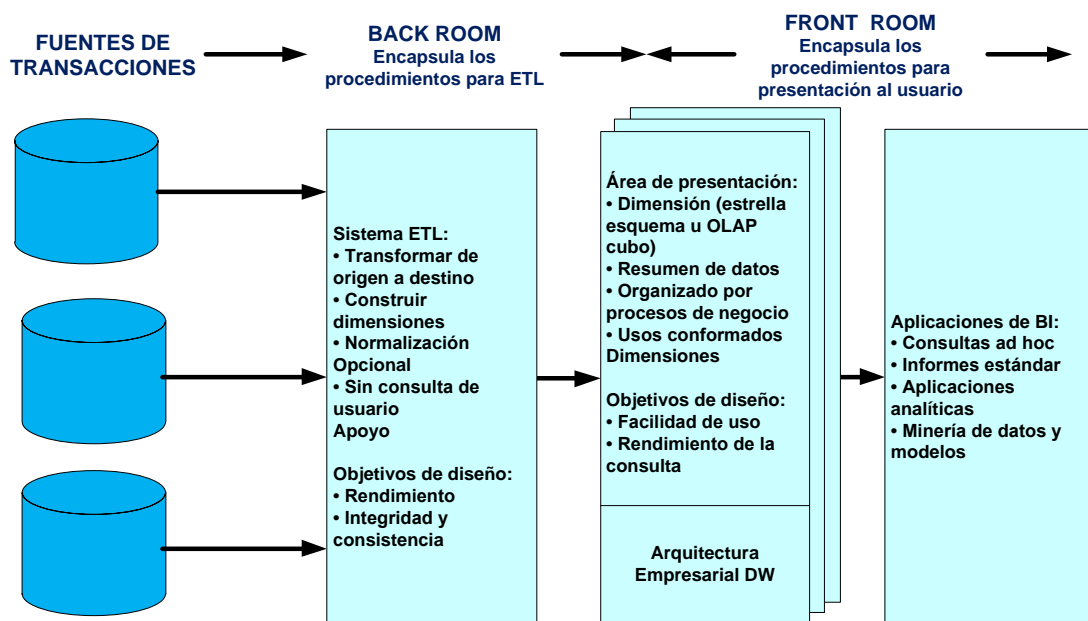


Figura 15. Arquitectura Kimball. Fuente: (Kimball & Ross, 2013)

En las siguientes líneas, se detalla cada uno de los momentos que conforman la propuesta metodológica de Kimball.

1.3.3.2.1. Planeamiento del Proyecto

En todo proyecto de tecnologías de información, el paso inicial es la planificación. Este proceso de planeación posee los siguientes pasos: 1) Preparación, en esta parte, se debe contar con un líder que conozca ampliamente las necesidades del proyecto y los requerimientos de información para la construcción de la data warehouse. 2) Evaluación de factibilidad técnica, esta fase posee una complejidad significativa, ya que se debe determinar si los datos obtenidos de las operaciones son suficientes para los requerimientos del negocio. Para ello, se necesita trabajar con datos confiables a bajo un nivel específico de granularidad. 3) Alcance, en esta parte se determinan los límites que existen en torno al proyecto. Los aspectos antes mencionados son definidos entre los directivos y el equipo de business intelligence. 4) Justificación, consiste en realizar una estimación de los costos e impacto del proyecto (Gutiérrez, 2012).

1.3.3.2.2. Requerimientos del negocio

Para conseguir los requerimientos, se debe proponer un método que establezca la forma en cómo serán obtenidos. Para ello, se cuentan con dos técnicas validadas para el registro de requerimientos: las entrevistas y focus group. Estas técnicas deben realizarse con dos actores del negocio: directivos del negocio o alta dirección y equipo de tecnologías de información. La intención es conseguir información sobre los procesos claves del negocio y relacionar las propuestas con los datos (Gutiérrez, 2012). Para lograr ello, se debe seleccionar a las personas adecuadas para cumplir el rol de entrevistador.

1.3.3.2.3. Diseño de la arquitectura

Consiste en la determinación de las características con las que contará el diseño, el cual debe tomar los elementos necesarios para la construcción de la data warehouse. Los elementos con los que contará el DW serán representados mediante un modelo que cuente con los requerimientos identificados y que

deben ser incluidos de forma inmediata (Gutiérrez, 2012). Conforme a lo determinado en la planificación y diseño de su arquitectura, se debe buscar un producto de software que se ajuste a lo planificado. El proceso de selección inicia con la construcción de una tabla de evaluación con las necesidades y prioridades para una posterior exploración de opciones en el mercado. Luego, se comparan las alternativas que mejor se ajusten al proyecto; se solicitan algunas pruebas bajo prototipos. Finalmente, se selecciona la mejor opción y se realiza la implementación y pruebas (Gutiérrez, 2012).

1.3.3.2.4. Modelado dimensional

En el proceso de modelado dimensional se debe determinar la información cualitativa y cuantitativa que ofrecen una comprensión del negocio, la cual es definida como dimensiones. La obtención de este modelado se realiza de la siguiente forma: Primero, se elabora una lista de las potenciales dimensiones con sus respectivas interacciones. Segundo, se deben identificar claramente los procesos del negocio. Tercero, se realiza una evaluación de la consistencia, de los valores válidos y la disponibilidad de los datos que confirman las dimensiones. Cuarto se construye un esquema dimensional base. Quinto, se realiza la validación del esquema en base a indicadores y concluye con la documentación del modelo (Gutiérrez, 2012).

1.3.3.2.5. ETL

El sistema ETL de la data warehouse para una solución Business Intelligence consiste en un subproceso o área de trabajo específica para el tratamiento de las estructuras de datos. Se ubica entre las fuentes de datos transaccionales y la presentación en BI (Kimball & Ross, 2013). El primer paso consiste en la extracción de los datos del almacén. Esto implica que debe ser capaz de comprender los datos fuentes y copiarlos dentro del flujo ETL para una mejor gestión. Una vez que los datos pasaron al sistema ETL, se realizan una serie de actividades, las cuales involucran limpieza de los datos y corrección de los

mismos con el objetivo de ser procesados y analizados posteriormente para evitar redundancia (Kimball & Ross, 2013).

Muchos de los subsistemas definidos se enfocan en el procesamiento de la tabla de dimensiones, realizando tareas como asignación de claves y búsqueda de códigos para una mejor presentación de los datos. También permite unir a las estructuras las tablas de forma normal subyacentes en dimensiones no normalizadas. Por el contrario, las tablas de hechos son típicamente grandes y requiere mayor tiempo para cargar (Kimball & Ross, 2013).

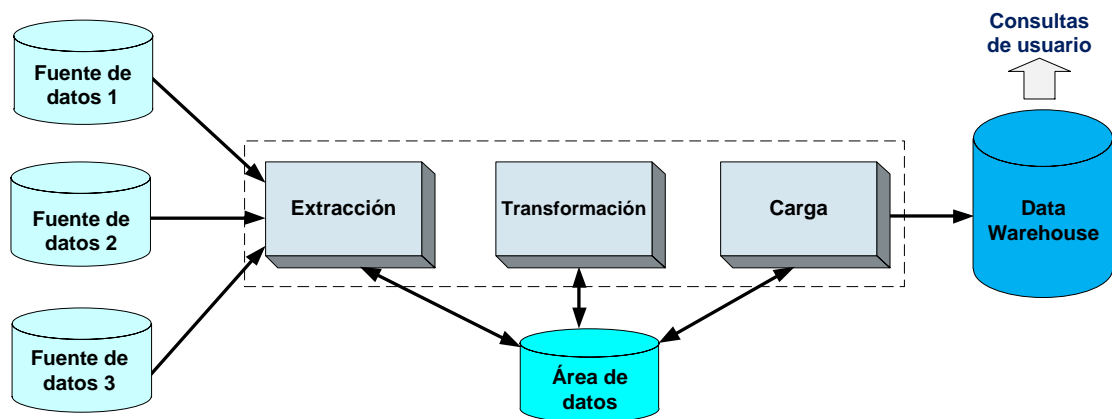


Figura 16. Proceso ETL. Fuente: (Chen, Baoran, & Yang, 2016)

a) Extracción

El proceso de extracción de datos se puede realizar manualmente o por medio de una aplicación ETL. El método manual requiere la programación de líneas de código en un lenguaje de programación con el objetivo de automatizar dicho proceso; las instrucciones permiten la extracción de los datos de las diversas fuentes de información. La finalidad de este proceso es lograr la mejor extracción de los datos de los sistemas transaccionales para luego integrarlos dentro del flujo ETL (Cano, 2007).

Durante la primera fase, los datos se extraen de las fuentes internas y externas disponibles. Es posible realizar una distinción lógica entre la extracción inicial,

donde los datos disponibles son relativos a todos los períodos pasados se introducen en el almacén de datos vacío, y las posteriores extracciones incrementales que actualice el almacén de datos utilizando nuevos datos que estén disponibles a lo largo del tiempo (Vercellis, 2009). El modelo seguido para la importación de los datos es el de Data Warehouse, el cual está sujeto a los requerimientos de información de la solución BI, lo cual es un reflejo de la necesidad de información de la organización para la toma de decisión.

b) Limpieza

Los datos obtenidos de los sistemas transaccionales necesitan ser limpiados antes de su procesamiento. El mercado actual, cuenta con herramientas ETL que facilitan este proceso. En proyectos de mayor complejidad como los CRM (Customer Relationship Management), es realmente importante contar con un proceso de limpieza de datos confiable. Los datos de las ventas, clientes, proveedores, pedidos, entre otros, deben carecer de redundancia. Si no se lleva a cabo este subproceso de forma cuidadosa, se perderá confiabilidad respecto a la información obtenida (Cano, 2007).

c) Transformación

La finalidad de la fase de limpieza y transformación es incrementar la calidad de los datos extraídos de las diversas fuentes de información. En este proceso se resuelven las inconsistencias e imprecisiones. Algunos de los aspectos identificados y rectificados en esta fase son los siguientes: datos duplicados, valores no permitidos o valores que fueron registrados en distintos atributos, pero que tienen el mismo significado (Vercellis, 2009).

En el proceso de limpieza se utilizan reglas automatizadas para la corrección y reducción de errores frecuentes. En muchos casos, se usan diccionarios con términos válidos para sustituir los términos supuestamente incorrectos, en función del nivel de similitud. Asimismo, durante la fase de transformación, las conversiones de datos adicionales pueden presentarse para garantizar la

estandarización de los datos. Por último, la agregación y consolidación de datos son realizados para obtener los resúmenes que reduce el tiempo de respuesta requerido por consultas y análisis posteriores para los cuales el almacén de datos fue destinado (Vercellis, 2009).

d) Carga

Es el último paso en su integración al DW. En esta fase se cargan los datos y se debe comprobar si la información coincide con lo existente en los sistemas operacionales o transaccionales de la organización. Asimismo, se realiza la validación de los valores, lo cuales deben mantener correspondencia con los determinados en la data warehouse. Esta fase es altamente sensible, pues de lo contrario puede contener información poco confiable en la solución BI que conlleve a errores en la toma de decisiones finales (Cano, 2007).

e) Actualización

En esta fase, se determina la frecuencia con la que se realizarán nuevas cargas de datos a la DW. Como en todo proyecto, esta variable es definida de acuerdo a los requerimientos organizacionales y de sus usuarios (Cano, 2007).

1.3.4. Minería de datos (DM)

La minería de datos es una parte fundamental del proceso BI, con la cual se obtiene información relevante. Esta información se obtiene mediante la extracción de grandes bases de datos y es utilizada para tomar decisiones importantes en los negocios. Dicho de otra forma, consiste en la obtención de conocimiento de volúmenes de datos. La minería de datos utiliza técnicas informáticas que son empleadas en la solución de problemas reales de gestión de información, la aplicación de dichas técnicas permite predecir tendencias y comportamientos para la toma de decisiones. Para ello, existen diversas técnicas, siendo algunas de las más importantes: agrupamiento automático, clasificación e identificación de patrones (Azoumana, 2013).

Los proyectos de minería de datos se diferencian en diversos aspectos tanto de la estadística clásica como de los análisis OLAP. La diferencia más importante consiste en la orientación activa que ofrecen los métodos de aprendizaje inductivo, en comparación con los métodos convencionales estadísticas y OLAP. Asimismo, en los análisis estadísticos, los responsables de la toma de decisiones formulan una hipótesis que luego debe ser confirmada sobre la base de la evidencia. (Vercellis, 2009).

La diferencia más importante consiste en la orientación activa que ofrecen los métodos de aprendizaje inductivo, en comparación con los métodos convencionales estadísticas y OLAP. Asimismo, en los análisis estadísticos, los responsables de la toma de decisiones formulan una hipótesis que luego debe ser confirmada sobre la base de la evidencia.

Las metodologías de minería de datos se pueden aplicar a una variedad de dominios, desde el control del proceso de comercialización y fabricación hasta el estudio de los factores de riesgo en el diagnóstico médico, desde la evaluación de la efectividad de nuevos medicamentos hasta la detección de fraudes.

Respecto a la representación de datos de entrada, en la mayoría de los casos, la entrada a un análisis de minería de datos toma la forma de una tabla bidimensional, la cual recibe el nombre de conjunto de datos, independientemente de la lógica real y la representación material adoptada para almacenar la información en archivos y bases de datos.

1.3.4.1. Métodos de minería de datos

El estudio de la minería de datos constituye un nuevo enfoque de análisis de datos, en el cual se utilizan algoritmos sofisticados para la detección de patrones y obtención de conocimiento. Las técnicas empleadas son altamente eficientes para la extracción de información de repositorios con grandes cantidades de datos. Las técnicas de minería de datos evolucionaron a lo largo de los años y

se integraron exitosamente con los procesos business intelligence. Algunas de las aplicaciones más utilizadas son el reconocimiento de patrones y desarrollo de sistemas expertos. Los algoritmos de minería se dividen en dos grupos: algoritmos supervisados, utilizados para la predicción y no supervisados, utilizados para la generación de conocimiento (Azoumana, 2013).

1.3.4.1.1. Métodos descriptivos

Los métodos descriptivos brindan una comprensión y entendimiento de los datos más eficiente. Para ello, utilizan técnicas descriptivas, las cuales incluyen el análisis de clúster. Una aplicación frecuente es la agrupación que se desarrolla en una base de datos de ventas, la cual pasa por un proceso de segmentación; en ella, se establecen las relaciones con otros elementos como productos y clientes. Las herramientas de minería de datos recorren las bases de datos para identificar modelos. Este tipo de métodos también es utilizado para la detección de fraudes de crédito u otras aplicaciones similares dentro del sector financiero (Azoumana, 2013).

1.3.4.1.2. Métodos predictivos

Son métodos orientados a la construcción de modelos de comportamiento. Mediante estos métodos es posible realizar predicción de comportamientos y valores de variables a partir de una muestra. Algunas de estas técnicas tienen como base el aprendizaje inductivo, en el cual el modelo es construido a través de la generalización de una muestra de entrenamiento (Azoumana, 2013). El principio que rige este modelo es que los resultados obtenidos del modelo entrenado pueden ser extrapolados a todo el conjunto de datos; para la verificación se utilizan las técnicas de la estadística convencional: bondad de ajuste y técnicas de predicción para grandes volúmenes de datos.

1.3.4.2. Técnicas de minería de datos

La clasificación tradicional de las técnicas de minería de datos nos brinda una primera clasificación en técnicas predictivas, en las cuales existe una distinción de las variables como dependientes e independientes y otra clasificación en técnicas descriptivas; en estas últimas, las variables tienen al inicio el mismo estado (Pérez & Santín, 2008). Bajo la clasificación anterior, se encuentran una variedad de técnicas. En primer lugar, dentro de los métodos predictivos están las técnicas de regresión, métodos bayesianos, series temporales, algoritmos genéticos, árboles de decisión, análisis de varianza y covarianza y redes neuronales. En segundo lugar, dentro de los métodos descriptivos se encuentran las técnicas de escalamiento multidimensional, segmentación, reducción de la dimensión, asociación y análisis exploratorio (Pérez & Santín, 2008).

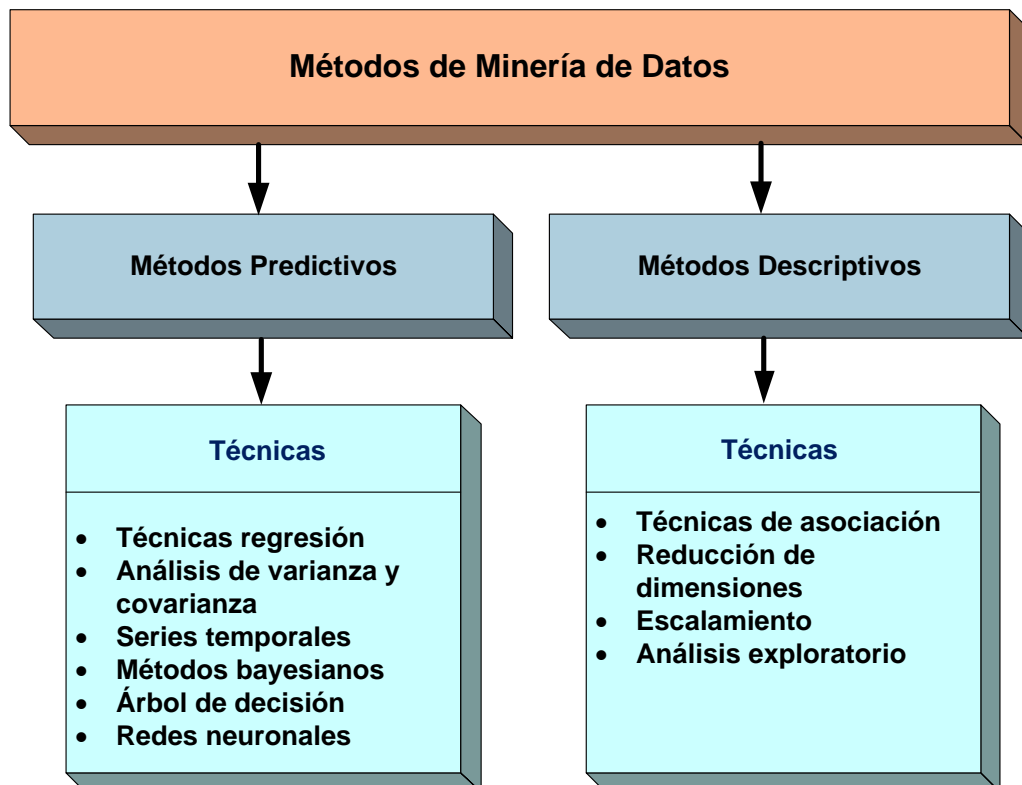


Figura 17. Métodos de minería de datos. Fuente: (Pérez & Santín, 2008)

A continuación, se presentan algunos de los algoritmos que fueron probados en soluciones business intelligence aplicando técnicas de minería de datos:

1.3.4.2.1. Método de clasificación: Árbol de decisión (DT)

El método de árbol de clasificación es ampliamente utilizado dentro de las técnicas de minería de datos; una de las razones de su amplio uso es la facilidad de comprensión de la lógica que utiliza, sus conceptos y las reglas generadas. Sus reglas permiten comprender las relaciones entre la variable de tipo objetivo y las variables predictivas (Vercellis, 2009). Este tipo de método se encuentra guiado por procedimientos heurísticos de árbol de decisión (DT). El árbol puede desarrollarse siguiendo distintas lógicas. A continuación, se presenta el procedimiento descendente de árboles de decisión: Primero, en la inicialización se agregan las observaciones al nodo raíz del árbol. Segundo, en caso de que la lista se encuentre vacía se detiene el procedimiento; en caso contrario, se elige un nodo de la lista y se elimina de ella. Tercero, se determinan y aplican las reglas de división; de esta forma los nodos son generados a partir de las observaciones divididas. Si el nodo cumple con las condiciones se convierte en hoja de tipo objetivo. Este procedimiento es recurrente (Vercellis, 2009).

La técnica de árbol de decisión es utilizada en el aprendizaje supervisado, y consiste en la división o segmentación de un conjunto de datos recurrentes ejemplo (Azoumana, 2013). En la siguiente imagen se muestra la estructura de un árbol de decisión.

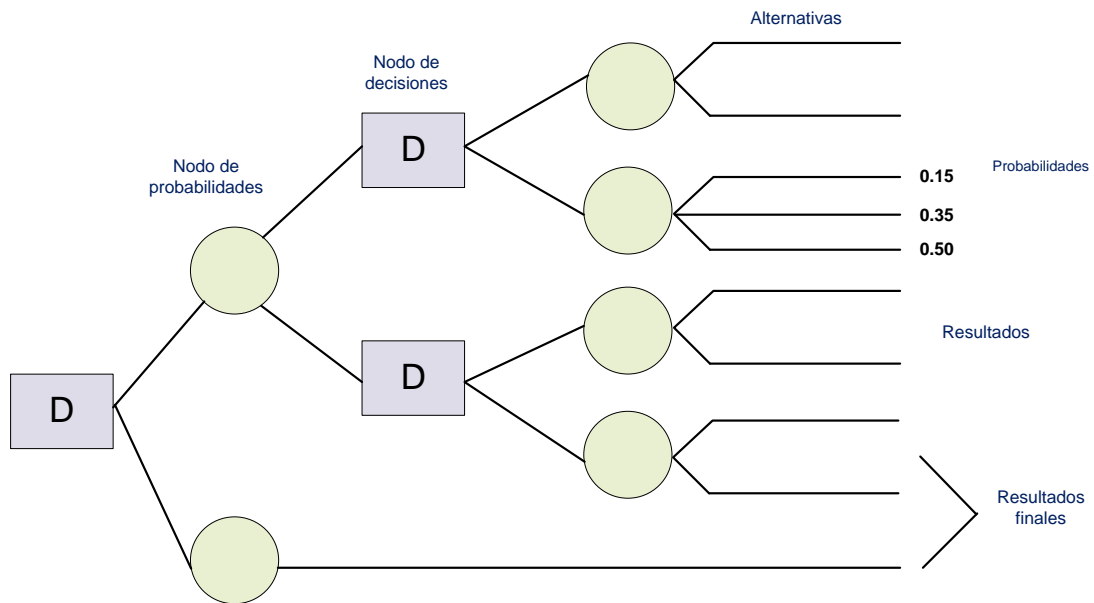


Figura 18. Estructura de un árbol de decisión. Fuente: (Vercellis, 2009).

Asimismo, el algoritmo base DT puede ser representado mediante las siguientes líneas de pseudocódigo:

```

Input: D conjunto de N patrones etiquetados, cada uno de los
cuales está caracterizado por n variables de predicción
 $X_1 \dots X_n$  y la variable clase C.
Output: Árbol de clasificación
Begin: DT
  If (todos los patrones de D pertenecen a la misma clase C)
    Then
      Resultado de la inducción es un nodo simple
      (hoja) etiquetado como C
    Else
      Begin
        1. Seleccionar la variable más informativa  $X_r$ 
        con valores  $x_r^1 \dots x_r^{n_r}$ 
        2. Particionar D de acuerdo a los  $n_r$  valores
        de  $X_r$  en  $D_1 \dots D_{n_r}$ 
        3. Construir  $n_r$  sub árboles  $T_1 \dots T_{n_r}$  para
         $D_1 \dots D_{n_r}$ 
        4. Unir  $X_r$  y lo  $n_r$  subárboles  $T_1 \dots T_{n_r}$  con los
        valores  $T_1 \dots T_{n_r}$ 
      End
    Endif
  End DT

```

Figura 19. Pseudocódigo del algoritmo árbol de decisión (DT). Fuente: (Vercellis, 2009)

1.3.4.2.2. Método bayesiano: Naive Bayes (NB)

Los métodos bayesianos pertenecen a los modelos de clasificación basada en probabilidades. Se calculan las probabilidades de la observación mediante el teorema de Bayes. La particularidad de este método es que requiere que el usuario calcule la probabilidad de ocurrencia de una observación. En estos casos, la etapa de aprendizaje de un clasificador bayesiano se vincula con el análisis inicial de las observaciones de un grupo de entrenamiento para una posterior clasificación (Vercellis, 2009).

El teorema Bayes es utilizado para la estimación de la probabilidad posterior $P(y|x)$. Sin embargo, la estimación muestral equivalente de las probabilidades de las clases antes mencionadas no se puede conseguir debido a la complejidad en términos computacionales y al requerimiento de muestra (Vercellis, 2009). Con el objetivo de superar la dificultad anterior, se realiza la introducción de hipótesis que conducen a clasificadores Naive Bayes. El modelo propuesto en este teorema, puede representarse matemáticamente de la siguiente manera:

$$p(C|F_1, \dots, F_n) = \frac{p(C) p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}$$

El objetivo del modelo es reformular la tabla de probabilidades para hacerlo más manejable, la fórmula anterior podría simplificarse entendiendo la probabilidad posterior (p) como el resultado de dividir (anterior x probabilidad) entre Evidencia. La característica más resaltante de este clasificador es la suposición de que cada atributo es independiente. De esta manera, la hipótesis asumida da origen a una red bayesiana (Azoumana, 2013).

1.3.4.2.3. Método de partición K-Means

Los métodos de partición comienzan con una asignación inicial de las observaciones de una agrupación (K). Después, se utiliza una técnica de

reasignación con el objetivo de ordenar las observaciones en un grupo distinto para mejorar su calidad (Vercellis, 2009).

El método K-Means es conocido por la eficiencia del algoritmo que utiliza, el cual es un algoritmo de partición. Este algoritmo recibe como input un conjunto de datos (D), un número de clúster a generar (K) y una función, la cual expresa la matriz de distancias (D) entre las observaciones (Vercellis, 2009). Los objetos son representados mediante vectores de dimensiones (d). La representación matemática del algoritmo es la siguiente:

$$\min_s E(\mu_i) = \min_s \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

En donde S es el grupo de datos cuyos elementos que lo conforman son los objetos X_j . El objetivo del algoritmo es construir grupos (K) con el objetivo de minimizar la suma de distancias de los objetos dentro de cada conjunto.

El procedimiento del algoritmo K-means es el siguiente: Primero, en la etapa inicial o de inicialización se eligen de forma aleatoria K observaciones en D (dataset) como centro de las agrupaciones. Segundo, las observaciones son asignadas de forma iterativa al grupo cuyo centro presenta mayor similitud con la observación y con ello minimizar la distancia del registro. En caso de no asignar ninguna observación a un grupo, el algoritmo se detiene. Finalmente, para cada uno de los grupos, el nuevo centro se obtiene como la media de los valores obtenidos en las observaciones; luego regresa al paso 2 y ejecuta el mismo flujo (Vercellis, 2009).

1.3.4.2.4. Regresión logística

La técnica de regresión logística tiene principios estadísticos. Consiste en un tipo de análisis de regresión aplicado a la predicción de variables que pueden adoptar un número limitado de características. Este tipo de técnica es utilizado

ampliamente en aprendizaje máquina y minería de datos para la predicción. Su representación matemática es la siguiente:

$$p(x) = \frac{1}{1 + e^{-(B_0 + B_1x)}}$$

En la fórmula anterior, $p(x)$ la probabilidad de éxito cuando el valor de la variable de predicción es x . $e^{-(B_0 + B_1x)}$ es la razón de momios o razón de probabilidades, la cual se encuentra comprendido entre 0 y 1.

1.3.4.2.5. Máquinas de vectores de soporte (SVM)

La técnica de minería de datos SVM consiste en un conjunto de algoritmos de aprendizaje supervisado. El método que utiliza está relacionado a los problemas de clasificación y regresión, mediante el cual se brinda un conjunto de entrenamiento denominado muestra con la finalidad de realizar predicciones.

De forma más específica, la técnica SVM construye un conjunto de planos en un espacio dimensional que puede ser usado para resolver problemas de clasificación y regresión. Este modelo está directamente relacionado con las redes neuronales.

Dentro del espacio p -dimensional, un hiperplano se determina como un subespacio plano y afín de dimensiones $p-1$. La palabra afín indica que el subespacio no debe pasar por el origen. La representación matemática de un hiperplano de dos dimensiones es la siguiente:

$$B_0 + B_1x_1 + B_2x_2 = 0$$

Considerando los parámetros B_0, B_1 y B_2 todos los pares de valores $x = (x_1, x_2)$ para los que se cumplen con la igualdad serán los puntos del hiperplano.

1.4. Formulación del Problema.

Después de realizar un análisis de los conceptos y modelos relacionadas a business intelligence y la importancia de la aplicación de minería de datos, lo cual fue revisado en los antecedentes de estudio, la presente investigación propuso aplicar las técnicas de minería de datos para la predicción de rendimiento académico de estudiantes de una universidad peruana y responder la siguiente pregunta ¿Cuál es la técnica de minería de datos de mejor rendimiento aplicada a una solución business intelligence para predecir el rendimiento académico?

1.5. Justificación e importancia del estudio.

1.5.1. Justificación teórica

Desde un enfoque teórico, la investigación se justifica, ya que se utilizaron los conceptos y principios que sustentan el desarrollo de soluciones business intelligence (BI), principalmente los desarrollados por Kimball e Inmon. Profundiza en el análisis de la minería de datos (DM) aplicadas a este tipo de soluciones y permite determinar las técnicas de mejor rendimiento. Contribuye a su profundidad teórico, ya que existen diversas investigaciones relacionadas a su aplicación en casos de estudio con el objetivo de obtener predicciones y anticipar escenarios. No obstante, la literatura en torno a su uso en el desarrollo de soluciones business intelligence es limitada. Es presente estudio busca contribuir al conocimiento de esta área, brindando un marco de referencia sobre la utilización de las técnicas de minería de datos en soluciones BI, lo cual tiene impacto en el proceso de toma de decisiones de las empresas.

1.5.2. Justificación práctica

A nivel práctico, la investigación contribuye a que los desarrolladores de soluciones business intelligence tengan un marco de referencia al momento de realizar la elección de una técnica de minería de datos, ya sea predictiva o

descriptiva. La elección adecuada de una técnica de minería de datos permite que la solución final tenga una mayor utilidad para el usuario final y así optimizar la toma de decisiones dentro de una organización. Con el incremento de las organizaciones, volúmenes de datos y desarrollo de soluciones para la toma de decisión, Es necesario contar con procedimientos que garanticen la elección adecuada de técnicas que faciliten el análisis de los datos obtenidos de diversas fuentes de información.

1.5.3. Justificación metodológica

La investigación cuenta con una utilidad metodológica, ya que a partir del método propuesto para la selección y evaluación de las técnicas que serán implementadas en la realización de pruebas, permite obtener resultados objetivos y confiables respecto a su rendimiento en soluciones de inteligencia de negocios. Asimismo, permite validar la utilización de métricas para su evaluación, las mismas que podrán ser replicadas en otros estudios de similares características. De esta forma, la metodología seguida contribuye a la reducción de la brecha de conocimiento relacionada a la aplicación de las técnicas de minería de datos.

1.6. Hipótesis.

La mejor técnica de minería de datos aplicada a una solución business intelligence para predecir el rendimiento académico es Naive Bayes.

1.7. Objetivos.

1.7.1. Objetivo general.

Analizar comparativamente el rendimiento de técnicas de minería de datos aplicadas a una solución business intelligence para predecir el rendimiento académico.

1.7.2. Objetivos específicos.

- a) Seleccionar dos técnicas de minería de datos con mejor rendimiento utilizadas en sistemas para la toma de decisión.
- b) Determinar los indicadores de evaluación de las técnicas de minería de datos que serán utilizadas en la realización de pruebas.
- c) Definir el método business intelligence para la aplicación de las técnicas de minería de datos
- d) Proponer una base de datos para la realización de las pruebas.
- e) Implementar las técnicas de minería de datos seleccionadas para el análisis de datos obtenidos del proceso business intelligence.
- f) Comparar y evaluar los resultados de las técnicas de minería de datos utilizadas.

II. MATERIAL Y MÉTODO

2.1. Tipo y Diseño de Investigación.

2.1.1. Tipo de investigación

La investigación es de tipo aplicada, ya que utiliza el conocimiento científico existente sobre base de datos y técnicas de minería de datos, los cuales son cuantificables y pueden ser aplicados en el desarrollo de soluciones informáticas que requieran el procesamiento y análisis de grandes volúmenes de datos.

El campo de aplicación específico es el desarrollo de sistemas de información de soporte a la toma de decisión, dentro de los que se encuentran las soluciones business intelligence, los cuales utilizan las técnicas de minería de datos para la obtención de conocimiento mediante el procesamiento automático de los datos recopilados de diferentes fuentes de información.

2.1.2. Diseño de investigación

El diseño de la investigación es del tipo cuasi experimental, ya que la población de la investigación y la muestra fueron determinadas por el investigador con base a la existencia actual de técnicas de minería de datos, las cuales fueron obtenidas de diversos trabajos de investigación en el campo de la ingeniería de sistemas.

Asimismo, se realizó la manipulación controlada con el objetivo de medir el rendimiento de las técnicas de minería de datos aplicadas a business intelligence y realizar un análisis comparativo de sus resultados. Para ello, se han utilizado métricas estandarizadas y validadas en investigaciones de ingeniería de sistemas y computación e informática.

2.2. Población y muestra.

2.2.1. Población

La población está compuesta por ocho técnicas de minería de datos utilizadas en sistemas de soporte para la toma de decisión o business intelligence. Para la obtención de la población se revisaron artículos científicos de las bases de datos científicas IEEE y ScienceDirect, de las cuales se obtuvieron los resultados de su aplicación. En la siguiente tabla se presenta la relación de las técnicas de minería de datos.

Tabla 1

Población de técnicas de minería de datos

N°	Técnica de minería de datos
1	Árbol de Decisión (DT)
2	Naive Bayes (NB)
3	K Nearest Neighbor (KNN)
4	J-RIP
5	J48
6	Redes Neuronales (RN)
7	Random Tree (RT)
8	Support Vector Machine (SVM)

Fuente: Elaborado por el autor con base a las investigaciones obtenidas de las bases de datos científicas IEEE y ScienceDirect.

2.2.2. Muestra

La muestra de la investigación fue obtenida bajo el método no probabilístico del tipo muestreo por conveniencia; para ello, se tomó en consideración los resultados en aplicaciones previas. El proceso de selección de la muestra inició con la investigación de técnicas de minería de datos en bases de datos

científicas y posteriormente se procedió con la tabulación de los resultados bajo la clasificación de bajo rendimiento, rendimiento aceptable y buen rendimiento.

En ese sentido, la muestra está compuesta por dos algoritmos de minería de datos: Naive Bayes (NB) y Árbol de Decisión (DT), las cuales fueron utilizadas en más del 50% de las 21 investigaciones consultadas, las cuales se aprecian posteriormente en la Tabla N° 05. Dichas técnicas obtuvieron un rendimiento superior al 90% de precisión en el análisis de datos.

Asimismo, la elección de la muestra para el método no probabilístico por conveniencia se sustenta en la disponibilidad de las técnicas de minería de datos en diversos lenguajes de programación, lo cual permite su implementación para la realización de pruebas y obtención de resultados para su comparación. Entre algunos de los lenguajes que permiten codificar ambas técnicas de minería de datos se encuentran: Python, JAVA, C# y R.

2.3. Variables, Operacionalización.

2.3.1. Variables

2.3.1.1. Variable independiente

Técnicas de minería de datos

2.3.1.2. Variable dependiente

Solución business intelligence

2.3.2. Operacionalización de variables

Tabla 2

Operacionalización de variable independiente

Variable	Dimensión	Indicadores	Ecuación	Descripción
Variable independiente: Técnicas de minería de Datos	Error y Precisión	Error absoluto medio (MAE)	$MAE = \frac{\sum_{i=1}^n y_i - x_i }{n}$	Diferencia absoluta entre el valor objetivo y el valor obtenido
		Error cuadrático medio (MSE)	$MSE = \frac{\sum \text{Error de pronóstico}^2}{n}$	Dispersión del error pronóstico
		Error absoluto relativo (MAPE)	$MAPE = \frac{\sum_{i=1}^n Real_i - Pronóstico_i }{Real_i \cdot n}$	Desviación del error a nivel porcentual
		Precisión	$Precisión = \frac{TP}{TP + FP}$	Porcentaje de predicciones correctas

Fuente: Elaborado por el autor con base a métricas estándares para medición de error y obtención de indicadores business intelligence.

Tabla 3

Operacionalización de variable dependiente

Variable	Dimensión	Indicadores	Ecuación	Descripción
Variable dependiente: Business intelligence	KPI's del negocio	Procesamiento de datos en Cifra por Carreras (CA)	<p>KPI</p> $CA = \sum \text{Aprobados por carrera}$ <p>Procesamiento %:</p> $\% \text{ Proc. éxito CA} = \frac{\text{Total datos proces.}}{\text{Total de datos}} * 100$	Explica el procesamiento total de datos del total de aprobados por carrera
		Procesamiento de datos en Cifra de Modalidad (CM)	<p>KPI</p> $CA = \sum \text{Aprobados por modalidad}$ <p>Procesamiento %:</p> $\% \text{ Proc. éxito MC} = \frac{\text{Total datos proces.}}{\text{Total de datos}} * 100$	Explica el procesamiento total de datos del total de aprobados por modalidad
	Tiempo	Tiempo promedio de proceso	$TP = \frac{\sum \text{Tiempo observado}}{\text{Número de observaciones}}$	Tiempo del proceso para la obtención de datos de entrada

Fuente: Elaborado por el autor con base a métricas estándares para medición de error y obtención de indicadores business intelligence

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

En el presente trabajo de investigación, se utilizó la técnica de observación y como instrumento la ficha técnica. A continuación, se explica en qué consiste dicha técnica y el instrumento.

2.4.1. Técnica de recolección de datos

2.4.1.1. Observación

Esta técnica consiste en la observación del fenómeno o hecho para registrar la información y analizarla; permite recabar un mayor número de datos respecto a las variables de estudio. En este caso, se busca observar el rendimiento de las técnicas de minería de datos para realizar posteriormente el análisis comparativo de los resultados.

2.4.2. Instrumento de recolección de datos

2.4.2.1. Ficha de registro

El instrumento utilizado en la presente investigación fue la ficha de registro de información, el cual es un instrumento prediseñado en el que se detallan los aspectos a observar. Su finalidad es tener un registro ordenado de los datos. En este caso, se registran los resultados de obtenidos de las técnicas de minería de datos implementadas, de acuerdo a las métricas establecidas para evaluar su rendimiento. El instrumento utilizado se puede observar en el Anexo 3.

2.5. Procedimiento de análisis de datos.

Para el análisis de datos se consideraron los métodos estadísticos de promedio simple, error absoluto, error medio cuadrado y error absoluto relativo. Para el procesamiento de los datos se utilizaron los softwares estadísticos Statistical Package for the Social Sciences (SPSS) y MS Excel.

Asimismo, para la presentación de los datos se utilizaron las tablas con los resultados obtenidos e histogramas con valores relativos y acumulados; dichos gráficos serán obtenidos de los softwares estadísticos.

2.6. Criterios éticos.

a) Veracidad

Las técnicas e instrumentos para la recolección de datos fueron aplicados de acuerdo a procedimientos que garantizan la objetividad de los resultados. Ambas técnicas de minería de datos fueron implementadas bajo un entorno de iguales condiciones para obtener de forma objetiva su rendimiento. Asimismo, la información consignada a lo largo del trabajo de investigación puede ser corroborada desde sus fuentes primarias.

b) Originalidad

En la presente investigación se citaron la totalidad de las fuentes consultadas, las cuales también se encuentra en la bibliografía general. Las ideas consignadas pueden ser verificables desde los documentos originales. Asimismo, el trabajo se sostiene en su originalidad, no siendo replica total ni parcial de ninguna otra investigación.

c) Confidencialidad

Durante el proceso de investigación, realización de pruebas se mantuvo la privacidad de la información y reserva de datos que contenía la base de datos de prueba.

2.7. Criterios de Rigor Científico.

a) Validez

La operacionalización de las variables independiente y dependiente y sus dimensiones descritas, se evaluaron de acuerdo a lo establecido en las técnicas de recolección y procesamiento de datos. Asimismo, el análisis e interpretación de los datos son totalmente objetivos.

b) Fiabilidad

Para asegurar la fiabilidad de los resultados obtenidos se estableció un entorno de pruebas que cumpla con las condiciones necesarias, teniendo como referencia los métodos y procedimientos seguidos en investigaciones similares. Para ello, se utilizaron indicadores y procedimientos verificables.

c) Objetividad

Los datos obtenidos del proceso de experimentación de las técnicas de minería de datos se encuentran libres de cualquier apreciación subjetiva del investigador y sus conclusiones se basan plenamente en los resultados obtenidos en la fase de prueba.

III. RESULTADOS.

3.1. Resultados en Tablas y Figuras.

La presente investigación propuso como objetivo comparar el rendimiento de las técnicas de minería de datos en una solución business intelligence, de la cual se obtuvieron los datos de entrada los cuales fueron procesados por las técnicas Árbol de decisiones y Naive Bayes para la predicción de rendimiento académico. En tal sentido, se presentan a continuación los resultados de forma comparativa:

El primer indicador es MAE es la diferencia absoluta ente el valor real y el valor obtenido en la predicción del modelo. En la siguiente ecuación, y_i es la predicción del modelo y x_i es el valor verdadero.

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

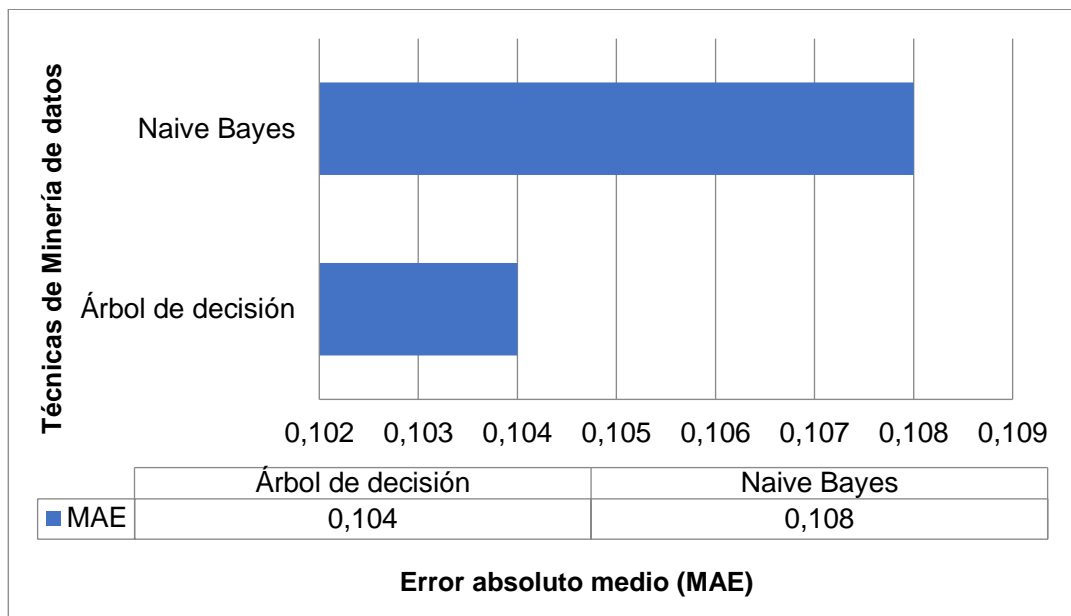


Figura 20. Resultados MAE obtenidos del modelo de predicción de rendimiento académico aplicando técnicas de minería de datos. Fuente: Elaboración propia.

La figura anterior muestra el error absoluto medio de ambas técnicas aplicadas al modelo de predicción de rendimiento académico. De esta forma, se cuantifica la precisión de la técnica, siendo la técnica de Árbol de decisión la que obtuvo menor MAE 0.104. Sin embargo, representa una diferencia poco significativa de 0.004.

El segundo indicador es MSE, el cual es una medida de dispersión que permite calcular el error cuadrático medio. Dicho de otra forma, mide la diferencia entre el estimador y lo que se estima. La fórmula utilizada para el cálculo es la siguiente:

$$MSE = \frac{\sum Error\ de\ pronóstico^2}{n}$$

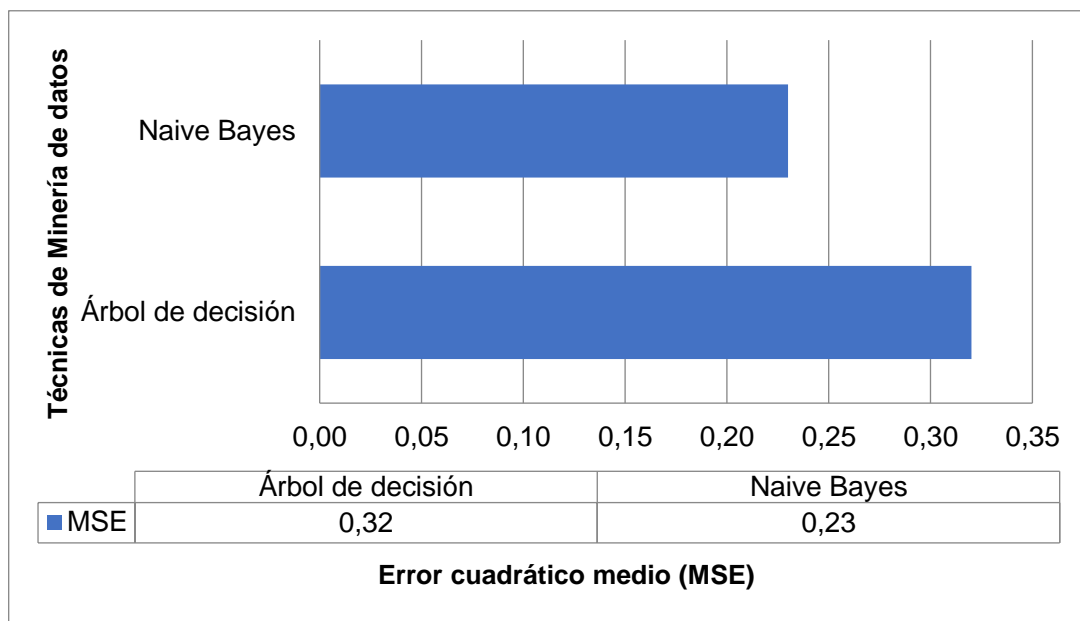


Figura 21. Resultados MSE obtenidos del modelo de predicción de rendimiento académico aplicando técnicas de minería de datos. Fuente: Elaboración propia.

La figura anterior muestra el error cuadrático medio de ambas técnicas. El modelo implementado obtuvo un mayor MSE en la técnica Árbol de decisión. Si bien se espera un error cercano a 0; en un análisis comparativo la técnica Naive Bayes tiene un mejor rendimiento con 0.23.

El tercer indicador es MAPE, el cual permite calcular el error porcentual medio. Consiste en la diferencia entre el valor real y valor esperado o pronóstico, pero expresado en términos porcentuales. La fórmula para el cálculo es la siguiente:

$$MAPE = \frac{\sum_{i=1}^n |Real_i - Pronóstico_i|}{Real_i} \cdot \frac{1}{n}$$

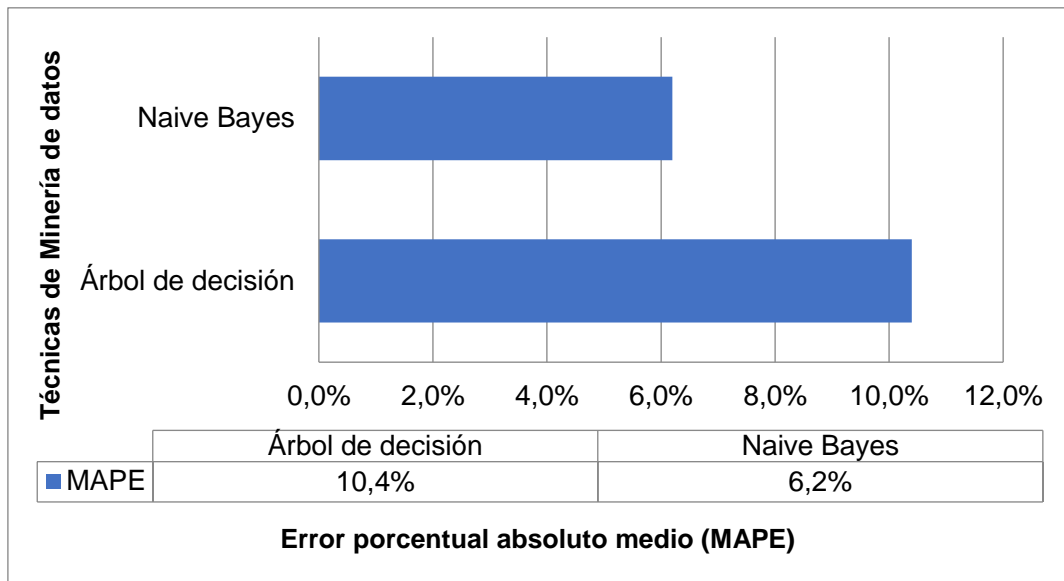


Figura 22. Resultados MAPE obtenidos del modelo de predicción de rendimiento académico aplicando técnicas de minería de datos. Fuente: Elaboración propia.

La figura anterior muestra el error porcentual absoluto medio. La técnica con menor MAPE es Naive Bayes con 6.2%. Esto evidencia que el modelo implementado y procesado bajo esta técnica obtiene un menor tamaño de error en comparación con Árbol de decisiones.

Asimismo, se consideró el Tiempo de Procesamiento (TP); la fórmula para el cálculo está basado en la sumatoria del tiempo promedio observado dividido entre el número de observaciones:

$$TP = \frac{\sum \text{Tiempo observado}}{\text{Número de observaciones}}$$

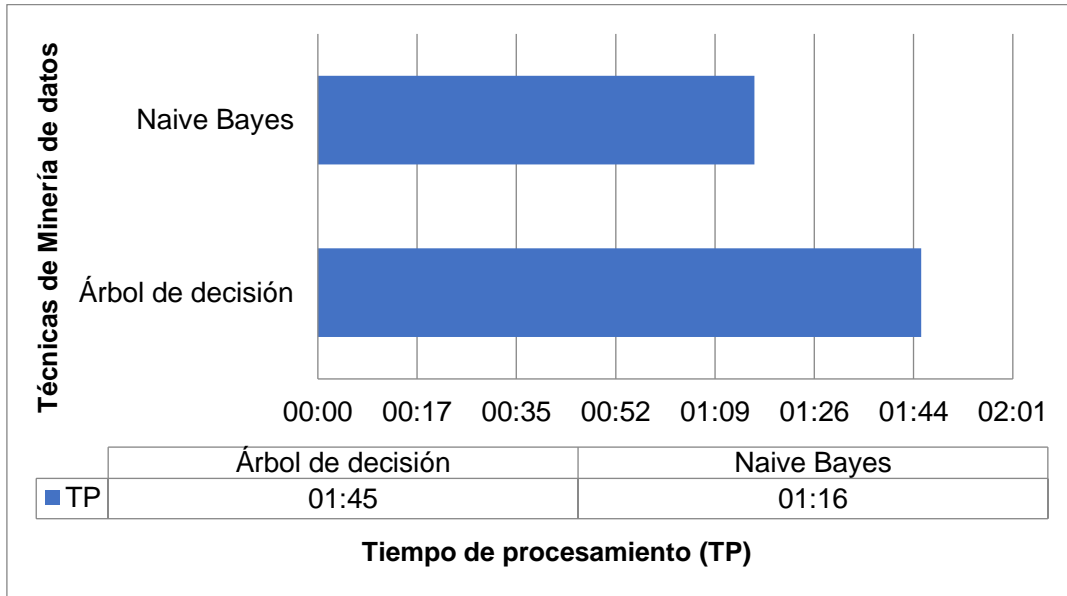


Figura 23. propia Resultados Tiempo de Procesamiento obtenidos de la observación durante la aplicación de las técnicas de minería de datos. Fuente: Elaboración.

En la figura anterior muestra el tiempo de procesamiento de las técnicas de minería de datos, constituye el tiempo de procesamiento de los datos destinados para entrenamiento y pruebas. En tal sentido, el tiempo del proceso observado fue menor en la técnica Naive Bayes. La información respecto a las características del equipo utilizado para la realización de las pruebas se encuentra detallada en el Anexo 01.

Otro indicador a considerar es el porcentaje de procesamiento de los datos utilizados para la obtención del indicador de Cifras por Modalidad (CM), cuyos gráficos estadísticos obtenidos a partir de la solución business intelligence se encuentran en las figuras 38 y 39.

$$\% \text{ Proc. éxito MC} = \frac{\text{Total datos proces.}}{\text{Total de datos}} * 100$$

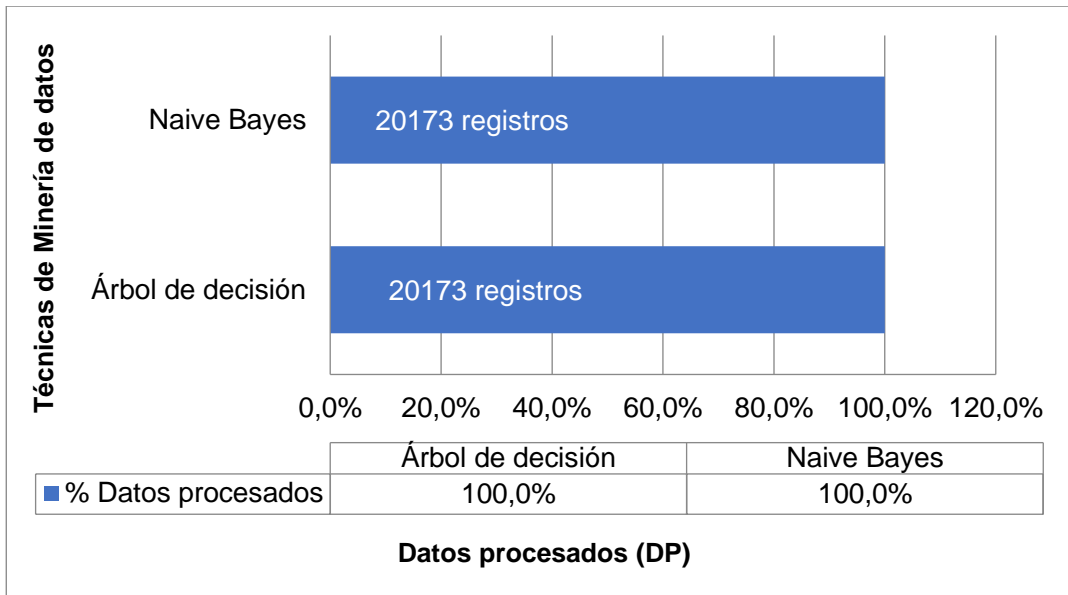


Figura 24. Porcentaje de datos procesados en la aplicación de las técnicas de minería de datos. Fuente: Elaboración propia.

Como se puede apreciar en la figura anterior, en ambas técnicas se obtuvo un porcentaje de datos procesados de 100%. El principal motivo de haber obtenido un porcentaje tal alto fue la ejecución del proceso ETL en la solución business intelligence.

Finalmente, con la finalidad de comparar los resultados contenidos con los obtenidos en investigaciones que han utilizado técnicas de minería de datos, se ha considerado el cálculo de la precisión, el cual indica el porcentaje de predicciones correctas, donde TP son los valores verdaderos positivos y FP son los valores falsos positivos, los cuales se encuentran en la matriz de confusión. La fórmula es la siguiente.

$$Precisión = \frac{TP}{TP + FP}$$

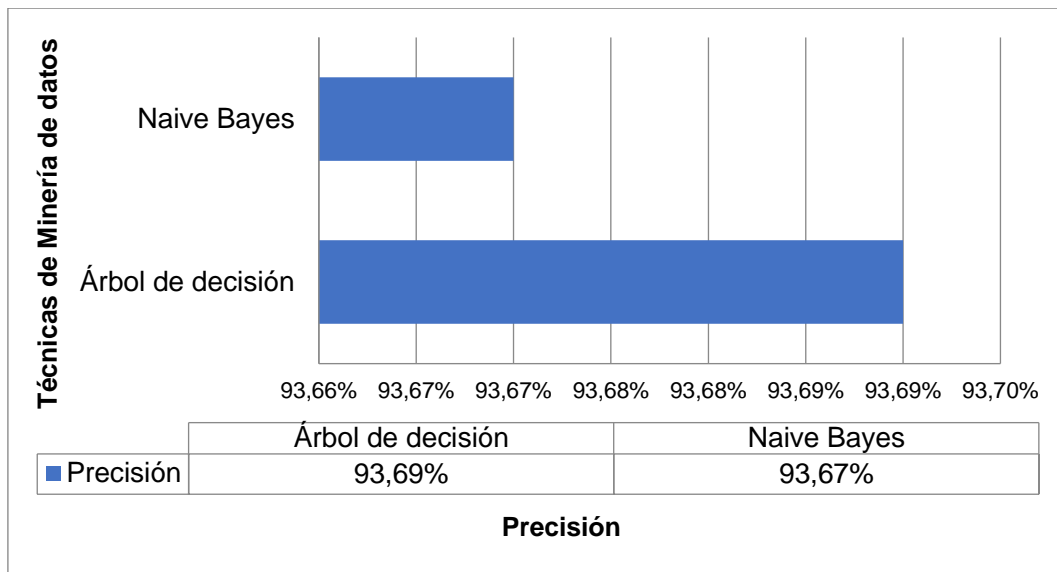


Figura 25. Resultados precisión obtenidos de la aplicación de las técnicas de minería de datos utilizando matriz de confusión. Fuente: Elaboración propia.

Como se puede observar en el gráfico anterior, la técnica Árbol de decisión obtuvo una precisión de 93.69% para predecir el rendimiento académico de los alumnos, la cual es superior en 5.1%. Teniendo en consideración que ambas técnicas fueron aplicadas al mismo caso de estudio con un total de 20,173 registros obtenidos del proceso business intelligence.

3.2. Discusión de resultados.

Los resultados obtenidos permitieron realizar comparaciones con investigaciones que aplicaron técnicas de minería de datos, específicamente Árbol de decisiones y Naive Bayes. Los resultados de Schuh, Prote, & Hünnekes (2020) utilizando el método de validación cruzada mostró una precisión del 90.6%. Asimismo, los resultados de Viloría, Rodríguez, Payares y Vargas (2019) aplicando la misma técnica de minería de datos, Árbol de decisiones, a datos de una organización educativa para predecir la interacción en entornos virtuales obtuvo una precisión de 92%. El modelo implementado a partir de la obtención de datos de entrada mediante un proceso business intelligence permitió obtener una precisión del 93.69% en la predicción del rendimiento académico de los estudiantes aplicando Árbol

de decisiones. Si bien el modelo implementado evidencia una precisión mayor, es importante considerar que son dos factores los que influyen ello. Por un lado, la mejor calidad de los datos a partir de los procesos ETL y construcción de la base de datos dimensional, lo cual permitió obtener un conjunto de datos de entrada para los procesos de entrenamiento y prueba. Por otro lado, la correcta selección de los variables del modelo que conformaron el dataset, dentro de las que se encontraba el promedio y la modalidad de estudio, siendo en este caso Presencial y Semipresencial.

Con respecto a la técnica de minería de datos Naive Bayes, existe también un mejor rendimiento en comparación a los resultados obtenidos en investigaciones realizadas en otros contextos. La investigación realizada por Ghazzawi & Alharbi (2019), en la cual aplicaron dicha técnica para el análisis de reclamos de clientes obtuvo una precisión de 86.5%. Asimismo, la investigación de Parama (2018) en la que utilizó técnicas de minería de datos a un modelo de inteligencia de negocios obtuvo una precisión de 74.6%. El modelo implementado aplicando la técnica de minería de datos Naive Bayes muestra un resultado de precisión de 93.67%; dichos resultados fueron obtenidos mediante matriz de confusión. Si bien el modelo implementado obtenido a partir de datos de entrada mediante un proceso business intelligence es superior a investigaciones de contextos similares. No obstante, un factor que influye es el volumen de datos, ya que en la etapa experimental se procesaron únicamente 20,173 registros, lo cual constituye un número menor en comparación a otras investigaciones.

Finalmente, las métricas obtenidas de cada técnica permitieron realizar la comparación de su rendimiento aplicadas al mismo caso de estudio, siendo en este caso la predicción de rendimiento académico. Primero, el error medio absoluto (MAE), es menor en la técnica de Árbol de decisión 0.104. Sin embargo, no es una diferencia realmente significativa en comparación con Naive Bayes, ya que solo se diferencia en 0.004. Segundo, en relación el error cuadrático medio (MSE), la técnica de minería de datos Naive Bayes evidenció un rendimiento menor con 0.23. Finalmente, el análisis del error porcentual absoluto medio (MAPE), mostro que la técnica con menor MAPE

fue Naive Bayes, la cual obtuvo 6.2%. El modelo de predicción implementado, en la cual se aplicaron las técnicas de minería de datos al caso de estudio, presentaría mejor rendimiento al aplicar la técnica Naive Bayes.

3.3. Aporte práctico.

Proceso de selección de técnicas de minería de datos

Para la selección de las técnicas de minería de datos, se realizó un proceso de investigación en las bases de datos especializadas IEEE Xplore y ScienceDirect. Las investigaciones seleccionadas contemplan resultados obtenidos mediante un proceso de experimentación en el área de ingeniería en los últimos cinco años (2015 – 2020), ya sea a nivel comparativo de dos o más técnicas de minería de datos o aplicada a un caso de estudio.

Las bases de datos especializadas fueron seleccionadas, por su prestigio y contribución en el campo de la Ingeniería. Asimismo, se consideró el número de publicaciones vinculadas al tema de investigación bajo las palabras claves “Data mining” y “Data mining techniques”. Por un lado, IEEE Xplore cuenta con más de 300 revisas indexadas a su base de datos, teniendo como principales contribuciones las investigaciones del Institute of Electrical and Electronics (IEEE); en esta base de datos se encontraron más de 20,000 publicaciones relacionadas. Por otro lado, ScienceDirect cuenta con más de 2,500 revisas indexadas a su base de datos bajo el modo de revisión por pares y cuenta con una sección principal de ingeniería, en la que se obtuvo más de 5,000 resultados relacionados al tema de investigación.

Para la selección de las investigaciones, se consideró investigaciones aplicadas publicadas desde el año 2015 hasta el 2020. Asimismo, se requería que dichos artículos describieran claramente el método utilizado para la obtención de los resultados. De esta forma, se redujo a 21 investigaciones, cuyos resultados fueron organizados para la selección de la muestra. Los resultados de las investigaciones consideradas fueron tabulados en una matriz y posteriormente se realizó una selección de dos

técnicas con base a los mejores rendimientos evidenciados. El proceso de selección de muestra en la siguiente figura:

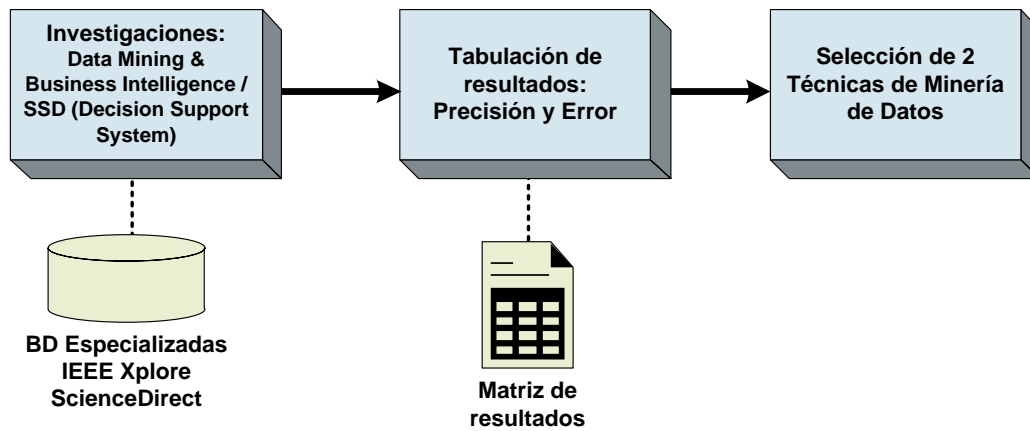





Figura 26. Proceso de selección de técnicas de minería de datos. Fuente: Elaboración propia

Para la construcción de la matriz, se consideró la siguiente información: Algoritmo utilizado en la técnica de minería de datos, investigador, año de la investigación, escenario de aplicación, resultados obtenidos y rendimiento promedio de la técnica, el cual fue calculado mediante promedio simple del rendimiento obtenido por cada tipo de algoritmo: $RP = \frac{\sum \text{Rendimiento obtenido}}{\text{total de investigaciones}}$ el cálculo se realizó por cada una de las técnicas. Asimismo, para brindar una calificación cualitativa a los resultados, se propusieron tres escalas, teniendo en consideración los hallazgos de las investigaciones realizadas.

Las técnicas de minería de datos que obtuvieron un rendimiento menor o igual a 80% ($\text{Rendimiento} \leq 80\%$) fueron consideradas de “Bajo rendimiento”, mientras que aquellas con resultados mayor a 80% y menor o igual a 90% ($\text{Rendimiento} > 80\% \text{ y } \leq 90\%$) se fueron consideradas de “Rendimiento aceptable”; por último, aquellas técnicas que obtuvieron un resultado superior al 90% en su rendimiento ($\text{Rendimiento} > 90\%$) fueron consideradas de “Buen rendimiento”. En la siguiente tabla se presenta la escala de valoración considerada para la evaluación de su rendimiento.

Tabla 4

Calificación de las técnicas de minería de datos en publicaciones científicas de los últimos 5 años (2015- 2020)

Rendimiento	Calificación	Color
$\leq 80\%$	Bajo rendimiento	
$> 80\%$ y $\leq 90\%$	Rendimiento aceptable	
$> 90\%$	Buen rendimiento	

Fuente: Elaboración propia

A continuación, se presenta la tabla en la cual se organizaron los resultados de las investigaciones más relevantes encontradas en las bases de datos académicas descritas previamente.

Tabla 5

Matriz de evaluación de técnicas de minería de datos resultantes del proceso de selección de investigaciones publicadas del 2015 al 2020

Técnica de Minería de datos	Investigador y año de publicación	Escenario de aplicación	Eficiencia de rendimiento	Calificación	Rendimiento promedio
Árbol de Decisión Decision Tree (DT)	(Siyuan, Xingsen, Renhu, & Shouzhen, 2019)	Identificación de factores para mejorar la tasa de productos calificados	95.7%	Buen rendimiento	82.3%
	(Viloria, y otros, 2019)	Identificación de interacción de alumnos en cursos virtuales	92.9%	Buen rendimiento	
	(Ozyirmidokuz, Kumru, & Mustafa, 2015)	Análisis de volúmenes de datos de Marketing para toma de decisiones	92.6%	Buen rendimiento	
	(Schuh, Prote, & Hünnekes, 2020)	Análisis de datos para toma de decisiones en producción	90.6%	Buen rendimiento	
	(Pérez-Gutiérrez, 2020)	Identificación de deserción estudiantil	83%	Rendimiento aceptable	
	(Charris, y otros, 2018)	Análisis de datos biológicos	80%	Bajo rendimiento	

Árbol de Decisión Decision Tree (DT)	(Reuter, Brambring, Weirich, & Kleines, 2016)	Análisis de volúmenes de datos del proceso de producción para mejorar calidad	66.5%	Bajo rendimiento	
	(Parama, 2018)	Técnicas de minería de datos aplicadas a business intelligence	57.6%	Bajo rendimiento	
Naive Bayes (NB)	(Suharjito, 2015)	Análisis de datos en redes sociales para descubrir tendencias	90.3%	Buen rendimiento	85.1%
	(Mosquera, Parra-Osorio, & Castrillón, 2016)	Predicción de riesgo psicosocial	89%	Rendimiento aceptable	
	(Ghazzawi & Alharbi, 2019)	Procesamiento de datos de transporte público	86.5%	Rendimiento aceptable	
	(Parama, 2018)	Técnicas de minería de datos aplicadas a business intelligence	74.6%	Bajo rendimiento	
K Nearest Neighbor (KNN)	(Viloria, y otros, 2019)	Identificación de interacción de alumnos en cursos virtuales	98%	Buen rendimiento	82%
	(Ghazzawi & Alharbi, 2019)	Procesamiento de datos de transporte público	81%	Rendimiento aceptable	

	(Reuter, Brambring, Weirich, & Kleines, 2016)	Análisis de volúmenes de datos del proceso de producción para mejorar calidad	67%	Bajo rendimiento	
J-RIP	(Harley & Liu, 2017)	Análisis de datos para procesos de producción en manufactura	95.4%	Buen rendimiento	
	(Viloria, y otros, 2019)	Identificación de interacción de alumnos en cursos virtuales	94.4%	Buen rendimiento	94.9%
J48	(Mosquera, Parra-Osorio, & Castrillón, 2016)	Predicción de riesgo psicosocial	91%	Buen rendimiento	91%
Redes Neuronales (RN)	(Adnan, y otros, 2020)	Análisis de datos para toma de decisiones en producción	92.3%	Buen rendimiento	92.3%
Random Forest (RF)	(Ghazzawi & Alharbi, 2019)	Procesamiento de datos de transporte público	85.6%	Rendimiento aceptable	85.6%
Support Vector Machine (SVM)	(Parama, 2018)	Técnicas de minería de datos aplicadas a business intelligence	78.9%	Bajo rendimiento	78.9%

Fuente: investigaciones relevantes obtenidas de bases de datos científicas IEEE Xplore y ScienceDirect (2015 -2020)

Nota: Las técnicas fueron organizadas en la matriz de acuerdo al número de investigaciones realizadas en las cuales se utilizó dicha técnica.

Tal como se mencionó, en la revisión se consideraron 21 investigaciones, cuyo rendimiento se detalló en la tabla 4. Del total de investigaciones, ocho aplicaron la técnica de minería de datos Árbol de decisión (DT) y cuatro utilizaron Naive Bayes (NB). En la siguiente tabla, se presenta el total de investigaciones que aplicaron la técnica de minería de datos.

Tabla 6

Total de investigación por técnica de minería de datos

Técnica de minería de datos	Número de investigaciones aplicadas
Árbol de Decisión (DT)	8
Naive Bayes (NB)	4
K Nearest Neighbor (KNN)	3
J-RIP	2
J48	1
Redes Neuronales (RN)	1
Random Forest (RF)	1
Support Vector Machine (SVM)	1

Fuente: investigaciones relevantes obtenidas de bases de datos científicas IEEE Xplore y ScienceDirect (2015 -2020)

La finalidad de contabilizar el número de investigaciones que aplicaron una determinada técnica de minería de datos es obtener un ranking según su utilización. De acuerdo con ello, en la siguiente figura se muestran gráficamente las tres técnicas de minería de datos más utilizadas: Árbol de decisión o Decision Tree (DT), Naive Bayes (NB) y K Nearest Neighbor (KNN).

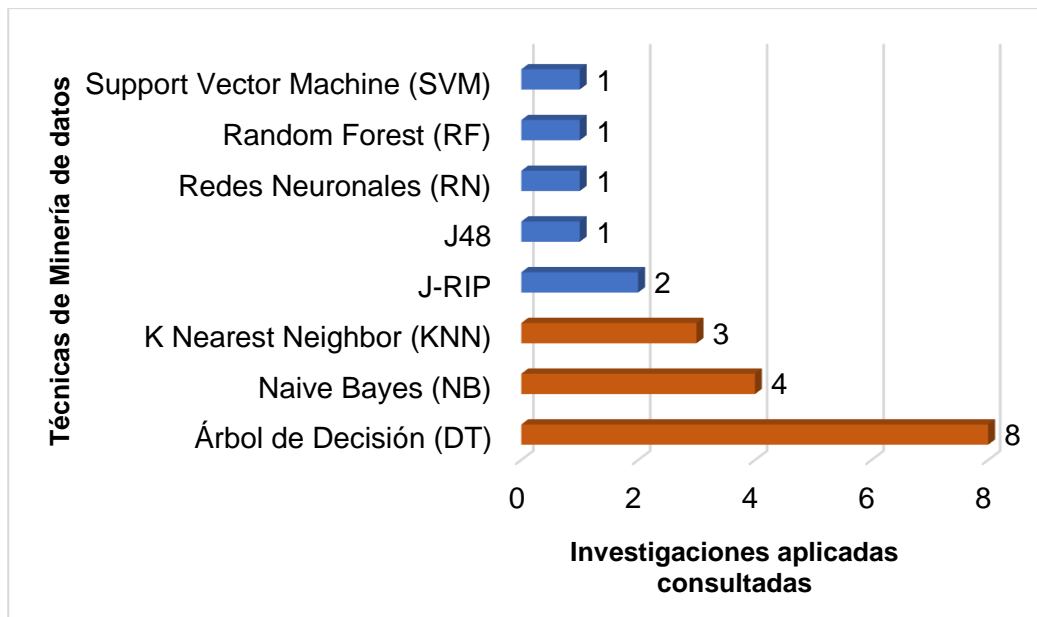


Figura 27. Top técnicas de minería de datos más utilizadas. Fuente: Elaboración propia

Finalmente, con base a los rendimientos obtenidos y el total de investigaciones en las que se utilizaron las técnicas de minería de datos, se determinó utilizar las técnicas Árbol de Decisión (DT) y Naive Bayes (NB). Como primer criterio se consideró que el rendimiento promedio de rendimiento sea superior a 80% de precisión. Como segundo criterio se evaluó que las técnicas de minería de datos hayan sido utilizadas en contextos similares; de esta forma, se identificó que las técnicas Árbol de decisión y Naive Bayes fueron utilizadas en 12 de las investigaciones, lo cual equivale al 55% de las investigaciones, tal como se puede apreciar en la siguiente figura:

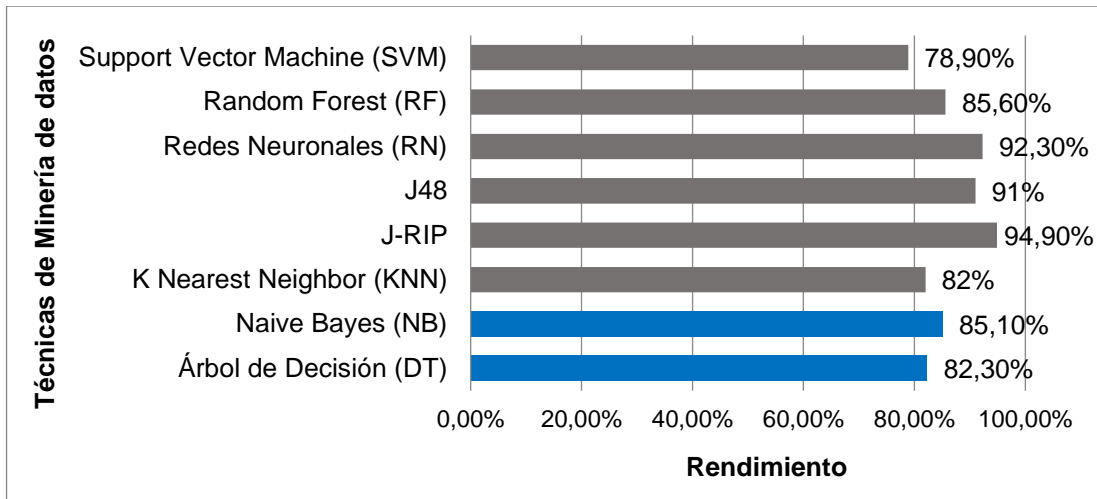


Figura 28. Técnicas de minería de datos que fueron utilizadas en el modelo.
Fuente: Elaboración propia.

Determinación de métricas para evaluación de técnicas de minería de datos

Para la evaluación de las técnicas de minería de datos se consideraron las mediciones estadísticas de error MAE, MSE y MAPE. Estas técnicas poseen una validación estadística y fueron utilizadas en investigaciones similares dada su practicidad y eficiencia para la medición.

El MAE es la diferencia absoluta ente el valor real y el valor obtenido en la predicción del modelo

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

En la ecuación anterior, y_i es la predicción del modelo y x_i es el valor verdadero. Otro factor de medición es MSE, el cual es una medida de dispersión que permite calcular el error cuadrático medio. Dicho de otra forma, mide la diferencia entre el estimador y lo que se estima. La fórmula utilizada para el cálculo es la siguiente:

$$MSE = \frac{\sum Error\ de\ pronóstico^2}{n}$$

Asimismo, se consideró el MAPE, el cual permite calcular el error porcentual medio. Consiste en la diferencia entre el valor real y valor esperado o pronóstico, pero expresado en términos porcentuales. La fórmula para el cálculo es la siguiente:

$$MAPE = \frac{\sum_{i=1}^n |Real_i - Pronóstico_i|}{\frac{Real_i}{n}}$$

Por otro lado, en relación a los indicadores de la solución business intelligence utilizados para la evaluación del rendimiento de las técnicas de minería de datos, se consideró el procesamiento de los datos. A manera de ejemplo, se consignó como un KPI la Cifra de Aprobados, la cual es calculada de la siguiente forma:

$$CA = \sum Aprobados \text{ por carrera}$$

Para el indicador de este tipo, se propone el cálculo del porcentaje de procesamiento de éxito de la cifra de ventas (% Proc. éxito CA), el cual es expresado de la siguiente forma:

$$\% \text{ Proc. éxito CA} = \frac{\text{Total datos proces.}}{\text{Total de datos}} * 100$$

Otro indicador de medición sería el procesamiento de los datos del margen comercial, cuya fórmula es la siguiente:

$$CA = \sum Aprobados \text{ por modalidad}$$

Por último, se consideró el Tiempo de Procesamiento (TP); la fórmula para el cálculo está basada en la sumatoria del tiempo promedio observado dividido entre el número de observaciones:

$$TP = \frac{\sum \text{Tiempo observado}}{\text{Número de observaciones}}$$

Con el objetivo de determinar la precisión de las técnicas de minería de datos se propuso utilizar la matriz de confusión, lo cual permitirá calcular el porcentaje de predicciones positivas correctas. La estructura de la matriz utilizada es la siguiente:

Matriz de confusión		Estimado por la técnica		
		Negativo (N)	Positivo (P)	
Real	Negativo	TN	FP	
	Positivo	FN	TP	“Precisión” % de predicciones positivas correctas

La precisión es el cociente obtenido entre el valor de verdaderos positivo (TP) dividido con la suma de los valores verdaderos positivos (TP) y falsos positivos (FP). La fórmula es la siguiente:

$$Precisión = \frac{TP}{TP + FP}$$

Diseño de método de aplicación

El método propuesto para la aplicación de las técnicas de minería de datos en una solución business intelligence consta de seis pasos: a) análisis y comprensión de las fuentes de datos del negocio, implementación de la base de datos en SQL Server, proceso ETL, aplicación de los algoritmos de minería de datos Árbol de decisión (DT) y Naive Bayes (NB), procesamiento de datos y obtención de métricas MAE, MSE y MAPE. En la siguiente figura se muestra la secuencia de pasos que se siguieron en la investigación:

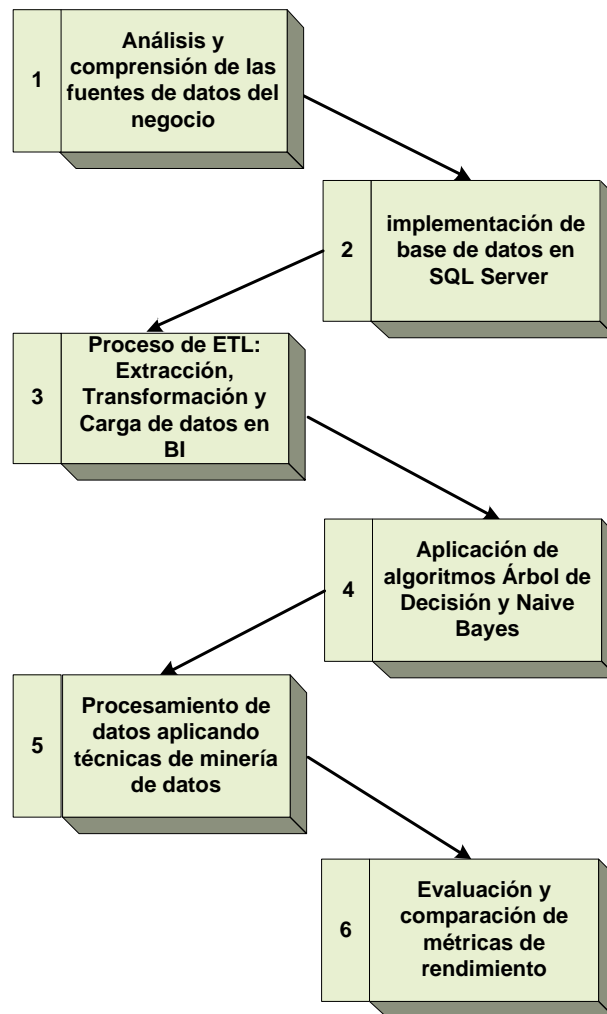


Figura 29. Método de aplicación de las técnicas de minería de datos en una solución business intelligence. Fuente: Elaboración propia

En la presente investigación, se trabajó con datos de una universidad peruana. Por acuerdo de privacidad y confidencialidad de datos, se reservada la razón social de dicha institución y se hará referencia en la investigación como universidad peruana. La base de datos de la universidad recibe información de cuatro áreas principales: comercial, operaciones, servicios universitarios y académica. Los datos son registrados mediante el uso de sistemas de información transaccionales. En la siguiente tabla, se presentan las áreas y los principales procesos que se realizan en cada una:

Tabla 7

Áreas y procesos principales de la universidad del caso de estudio

N°	Área	Procesos principales
01	Operaciones	<ul style="list-style-type: none">Logística y mantenimientoGestión de comprasSeguridad
02	Servicios universitarios	<ul style="list-style-type: none">Planificación y matrículaPagos
03	Comercial	<ul style="list-style-type: none">Gestión de postulantesAdmisión
04	Académicos	<ul style="list-style-type: none">Diseño curricular y programaciónGestión de docentesGestión de alumnos

Fuente: información obtenida de la documentación del área de Gestión y Procesos de la universidad del caso de estudio en el año 2020.

De acuerdo con el cuadro anterior, la información de la Universidad del caso de estudio se encuentra en distintas fuentes. Sin embargo, para fines de determinar la predicción de rendimiento académico, solo se utilizará la información relevante correspondiente a las áreas de Servicios Universitarios y Académica. La información correspondiente a compras, finanzas, logística, entre otras, no es relevante para predecir el rendimiento académico.

Implementación de base de datos académica y proceso business intelligence

El proceso inició con la obtención de un reporte de 20,173 registros correspondientes a datos del área académica, los cuales fueron analizados para una reconstrucción de la base de datos mediante un proceso de ingeniería inversa en SQL Server. El proceso de creación del modelo de base de datos se obtuvo a partir de los datos existentes en una hoja de

cálculo, los cuales fueron revisados para presentar de forma gráfica las relaciones entre sus tablas. De esta forma, se seleccionaron ocho tablas: ALUMNO, SEDE, MODALIDAD, CURSO, DOCENTE, NOTAS, CARRERA y FACULTAD. La finalidad de dicho proceso fue presentar en una base de datos normalizada las tablas y columnas deducidas del reporte general.

La creación de la base de datos y las tablas se realizó mediante el uso de comando SQL. En la siguiente imagen, se muestra el código utilizado para

```
create database UNIVERSIDAD01
use UNIVERSIDAD01

create table Alumno(
id_alumno integer not null identity ( , ) primary key,
cod_alumno varchar ( ),
apellidos varchar( ),
nombres varchar ( ),
genero varchar ( ),
fecha_nacim date,
ciclo int,
retirado varchar ( ),
id_carrera int,
id_sede int
constraint FK_Carrera_Id_Carrera FOREIGN KEY (id_carrera) references
Carrera(id_carrera),
constraint FK_Sede_Id_Sede FOREIGN KEY (id_sede) references
Sede(id_sede)
)

create table Curso (
id_curso integer not null identity ( , ) primary key,
cod_curso varchar( ),
nombre_curso varchar( ),
seccion varchar ( ),
periodo varchar( ),
id_docente int,
id_facultad int,
id_modalidad int
constraint FK_Docente_Id_Docente FOREIGN KEY (id_docente) references
Docente(id_docente),
constraint FK_Facultad_Id_Facultad FOREIGN KEY (id_facultad) references
Facultad(id_facultad),
constraint FK_Modalidad_Id_Modalidad FOREIGN KEY (id_modalidad)
references Modalidad(id_modalidad)
)
```

la creación de 2 tablas: Alumno y Curso, cuyo procedimiento fue replicado para la generación de las tablas descritas anteriormente.

Figura 30. Script para creación de base de datos académica en SQL Server 2012. Fuente: Elaboración propia.

Posterior a la creación de la base de datos, se elaboró el diagrama de entidad relación en SQL Server, para lo cual se seleccionaron las ocho tablas generadas. El diagrama considera como claves primarias los ID de cada tabla. La base de datos fue poblada teniendo como origen de datos el archivo en Excel. En la siguiente imagen se presenta el diagrama obtenido de la base de datos normalizada:

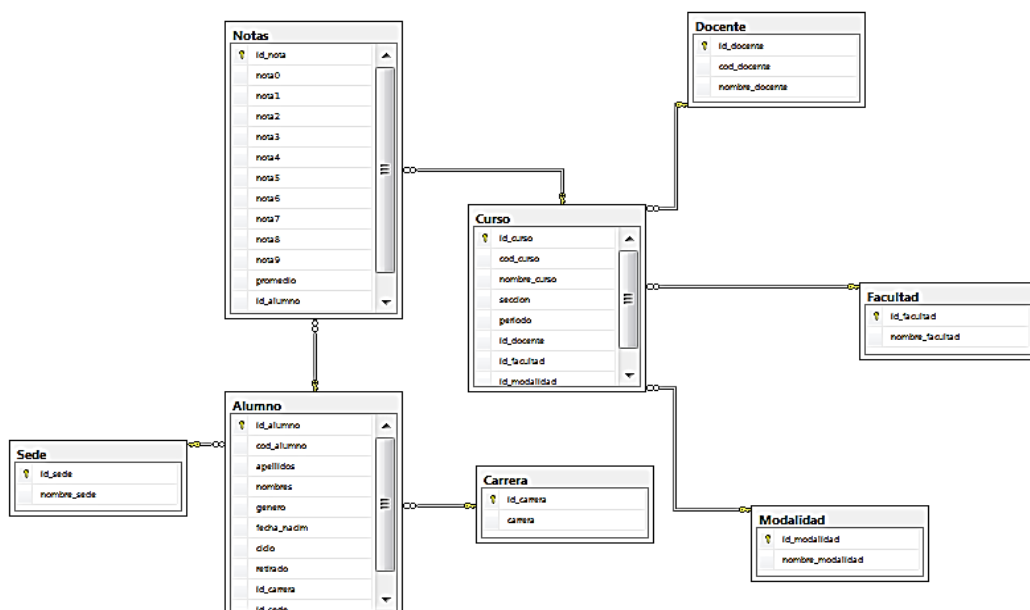


Figura 31. Diagrama de Entidad - Relación de base de datos académica de una universidad peruana. Fuente: Elaboración propia.

Después de la creación del diagrama E-R, se determinaron los indicadores del negocio KPI's, con el objetivo de que esta información se refleje posteriormente en la construcción de la data mart, el cual constituye un almacén de datos de un departamento o área específica. Para poder determinar dichos indicadores, se definió el problema del caso de estudio y el objetivo del proyecto.

a) Descripción del problema

El área académica de dicha universidad peruana cuenta con información sobre el rendimiento académico de los estudiantes, esta información es obtenida de diversas fuentes de datos y no es analizada de forma integral, lo que dificulta realizar análisis de predicción respecto al rendimiento académico.

b) Objetivo

Comparar la precisión respecto a la predicción del rendimiento académico de los estudiantes, mediante técnicas de minería de datos a variables obtenidas de una base de datos dimensional generada bajo la metodología business intelligence.

c) KPI

El indicador clave del área académica del caso de estudio es el rendimiento académico de los estudiantes. Esta información debe ser visible para las áreas responsables de la toma de decisión de forma que se pueda ser observada según la modalidad de estudio y carreras profesionales o cursos, los cuales se encuentran bajo la gestión de direcciones académicas independientes.

Después de haber importado los datos a la base de datos académica, describir y determinar los objetivos e indicadores del caso propuesto, se procedió a crear la base de datos dimensional mediante la inserción de consultas en SQL. En la siguiente imagen se muestra el código utilizado:

```

create database UNIVERSIDAD01_DM
use UNIVERSIDAD01_DM

create table DIM_ALUMNO(
ID_ALUMNO int not null identity ( , ) primary key,
COD_ALUMNO varchar( ),
APELLIDOS varchar ( ),
NOMBRES varchar ( ),
FECH_NACIM date,
GENERO varchar ( ),
CARRERA varchar ( ),
SEDE varchar ( )
)

create table DIM_TIEMPO(
ID_TIEMPO int not null identity ( , ) primary key,
FECHA varchar ( )
)

create table FAC_NOTAS(
ID_NOTAS int not null identity ( , ) primary key,
ID_ALUMNO int,
ID_CURSO int,
ID_TIEMPO int,
PROMEDIO varchar ( )
)

constraint FK_DIM_ALUMNO_ID_ALUMNO FOREIGN KEY (ID_ALUMNO)
constraint FK_DIM_CURSO_ID_CURSO FOREIGN KEY (ID_CURSO)
constraint FK_DIM_TIEMPO_ID_TIEMPO FOREIGN KEY (ID_TIEMPO)
references DIM_TIEMPO (ID_TIEMPO)
)

```

Figura 32. Script para creación de base de datos dimensional en SQL Server 2012. Fuente: Elaboración propia.

En la siguiente tabla se presenta una descripción de las dimensiones consideradas para la base de datos dimensional:

Tabla 8

Descripción de las dimensiones consideradas para la construcción de la base de datos dimensional

Dimensión	Descripción
DIM_ALUMNO	<ul style="list-style-type: none">Está compuesta por los datos de los alumnos: apellidos, nombres, fecha de nacimiento, carreras y están identificados por ID_ALUMNO
DIM_TIEMPO	<ul style="list-style-type: none">Considera el ciclo académico de estudio; está identificado por ID_TIEMPO
DIM_CURSO	<ul style="list-style-type: none">Registra la información: nombre de curso, código, sección, facultad y docente y tiene como identificador principal ID_CURSO.
FAC_NOTAS	<ul style="list-style-type: none">Esta es la tabla de hechos, la cual considera la relación mediante las siguientes claves primarias: ID_ALUMNO, ID_TIEMPO, ID_CURSO, ID_NOTAS. Asimismo, incluye la columna PROMEDIO la cual contiene solo dos etiquetas “Aprobado” y “No aprobado”.

Fuente: base de datos implementada en SQL Server con datos del área académica de una universidad peruana en el año 2020. Fuente: Elaboración propia.

Después de definir las dimensiones, se generó el diagrama de entidad relación, el cual se muestra en la figura 25. Dicho diagrama corresponde a la data mart del área académica y está conformado por cuatro dimensiones y una tabla de hechos, vinculadas mediante claves primarias: ID_TIEMPO, ID_NOTAS, ID_CURSO e ID_ALUMNO.

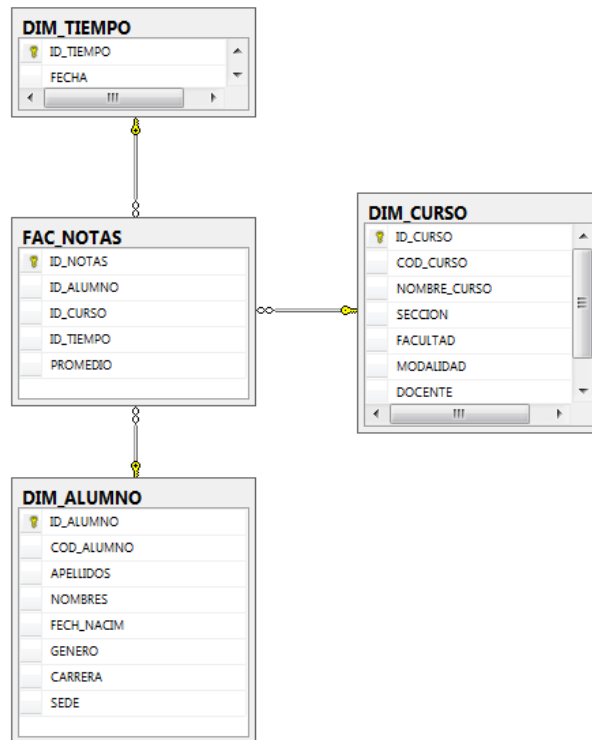


Figura 33. Diagrama de Entidad - Relación de base de datos dimensional.
Fuente: Elaboración propia.

Una vez obtenida la base de datos dimensional y el diagrama entidad – relación, se generó un proyecto de inteligencia de negocios en Microsoft Visual Studio 2015 para realizar el proceso ETL. Primero se diseñó el flujo de control y flujo de datos con SQL Server con el objetivo de poblar la base de datos dimensional. En la figura 26 se muestra la estructura del flujo de control que contempla a las cuatro dimensiones:

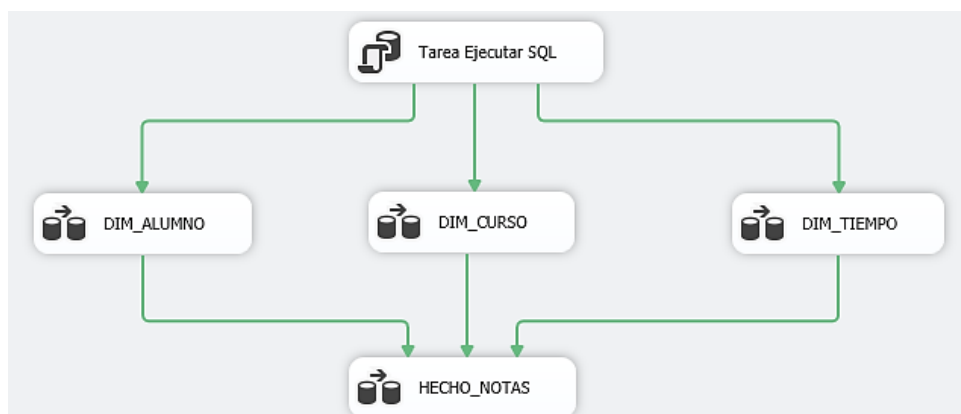


Figura 34. Diseño de flujo de control de la base de datos dimensional.
Fuente. Elaboración propia.

La finalidad de generar un flujo de control de la base de datos dimensional del área académica (UTP01-DM) es consignar las tareas que aceptan flujo y preparación de datos para inteligencia de negocios; para ello se utilizó el complemento de SQL Server Integration Services. Asimismo, se estableció el flujo de dato de las dimensiones: ALUMNO, CURSO, TIEMPO y NOTAS, los cuales se muestran en la siguiente imagen:

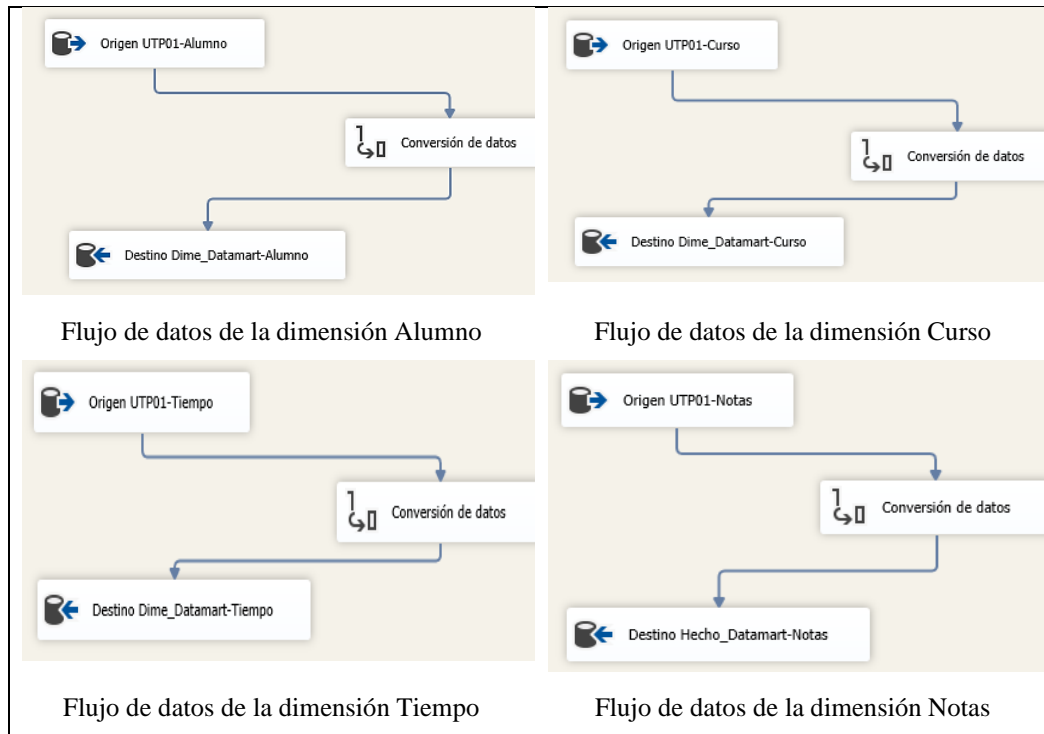


Figura 35. Diseño de flujo de datos de la base de datos dimensional. Fuente: Elaboración propia.

Finalmente, con el objetivo de visualizar los datos importados a la base de datos dimensional, se insertó una consulta SQL para la obtención de reporte general, el cual será el insumo posterior para los datos de entrada. La consulta se muestra en la siguiente imagen:

```

---consulta para reporte final---

SELECT alumno.COD_ALUMNO, alumno.APELLIDOS, alumno.NOMBRES,
alumno.FECH_NACIM, alumno.GENERO, CARRERA, alumno.SEDE,
curso.COD_CURSO, curso.NOMBRE_CURSO, curso.DOCENTE,
curso.SECCION, curso.MODALIDAD, curso.FACULTAD, tiempo.FECHA,
notas.PROMEDIO

FROM FAC_NOTAS notas inner join DIM_ALUMNO alumno on
(notas.ID_ALUMNO=alumno.ID_ALUMNO)
inner join DIM_CURSO curso on (notas.ID_CURSO=curso.ID_CURSO) inner
join DIM_TIEMPO tiempo on (notas.ID_TIEMPO=tiempo.ID_TIEMPO)

```

Figura 36. Script para la consulta de reporte general a la base de datos dimensional en SQL Server 2012. Fuente: Elaboración propia

Como se puede apreciar, dentro del marco del proceso de business intelligence es de suma importancia garantizar el correcto flujo de datos hacia la base de datos dimensional UTP01-DM, la cual debe ser poblada. Posteriormente, se construyó el cubo de business intelligence.

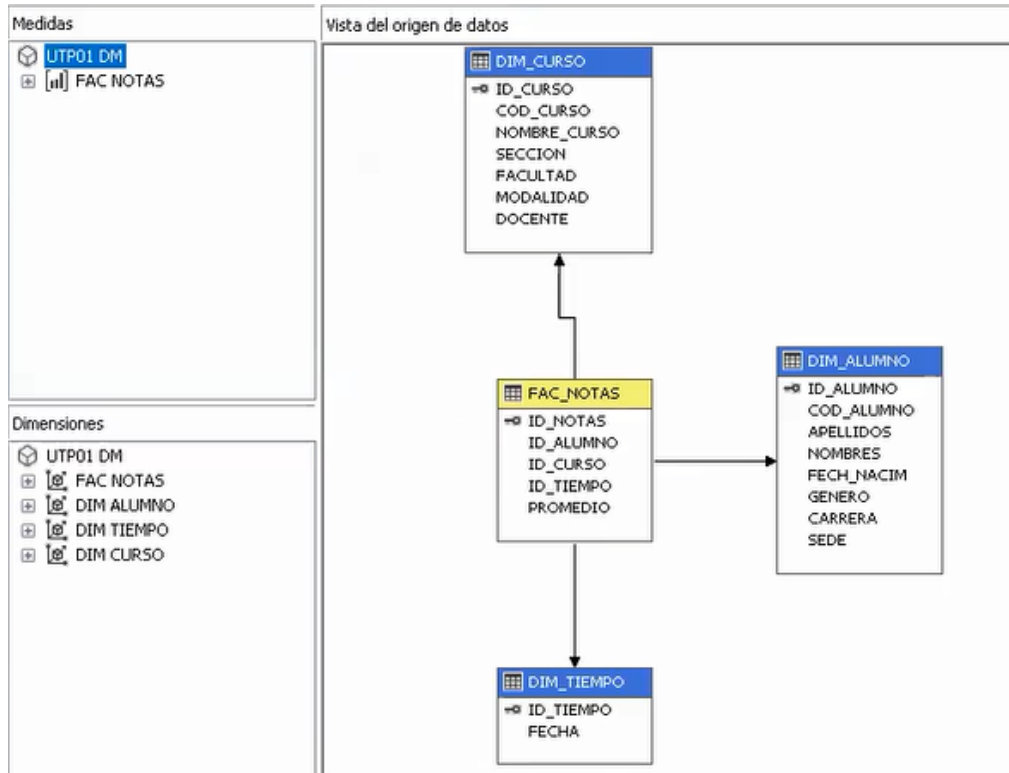


Figura 37. Generación del cubo business intelligence. Fuente: Elaboración propia.

Como producto final del proceso de business intelligence, se obtuvieron un conjunto de gráficos estadísticos del reporte obtenido del cubo BI. A manera de ejemplo, se presentan algunos de los gráficos:

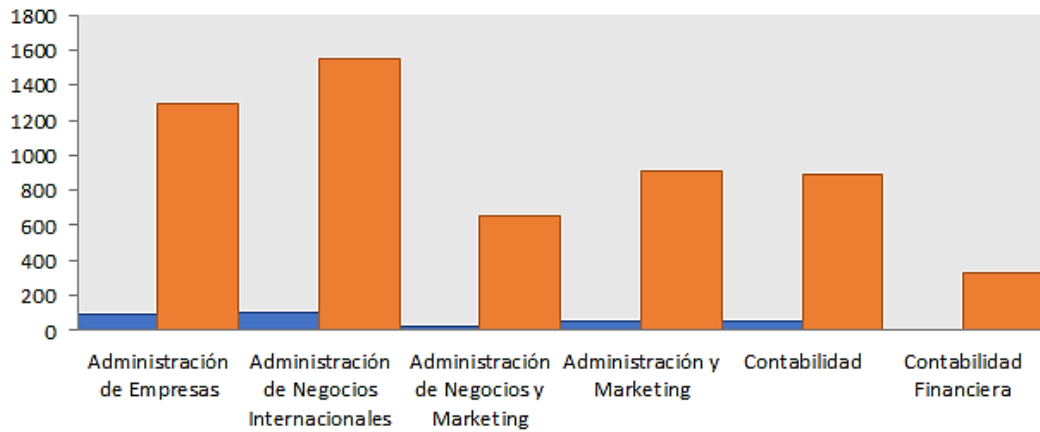


Figura 38. Modelo de reporte BI considerando KPI desaprobados por escuela. Fuente: Elaboración propia

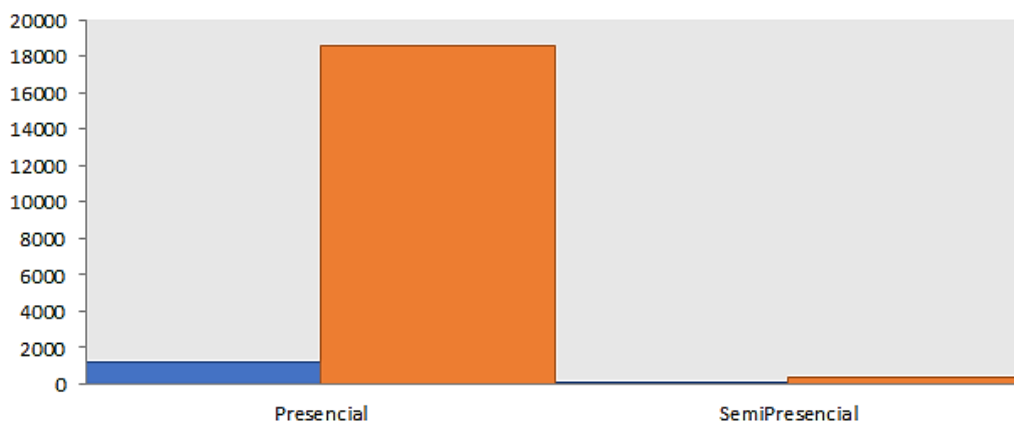


Figura 39. Modelo de reporte BI considerando KPI desaprobados por modalidad de estudio. Fuente: Elaboración propia.

El proceso business intelligence implementado permitió obtener un data mart del departamento académico con información relacionada al rendimiento académico de los estudiantes. Se generó un Cubo de business intelligence con las dimensiones necesarias para la generación de reportes, las mismas que fueron establecidas en la base de datos dimensional.

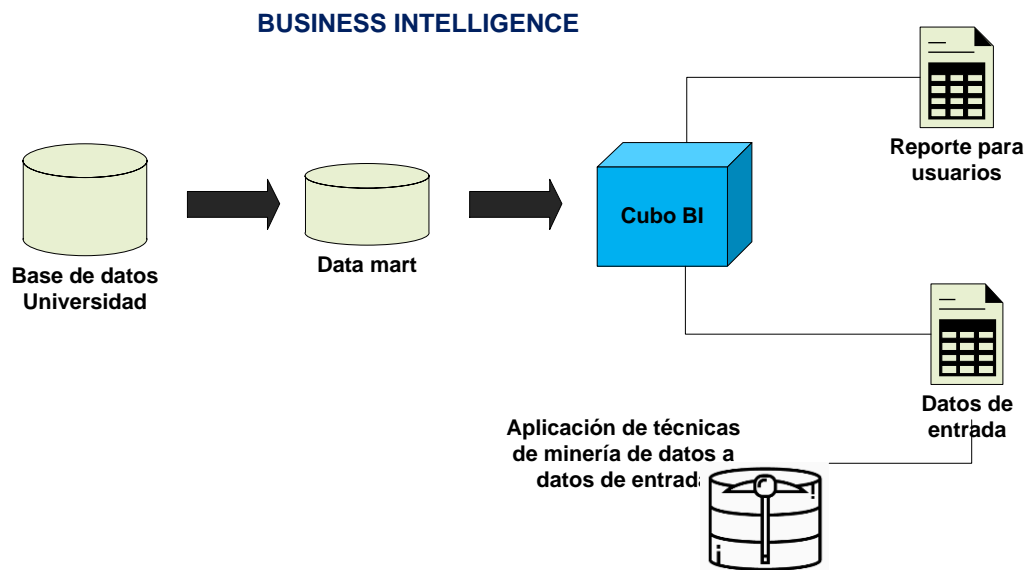


Figura 40. Proceso de obtención de datos de entrada para aplicación de técnicas de minería de datos a partir de un proceso business intelligence. Fuente: Elaboración propia.

La data mart creado contenía información del área académica, la cual fue extraída de la base de datos de la universidad. La base de datos dimensional obtenida del proceso business intelligence permitió generar los cubos para la obtención de los reportes estadísticos y los datos de entrada para la aplicación de las técnicas de minería de datos.

A partir de los datos de entrada obtenidos de la solución business intelligence, se implementó el método de aplicación propuesto para fines de comparación de rendimiento de las técnicas DT y NB para la predicción de rendimiento académico de estudiantes.

Técnicas de minería de datos Árbol de Decisión y Naive Bayes

Después de realizar el proceso business intelligence para la obtención de los datos de entrada, se procedió a codificar las técnicas de minería de datos. Para ello, se utilizó la herramienta RStudio, el cual es un entorno de desarrollo basado en el lenguaje de programación R. La herramienta fue seleccionada debido a que cuenta con la Licencia Pública General - GNU y está disponible para sistemas operativos Windows, el cual fue utilizado

durante todo el proceso de investigación. Asimismo, dicha herramienta ha sido utilizada en diversos trabajos de investigación sobre minería de datos. La versión RStudio utilizada fue 1.1.456, la cual permite instalar las librerías para Árbol de Decisión y Naive Bayes.

Para el proceso de comparación de los algoritmos de minería de datos. Se propuso un método de cuatro etapas: establecer un escenario de aplicación, seleccionar métricas estandarizadas, ejecución de las técnicas de minería de datos, registro de resultados en ficha de observación, procesamiento de resultados y generación de gráficos.

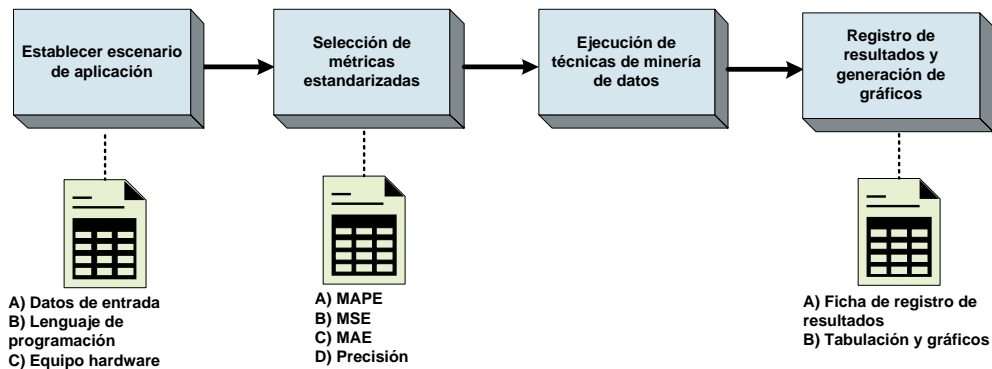


Figura 41. Método de comparación de técnicas de minería de datos propuesto. Fuente: Elaboración propia.

El proceso de implementación de la técnica de minería de datos Árbol de decisiones fue codificado en 34 líneas, incluyendo los comentarios y fue estructurado en cinco etapas. Primero se procedió con la instalación de las librerías de Árbol de decisiones; para el presente caso se utilizó la versión C5.0, el cual incorpora dentro de su algoritmo de clasificación la generación de árboles de decisión simples y modelos basados en reglas. Segundo, se realizó la preparación de la data set de entrenamiento; para ello, se envió la ubicación de la base de datos en formato .csv y se inició el generador de número aleatorios, el cual para este modelo es máximo 100. Tercero, se creó la data set de entrenamiento 70% y el restante 30% utilizado para pruebas; esto se realizó con el objetivo de tener un proceso de entrenamiento más fiable y teniendo como referencia trabajos similares con volúmenes de datos superior a los 10,000 registros. Cuarto, se aplicó la

técnica de minería de datos Árbol de decisiones tomando como eje central del modelo la variable "PROMEDIO". Por último, se obtuvo la matriz con las métricas, mediante la matriz de confusión. En la siguiente figura se muestra la estructura de codificación de Árbol de decisiones en el lenguaje R.

```

1  ##### PASO 1: Instalación de las librerías de Árbol de Decisiones
2  install.packages("c50")
3  #librería de Árbol de Decisiones -> C5.0
4  library(c50)
5  library(rpart)
6  library(rpart.plot)
7  ##### PASO 2: Preparación de data set
8  #Envía ubicación predeterminada donde se iniciará la lectura de datos
9  getwd()
10 #Lee la ruta de la base de datos en CSV
11 estudiantes <- read.csv("EntradaTecnicas.csv")
12 #inicializar el generador de números aleatorios 100
13 set.seed(100)
14 ##### PASO 3: Crea set de entrenamiento y de test
15 # Entrenamos los datos los datos el 70% y el restante 30% para usarlos
16 tamano.total <- nrow(estudiantes)
17 tamano.entreno <- round(tamano.total*0.70)
18 estudiantes.indices <- sample(1:tamano.total , size=tamano.entreno)
19 estudiantes.entreno <- estudiantes[estudiantes.indices,]
20 estudiantes.test <-estudiantes[-estudiantes.indices,]
21 str(estudiantes)
22 ##### PASO 4: Aplicación de Árbol de decisiones
23 #Aplicando Árbol de decisiones a la base de datos
24 modelo1 <- C5.0(PROMEDIO ~ ., data = estudiantes.entreno, trial = 50,
25                 control = C5.0control(noGlobalPruning = TRUE,CF = 0.70))
26 #Aplicando Predicción o uso de los datos restantes
27 prediccion <- predict(modelo1,newdata=estudiantes.test)
28 ##### PASO 5: Metricas de comparación
29 #Muestra los datos en una matriz de confusión
30 library(caret)
31 tabla1 <- table(prediccion,estudiantes.test$PROMEDIO)
32 confusionMatrix(tabla1)
33 #Muestra la cantidad de en una matriz
34 summary(modelo1)

```

Figura 42. Código de la implementación de técnica de minería de datos Árbol de Decisiones C5.0 en lenguaje de programación R utilizando la herramienta RStudio. Fuente: Elaboración propia.

Los resultados de la matriz de confusión en este modelo mostraron 93.68% de precisión. Asimismo, obtuvo 93% de exactitud, lo cual refleja el nivel de concordancia entre los valores verdaderos del rendimiento académico y los valores obtenidos. Por otro lado, la sensibilidad de 99% hace referencia al número de elementos positivos identificados correctamente del total de valores positivos verdaderos.

Asimismo, de forma complementaria para obtener las métricas de estimación de error consignadas para el análisis y comparación de las técnicas, se utilizó la herramienta WEKA:

```
=== Summary ===
```

Correctly Classified Instances	12650	89.5829 %
Incorrectly Classified Instances	1471	10.4171 %
Kappa statistic	0.1053	
Mean absolute error	0.1044	
Root mean squared error	0.3227	
Relative absolute error	89.8897 %	
Root relative squared error	133.7706 %	
Total Number of Instances	14121	

Figura 43. Resultados de métricas de estimación de error de la técnica Árbol de decisiones en WEKA

Con respecto a la técnica de minería de datos Naive Bayes, el proceso de implementación fue codificado en 33 líneas, incluyendo los comentarios y fue estructurado en cinco etapas. Primero se procedió con la instalación de las librerías de Naive Bayes. Segundo, se realizó la preparación de la data set de entrenamiento; para ello, se envió la ubicación de la base de datos en formato .csv y se inició el generador de número aleatorios. Tercero, se creó la data set de entrenamiento 70% y el restante 30% utilizado para predicción; esto se realizó con el objetivo de tener un proceso de entrenamiento más fiable y teniendo como referencia trabajos similares con volúmenes de datos superior a los 10,000 registros. Cuarto, se aplicó la técnica de minería de datos Naive Bayes tomando como eje central del modelo la variable "PROMEDIO". Por último, se obtuvo la matriz con las métricas, mediante la matriz de confusión. En la siguiente figura se muestra la codificación en el lenguaje R.

```

1  ###PASO 1: Instalar los Paquetes de Naive Bayes
2  install.packages("naivebayes")
3  install.packages("e1071")
4  #libreria de naive bayes -> e1071
5  library(e1071)
6  library(caret)
7  #### PASO 2: Preparamos set de datos
8  #Envía ruta Predetermianda donde se empezará la lectura de datos
9  getwd()
10 #Lee la ruta de nuestra Base de Datos en CSV
11 estudiantes <- read.csv("EntradaTécnicas.csv")
12 #Muestra Base de Datos
13 str(estudiantes)
14 #inicializar el generador de números aleatorios 2000
15 set.seed(2000)
16 #### PASO 3: Crea set de entrenamiento
17 # Entrena los datos los datos: el 70% y el restante 30% para predicción
18 entrenamiento <- createDataPartition(estudiantes$PROMEDIO, p = 0.70, list = F)
19 #### PASO 4: Aplicando Naive Bayes
20 #Aplica Naive Bayes a la base de datos
21 modelo <- naiveBayes(PROMEDIO ~ ., data = estudiantes[entrenamiento,])
22 #Apriori emplea una búsqueda por niveles para conjuntos de elementos frecuentes
23 modelo$apriori
24 #Aplicando Predicción o uso de los datos restantes
25 prediccion <- predict(modelo, newdata = estudiantes)
26 #Capturamos los datos en un conjunto
27 tabla1 <- table(prediccion,estudiantes$PROMEDIO)
28 #### PASO 5: Metricas de Comparación
29 #Mostramos los datos en una matriz de confusión (En ella encontramos las Métricas)
30 confusionMatrix(tabla1)
31 #Mostramos la cantidad en una matriz
32 modelo
33 summary(modelo)

```

Figura 44. Código de la implementación de técnica de minería de datos Naive Bayes en lenguaje de programación R utilizando la herramienta RStudio. Fuente: Elaboración propia.

Los resultados de la matriz de confusión en este modelo mostraron, 93.67% de precisión y 93% de exactitud, lo cual refleja el nivel de concordancia entre los valores verdaderos del rendimiento académico y los valores obtenidos. Por otro lado, la sensibilidad hace referencia al número de elementos positivos identificados correctamente del total de valores positivos verdaderos.

Asimismo, de forma complementaria para obtener las métricas de estimación de error consignadas para el análisis y comparación de las técnicas, se utilizó la herramienta WEKA:

=== Summary ===		
Correctly Classified Instances	18918	93.7788 %
Incorrectly Classified Instances	1255	6.2212 %
Kappa statistic	0.0332	
Mean absolute error	0.108	
Root mean squared error	0.2307	
Relative absolute error	92.9636 %	
Root relative squared error	95.7143 %	
Total Number of Instances	20173	

Figura 45. Resultados de métricas de estimación de error de la técnica Naive Bayes en WEKA. Fuente: Elaboración propia.

IV. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones.

- a) Durante el proceso de selección de las técnicas de minería de datos, se pudo comparar los resultados obtenidos en diversas investigaciones. De esta forma, se determinó que las técnicas de mejor rendimiento promedio en cuanto a su precisión fueron Árbol de decisiones con 85.1% y Naive Bayes con 82.3%. Asimismo, dichas técnicas habían sido utilizadas en más del 50% de las investigaciones consultadas.

- b) Respecto a la determinación de los indicadores de evaluación de las técnicas de minería de datos aplicadas al caso de estudio de predicción de rendimiento académico, se consideró el error absoluto medio (MAE), el error cuadrático medio (MSE), el error porcentual absoluto medio (MEPE), el tiempo de procesamiento observado (TP) y precisión. La selección de dichas técnicas permitió comparar los resultados con investigaciones similares aplicadas en otros contextos y comparar los resultados obtenidos entre ambas técnicas para determinar posteriormente la de mejor rendimiento en cuanto a error y precisión.

- c) El método de aplicación diseñado para el presente estudio estuvo conformado por cinco etapas: análisis y comprensión de las fuentes de datos, implementación de la base de datos en SQL Server, proceso ETL, implementación de los algoritmos de minería de datos a partir de los datos de entrada obtenidos del proceso business intelligence, procesamiento de datos y comparación de resultados. El método propuesto permitió obtener resultados de predicción de rendimiento académico y garantizar la calidad de los datos de entrada mediante el proceso ETL.

- d) Para la implementación de la base de datos normalizada se utilizó la herramienta SQL Server 2012, la cual permitió desarrollar el proceso business inteligente y generar la base de datos dimensional con los variables que serían utilizadas en el dataset para la aplicación de las técnicas de minería de datos. Teniendo en consideración el volumen de datos, el cual fue superior a los 20,000 registros, constituye una herramienta valiosa para la generación de datos de entrada de calidad.
- e) Para la implementación de las técnicas de minería de datos, se utilizó la herramienta RStudio, la cual trabaja con el lenguaje de programación R. La herramienta cuenta con licencia GNU y versión para Windows, cuyo sistema operativo fue utilizado durante todo el proceso de investigación. El uso de dicha herramienta permitió procesar los datos de entrada obtenidos del proceso business intelligence y obtener las métricas de error y precisión que serían utilizadas en el análisis de resultados.
- f) Para garantizar una adecuada comparación de las técnicas, se propuso un método de comparación compuesto por cuatro etapas: determinación de un escenario de pruebas, selección de métricas estandarizadas, aplicación de técnicas de minería de datos, registro de los resultados y representación gráfica. Dicho proceso permitió comparar y discutir objetivamente los resultados obtenidos con los hallazgos en otras investigaciones.
- g) Los resultados evidenciaron que el modelo propuesto, el cual utilizó datos de entrada obtenidos de un proceso business intelligence obtuvo un rendimiento en cuanto a su precisión superior al 90% en ambas técnicas de minería de datos. Árbol de decisiones obtuvo 93.69% y Naive Bayes 93.67%. Asimismo, en cuanto al análisis de error, Naive Bayes fue la que mejor resultados obtuvo, obteniendo un MAPE de 6.2%. Si bien la precisión fue mejor en comparación a otras investigaciones, dichas investigaciones procesaron volúmenes de datos mucho mayor.

4.2. Recomendaciones.

- a) Se recomienda la construcción de una matriz de resultados de investigaciones previas antes de decidir utilizar una técnica de minería de datos. La evaluación de su rendimiento ayudará a una mejor elección, pese a que hayan sido utilizadas en contextos diferentes.
- b) Se recomienda utilizar métricas de rendimiento aplicadas siempre a minería de datos como precisión y MAPE. Algunos de los aportes teóricos explorados en la investigación ponen especial énfasis en el consumo de memoria y CPU. No obstante, ello dependerá de variables ajenas a las técnicas de minería de datos, como las características de hardware y sistema operativo utilizado.
- c) Para fines de obtener datos de calidad para la aplicación de las técnicas de minería de datos, se recomienda seguir el proceso business intelligence hasta la conformación y llenado de la base de datos dimensional de cual se obtendrán los reportes con las variables que conformarán el dataset. Es importante considerar que la información obtenida debe ser la misma que será utilizada como insumo para la generación de reportes estadísticos con los KPI's del negocio.
- d) En el proceso de limpieza de datos, se recomienda utilizar algunos criterios que permitan aprovechar la gran mayoría de registros que alimentarán la base de datos dimensional. Por ejemplo, para los campos en blanco se podrían utilizar algunas medidas de tendencia central como la mediana y moda.
- e) Con la finalidad de incrementar la precisión de las técnicas de minería de datos para la predicción de rendimiento académico, se recomienda incluir algunas variables adicionales como modalidad de ingreso y número de cursos matriculados por alumno. Asimismo, se recomienda utilizar información por lo menos de los últimos 3 años, lo cual equivale a 6 ciclos académicos.

REFERENCIAS.

- Adnan, M., Saqib, S., Alyas, T., Ur Rehman, A., Saeed, Y., Zeb, A., & Zareei, M. (2020). Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. *IEEE Access*, 8, 116013-116023. doi:<https://doi.org/10.1109/ACCESS.2020.3003790>
- Anastasios, K., Panos, V., & Alkis, S. (2013). Scheduling strategies for efficient ETL execution. *Information Systems*, 38(6), 927-945. doi:<https://doi.org/10.1016/j.is.2012.12.001>
- Anwar, Y., & Addin, O. (2019). Using Data Mining Techniques to Guide Academic Programs Design and Assessment. *Procedia Computer Science*, 163, 472–481. doi:<https://doi.org/10.1016/j.procs.2019.12.130>
- Azoumana, K. (2013). Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos. *Pensamiento Americano*, 6(10), 41-51. doi:<https://doi.org/10.21803/pensam.v6i10.133>
- Bustamante, A., Galvis-Lista, E., & Gómez, L. (2012). Modeling techniques for extraction transformation and load processes: a critical review. *2012 6th Euro American Conference on Telematics and Information Systems (EATIS)*, 41-47. doi:<https://doi.org/10.1145/2261605.2261611>
- Cano, J. (2007). *Business Intelligence: Competir Con Información*. Fundación Cultural, 397.
- Charris, L., Henriquez, C., Hernandez, S., Jimeno, L., Guillen, O., & Moreno, S. (2018). Comparative analysis of algorithms of decision trees in the processing of biological data. *Revista I+D en TIC*, 9(1), 26-34. Obtenido de <http://revistas.unisimon.edu.co/index.php/identific/article/view/3158>
- Chen, G., Baoran, A., & Yang, L. (2016). A novel agent-based parallel ETL system for massive data. *2016 Chinese Control and Decision Conference (CCDC)*, 3942-3948. doi:<https://doi.org/10.1109/CCDC.2016.7531673>
- Denis-Cătălin , A., Ioana-Gilia, D., & Miruna, O. (2019). Organizational development through Business Intelligence and Data Mining. *Database Systems Journal*, 10, 82-99. Obtenido de http://dbjournal.ro/archive/30/30_9.pdf

- Gallego, J., Navarro, L., & Castillo, L. (2015). Aplicación de técnicas de minería de datos en atención primaria en salud (APS) para el análisis de riesgos en mujeres gestantes de la población manizaleña atendida por ASSBASALUD. *Biosalud*, 14(2), 71-78. doi:<https://doi.org/10.17151/biosa.2015.14.2.7>.
- Ghazzawi, A., & Alharbi, B. (2019). Analysis of Customer Complaints Data using Data Mining Techniques. *Procedia Computer Science*, 163, 62-69. doi:<https://doi.org/10.1016/j.procs.2019.12.087>
- Gutiérrez, P. (2012). *Metodología de uso de herramientas de inteligencia de negocios como estrategia para aumentar la productividad y competitividad de una PyMe*. México D.F.: Instituto Politécnico Nacional.
- Harley, O., & Liu, Y. (2017). Towards Industry 4.0 Utilizing Data-Mining Techniques: a Case Study on Quality Improvement. *Procedia CIRP*, 63, 167-172. doi:<https://doi.org/10.1016/j.procir.2017.03.311>
- Inmon, B. (2002). *Building the Data Warehouse* (3 ed.). New York: Wiley Computer Publishing.
- Khanbabaei, M., Mahmood, A., & Radfar, R. (2019). Applying clustering and classification data mining techniques for competitive and knowledge-intensive processes improvement. *Knowledge and Process Management*, 26(2), 123-139. doi:<https://doi.org/10.1002/kpm.1595>
- Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: the complete guide to dimensional modelling*. (3 ed.). Indiana: Wiley.
- Mosquera, R., Parra-Osorio, L., & Castrillón, O. (2016). Metodología para la predicción del grado de riesgo psicosocial en docentes de colegios colombianos utilizando técnicas de minería de datos. *Información Tecnológica*, 27(6), 259-272. doi:<http://dx.doi.org/10.4067/S0718-07642016000600026>
- Ozyirmidokuz, E., Kumru, U., & Mustafa, O. (2015). A Data Mining Based Approach to a Firm's Marketing Channel. *Procedia Economics and Finance*, 27, 77-84. doi:[https://doi.org/10.1016/S2212-5671\(15\)00975-2](https://doi.org/10.1016/S2212-5671(15)00975-2)
- Parama, S. (2018). Business Intelligence Model to Analyze Social Media Information. *Procedia Computer Science*, 135, 5-14. doi:<https://doi.org/10.1016/j.procs.2018.08.144>

- Pérez, C., & Santín, D. (2008). *Minería de datos: técnicas y herramientas*. Madrid: THOMSON.
- Pérez-Gutiérrez, B. (2020). Comparison of data mining techniques to identify signs of student desertion, based on academic performance. *Revista UIS Ingenierías*, 19(1), 193-204. doi: <https://doi.org/10.18273/revuin.v19n1-2020018>
- Phanikanth, K., & Sudarsan, S. (2017). A big data perspective of current ETL techniques. *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 330–334. doi:<https://doi.org/10.1109/ICACCE.2016.8073770>
- Reuter, C., Brambring, F., Weirich, J., & Kleines, A. (2016). Improving Data Consistency in Production Control by Adaptation of Data Mining Algorithms. *Procedia CIRP*, 545-550. doi:<https://doi.org/10.1016/j.procir.2016.10.107>
- Schuh, G., Prote, J.-P., & Hünnekes, P. (2020). Data mining methods for macro level process planning. *Procedia CIRP*, 80, 48-53. doi:<https://doi.org/10.1016/j.procir.2020.05.009>
- Shaker, A., & Abdeltawab, A. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences*, 23, 91–104. doi:<https://doi.org/10.1016/j.jksuci.2011.05.005>
- Silberschatz, A., Korth, H., & Sudarshan, S. (2002). *Fundamentos de Bases de Datos* (4 ed.). Madrid: McGRAW-HILL.
- Siyuan, C., Xingsen, L., Renhu, L., & Shouzhen, Z. (2019). Extension data mining method for improving product manufacturing quality. *Procedia Computer Science*, 162, 146-155. doi:<https://doi.org/10.1016/j.procs.2019.11.270>
- Suharjito, J. (2015). Data mining of automatically promotion tweet for products and services using Naïve Bayes algorithm to increase twitter engagement followers atPT. Bobobobo. *Procedia Computer Science*, 254-261. doi:<https://doi.org/10.1016/j.procs.2015.07.550>
- Vélez, L. (2018). *Gestión de Bases de Datos 1.0*. Sevilla: Instituto de Educación Superior Luis Vélez de Guevara.
- Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. Chennai: WILEY.

Viloria, A., Rodríguez, J., Payares, K., Vargas-Mercado, K., Duran, S., Hernández-Palma, H., & Arrozola, M. (2019). Determinating Student Interactions in a Virtual Learning Environment Using Data Mining. *Procedia Computer Science*, 155, 587–592. doi:<https://doi.org/10.1016/j.procs.2019.08.082>

ANEXOS.

Anexo 1. Resolución de aprobación del proyecto de investigación

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO RESOLUCIÓN N°2322-2020/FIAU-USS

Pimentel, 17 de noviembre de 2020

VISTOS:

El Acta de reunión N° 2610-2020 del Comité de investigación de la Escuela profesional de INGENIERÍA DE SISTEMAS remitida el 12 de noviembre de 2020 mediante oficio N° 0237-2020/FIAU-IS-USS de la Dirección de Escuela profesional de INGENIERÍA DE SISTEMAS, 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 acuerda aprobar la modificación de los temas de Tesis a cargo de los estudiantes y/o egresados 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°: MODIFICAR, el tema de la Tesis perteneciente a la línea de investigación de INFRAESTRUCTURA, TECNOLOGÍA Y MEDIO AMBIENTE, a cargo de los estudiantes y/o egresados del Programa de estudios de INGENIERÍA DE SISTEMAS según se detalla en el anexo de la presente Resolución.

ARTÍCULO 2°: MODIFICAR, la Resolución de Facultad con la que se asigna Asesor especialista en el extremo del tema de la tesis quedando tal como se detalla en el anexo de la presente Resolución.

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

REGÍSTRESE, COMUNÍQUESE Y ARCHÍVESE



Dr. Mario Fernando Ramos Moscoso
Decano - Facultad de Ingeniería,
Arquitectura y Urbanismo
UNIVERSIDAD SEÑOR DE SIPÁN S.A.C.



MRA María Rosita Tobar Wines
Secretaria Asesorado / Facultad de Ingeniería,
Arquitectura y Urbanismo
UNIVERSIDAD SEÑOR DE SIPÁN S.A.C.

Cc: Interesado, Archivo

**FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO
RESOLUCIÓN N°2322-2020/FIAU-USS**

Pimentel, 17 de noviembre de 2020

ANEXO

N°	APELLIDOS Y NOMBRES	TEMA DE TESIS PRIMIGENIO	TEMA DE TESIS MODIFICADO
1	CABREJOS TORRES RAMIRO	INCIDENCIA DE LA METODOLOGÍA MAGERIT V3 EN LA SEGURIDAD DE INFORMACIÓN DE LA EMPRESA DECO INTERIORS SAC	INFLUENCIA DE LA METODOLOGÍA MAGERIT V3 EN LA SEGURIDAD DE INFORMACIÓN DE LA EMPRESA DECO INTERIORS SAC.
2	MARRUFO SALAZAR YOAN JOHEL	MEJORA DE LA USABILIDAD Y ACCESIBILIDAD EN PORTALES WEB EN INSTITUCIONES EDUCATIVAS DEL NIVEL SECUNDARIO	EVALUACIÓN DE LA USABILIDAD Y ACCESIBILIDAD EN PORTALES WEB DE INSTITUTOS DE EDUCACIÓN SUPERIOR PEDAGÓGICA DEL PERÚ
3	CAMONES AGUIRRE OSCAR BRYAN	EVALUACION DE LA VIABILIDAD DE IMPLEMENTACION DE TECNICA BASADA EN ALGORITMO GENÉTICO Y EL FRAMEWORK WS-ATTACKER PARA DETECTAR ATAQUES A SERVICIOS WEB EN UN AMBIENTE DE COMPOSICIÓN DINÁMICO	EVALUACIÓN DE TÉCNICA BASADA EN REGLAS Y ALGORITMO GENÉTICO PARA DETECTAR ATAQUES A SERVICIOS WEB EN UN AMBIENTE DE COMPOSICIÓN DINÁMICO
4	ALVAREZ GONZAGA BRAULIO RICARDO	ANÁLISIS COMPARATIVO DE TÉCNICAS DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE DATOS APLICADAS A BUSINESS INTELLIGENCE	ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS APLICADAS A BUSINESS INTELLIGENCE
5	SANDOVAL ODAR WILLIAM	COMPARACIÓN DE LAS TÉCNICAS DE SEGMENTACIÓN OPTIMIZACIÓN DE ENJAMBRES DE PARTÍCULAS Y AGRUPAMIENTO JERÁRQUICO EN AMBIENTES NO CONTROLADOS DE PLANTACIONES DE ARROZ	COMPARACIÓN DE ALGORITMOS DE SEGMENTACIÓN DE IMÁGENES DIGITALES DE PLANTAS DE ARROZ EN AMBIENTES NO CONTROLADOS
6	MOGOLLÓN GARCÍA MANUEL ESTEBAN	EVALUACIÓN DE HERRAMIENTAS DE SEGURIDAD INFORMÁTICA COMO SOPORTE DE LA NORMA ISO 27001 PARA GARANTIZAR LA GESTIÓN DE LA SEGURIDAD DE LA INFORMACIÓN EN UNA MUNICIPALIDAD DEL PERÚ	DESARROLLO DE UN MODELO DE GESTIÓN DE RIESGOS BASADO EN LA METODOLOGÍA MAGERIT PARA MINIMIZAR LOS RIESGOS DE LA IMPLANTACIÓN Y USO DE TI EN UNA MUNICIPALIDAD DEL PERÚ
7	CASTRO FERNÁNDEZ LEVI RONALD	IDENTIFICACIÓN DE INTRUSIONES A BASE DE DATOS DESDE APLICACIONES WEB UTILIZANDO LOS ALGORITMOS DE APRENDIZAJE AUTOMÁTICO	ANÁLISIS COMPARATIVO DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA IDENTIFICAR ATAQUES DE INYECCIÓN SQL A BASE DE DATOS EN APLICACIONES WEB
8	CALDERON TENORIO CESAR ROLEN	MODELO DE CONTROL AUTOMATICO PARA EL PROCESO DEL RIEGO EN EL CULTIVO DE LA PAPA EN EL CENTRO POBLADO DE CHAQUIL DISTRITO DE LA ESPERANZA PROVINCIA DE SANTA CRUZ REGION DE CAJAMARCA	DESARROLLO DE UN MODELO DE CONTROL AUTOMÁTICO BASADO EN INTELIGENCIA ARTIFICIAL PARA EL PROCESO DEL RIEGO EN EL CULTIVO DE LA PAPA PERUANA



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



"Año de la Universalización de la Salud"

Pimentel, miércoles 07 de diciembre de 2020

Señor(a):

DR. CHRISTIAN ABRAHAM DIOS CASTILLO
Director de Investigación UTP Región Norte
Universidad Tecnológica del Perú SAC
Ciudad.-

ASUNTO:

Presentación de estudiante para realizar caso de estudio.

Es grato dirigirme a usted para expresarle el saludo institucional a nombre de la Escuela Profesional de Ingeniería de Sistemas, perteneciente a la Facultad de Ingeniería, Arquitectura y Urbanismo, de la Universidad Señor de Sipán, a la vez presentar al estudiante del X ciclo, ALVAREZ GONZAGA BRAULIO RICARDO con código universitario 2121819830, e identificado con DNI 44967284, quien recogerá información relevante en la institución que usted representa, como parte de su trabajo de INVESTIGACIÓN, aprobado con resolución N°2322-2020/FIAU-USS, del proyecto titulado "ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS APLICADAS A BUSINESS INTELLIGENCE".

Para ello, solicitamos su autorización, esperando que el estudiante cumpla con todos los requerimientos necesarios.

En espera de su atención a la presente, aprovecho la oportunidad para expresarle mi consideración y estima personal.

Cordialmente,




Mag. Ing. Heber Ivan Mejía Cabrera
Director (e) de la Escuela Profesional
de Ingeniería de Sistemas

UNIVERSIDAD SEÑOR DE SIPÁN S.A.C.

Permiso para el uso de información de la UTP

Por la presente se otorga a BRAULIO RICARDO ALVAREZ GONZAGA, identificado con el número de DNI 44967284, el permiso correspondiente para la obtención de información relacionada al rendimiento académico de los estudiantes matriculados en UTP Campus Chiclayo, cuya información será utilizada para el desarrollo del trabajo de investigación titulado "ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS APLICADAS A BUSINESS INTELLIGENCE".

Cabe indicar que ha sido informado de las pautas correspondientes y que toda denominación a la UTP debe hacerse como "universidad privada de Lambayeque". Se expide este documento para fines del interesado.


Chiclayo, 17 de diciembre del 2020

Atentamente,



Dr. Christian Abraham Dios Castillo
Director de Investigación - Región Norte

Anexo 3. Ficha técnica del equipo utilizado para pruebas

Ficha técnica de equipo de cómputo			
Fabricante	LENOVO		
Procesador	INTEL CORE I3	Velocidad CPU	1.70 GHz
Memoria RAM	4 GB		
Disco duro	500 GB		
Sistema operativo	WINDOWS 7 PROFESSIONAL 64 BITS		
Imagen			
Comentarios			
1) Modelo específico de la laptop en la que se ejecutaron las pruebas es LENOVO B40. Se muestra imagen referencial.			

Anexo 4. Ficha de registro de resultados

Técnica aplicada	
Registros procesados	
Resultados	
MAE	
MAPE	
MAE	
Tiempo Procesamiento	
Precisión	
Comentarios	
1)	
2)	