Modelo predictivo para el procesamiento de datos académicos en Big Data en la educación Superior

Roger Ernesto Alarcón García
Jessie Leila Bravo Jaico
Janet del Rosario Aquino Lalupú
Carlos Alberto Valdivia Salazar
Nilton César Germán Reyes

Savez editorial

Modelo predictivo para el procesamiento de datos académicos en Big Data en la educación Superior

Modelo predictivo para el procesamiento de datos académicos en Big Data en la educación Superior

Roger Ernesto Alarcón García
Jessie Leila Bravo Jaico
Janet del Rosario Aquino Lalupú
Carlos Alberto Valdivia Salazar
Nilton César Germán Reyes

Savez editorial

Roger Ernesto Alarcón García
Jessie Leila Bravo Jaico
Janet del Rosario Aquino Lalupú
Carlos Alberto Valdivia Salazar
Nilton César Germán Reyes

Modelo predictivo para el procesamiento de
datos académicos en Big Data en la educación Superior

ISBN: 978-9942-603-58-6

Savez editorial
Título:
Modelo predictivo para el procesamiento de
datos académicos en Big Data en la educación Superior

Primera Edición: Julio 2022

ISBN: **978-9942-603-58-6**
Obra revisada previamente por la modalidad doble par ciego, en caso de requerir información sobre el proceso comunicarse al correo electrónico editor@savezeditorial.com

ÍNDICE DE CONTENIDOS

I. INTRODUCCIÓN 3
II. MARCO TEÓRICO 8
III. METODOLOGÍA 29
3.1. Tipo y diseño de investigación 29
3.2. Variables y Operacionalización 30
3.3. Población y Muestra 30
3.4. Técnicas e instrumentos de recolección de datos. 33
3.5. Procedimiento 33
3.6. Aspectos éticos 34
IV. ANÁLISIS SITUACIONAL 34
V. MODELO PREDICTIVO PARA EL PROCESAMIENTO DE DATOS ACADÉMICOS EN BIG DATA EN LA EDUCACIÓN SUPERIOR 48
VI. DISCUSIÓN 61
VII. CONCLUSIONES 64
VIII. RECOMENDACIONES 65
REFERENCIAS 66

I. INTRODUCCIÓN

A finales de la década se ha detectado que hay un alto crecimiento en el poder de generar datos de todo tipo y el de recolectarlos, esto debido al incremento en el poder que tienen las máquinas en el procesamiento de los mismos y la reducción de los altos costos de su almacenamiento. Sin embargo, toda esta data existente nos oculta una gran cantidad de información, la cual puede ser muy importante desde el punto de vista estratégico, a la que es imposible acceder utilizando las actuales técnicas de recuperación y generación de información.

Las actuales organizaciones están generando un alto volumen de datos, disponibles para su análisis, pero acarrea un problema, las limitantes capacidades humanas han sido superadas para lograr analizar los datos y generar conocimiento útil que le permita a la organización tomar las mejores decisiones. El uso de modelos matemáticos centrados en principios estadísticos, han permitido solucionar muchos de los problemas que aparecen en las ciencias, tanto de carácter empíricos como teóricos, motivando la utilización de técnicas y herramientas que proporcionen la posibilidad de generar nuevo conocimiento basado en el análisis de los datos.

Por lo tanto, actualmente se observa el requerimiento de gestionar grandes volúmenes de datos, las cuales puede estar en formato estructurado o no estructurado, producidas por el internet o por diversas organizaciones públicas o privadas, generando una enorme producción de información digital que terminan en un repositorio de almacenamiento, y no es procesada para beneficio interno, provocando incluso que se termine eliminando sin darse cuenta que puede ser útil a futuro.

Uno de los grandes activos que actualmente tienen las empresas públicas o privadas, es la producción de información, ésta ha generado no solo un alto desafío tecnológico, sino también, un desafío científico global en la que el desarrollo de generar, administrar, explotar, interpretar y clasificar la información se ha convertido en una pieza fundamental en las empresas.

En general, es de gran interés el descubrir patrones, tendencias, perfiles u otras relaciones que hasta ese entonces permanecían ocultas y no eran utilizadas por las empresas, por lo que las técnicas en minería de datos permiten apoyar en la generación de consultas y el análisis que ayuden en un adecuado tratamiento de los datos, siendo necesario extraerlos de almacenes en los sistemas tradicionales o los muy conocidos repositorios unificados para datos empresariales (Data Warehouse), o la Big data para información heterogénea o no estructurada que se extrae de múltiples fuentes.

Por lo que, la base para el conocimiento empresarial parte de un buen tratamiento de los datos, siendo por lo tanto una nueva necesidad el poder adaptarse a estos cambios en el contexto digital, llevándonos de confirmar o verificar a la producción de hallazgo a través del uso de modelos predictivos que se encargan de detectar los datos ocultos que se encuentran en los volúmenes de datos organizacionales.

Pese al escepticismo, hemos pasado a una realidad en la que los datos son vitales. Según Santos (2015) menciona que el vicepresidente de la consultora Gartner Peter Sondergaard, manifiesta: “La información es la gasolina del siglo XXI y el análisis de datos el motor de combustión”. Además, indica que el uso de nuevas y robustas tecnologías de aprendizaje automático a partir de datos, está permitiendo aplicar métodos que traducen textos, predecir trayectorias o inclusive recomendar búsquedas.

Álvarez Valle (2013) señala que el uso de grandes volúmenes de datos o más conocida como la Big data se está convirtiendo en una oportunidad millonaria de negocio permitiendo a las empresas ser más competitivas, a la vez que las encaminará a tomar mejores decisiones empresariales.

La Universidad en estudio ubicada en la zona norte del Perú, es una institución pública que no está ajena al tratamiento y procesamiento de los datos especialmente en el ámbito académico, en la cual hay gran cantidad de datos que tiene registrada por años a través de los diferentes semestres académicos, especialmente en los procesos de matrícula y notas de los distintos estudiantes

universitarios; sin embargo, después de haber realizado un análisis fáctico de estos procesos, ha permitido detectar las siguientes manifestaciones:

- La diversidad de los datos (estructurados y no estructurados) que implican nuevos enfoques de almacenamiento y de análisis haciendo uso de técnicas predictivas.

- La falta de manipulación de los datos almacenados, los cuales no son tratados y que permitan obtener nuevo conocimiento haciendo uso de técnicas predictivas.

- El incremento de la información adquirida por la institución cada nuevo semestre no es evaluado ni analizado.

Estas manifestaciones se sintetizan en el problema científico: el inadecuado procesamiento de datos usando técnicas predictivas aplicados a grandes volúmenes de datos y el análisis de la Big data limita el tratamiento de los datos académicos.

Lo que conlleva a plantear posibles causas del problema antes mencionado:

- Insuficiente referencia teórica sobre técnicas predictivas aplicadas a la Big data, en el proceso de extracción del conocimiento.

- Limitaciones de técnicas predictivas que den un soporte para la extracción de conocimiento de utilidad basado en grandes cantidades de datos.

- La falta de capacidad en el proceso de extracción de conocimiento de datos usando técnicas predictivas que generen conocimiento útil para la institución.

Estas valoraciones causales sugieren profundizar en el procesamiento de datos en la Big data, objeto de la presente investigación.

Según Vivas et al. (2015), la gran cantidad de datos con la que cuentan las organizaciones permiten, en base al procesamiento y generación de información, que las empresas tomen mejores decisiones, basándose en el uso de las bases de datos relaciones y otras alternativas con las que cuenten, enfocándose en estas tareas de negocios inteligentes analizando datos organizacionales y aplicando el uso de la Big Analytics.

Desde el punto de vista de Cano (2016), cita la representatividad de los datos como un aspecto que sigue siendo uno de los principales problemas cuando se trabaja con una muestra para analizar un fenómeno, dado que no solo debemos enfocarnos en el "cuántos", es decir, tener una data amplia de gran cantidad, sino también en el "cómo", siendo para esto muy importante un correcto control de los orígenes y validez de los datos, ya que pueden representarse como débiles, sesgados, triviales, no interpretables, entre otros, por lo que, tener mayor disponibilidad de datos no debe ser el único criterio que permita definir su valor, ya que se puede llegar a sucumbir al convertir datos masivos, en sesgos masivos, que no permitan una buena toma de decisiones por parte de las empresas.

También, Cano (2016) indica que otra característica que manipulan los datos es la conocida como la "maldición de la dimensión", en el sentido de que, se debe garantizar resultados con altos niveles de significancia (garantía probabilística) y no nos aturdamos con una sobreabundancia de datos que actualmente se produce en las organizaciones; por otro lado otros autores, han realizado experimentos simulando el tratamiento de datos relacionados con un fenómeno específico, basándose en algoritmos de generación de números pseudoaleatorios entendiendo que no hay correlación alguna, pero con estos experimentos se lograban encontrar de manera fácil buenas correlaciones, por lo que se puede llegar a la conclusión de que no era que hubiese un patrón en los datos, sino que no fuimos capaces de reconocerlo, dado que ahora se tiene al alcance la posibilidad de proliferar patrones, modelos y correlaciones.

Por otra parte, en la investigación de García et al. (2016), encuentra que las capacidades de procesamiento en los sistemas clásicos que utilizan la minería de datos ha sido superada rápidamente por el volumen actual de los datos, por lo que se está ingresando a una nueva era, el de los datos masivos, en donde el volumen, variedad y velocidad son características primordiales de la Big data.

El procesamiento de datos, pasa por diversas fases, primero se debe extraer los datos, los cuales generalmente están ubicados en diferentes fuentes de datos, posterior a esta fase, se realiza el preprocesamiento, etapa muy esencial en el proceso de descubrimiento de conocimiento, aquí se realiza en primera instancia la limpieza de datos, luego pasamos a integrar los datos, posteriormente a su transformación y culminamos con la reducción de los datos, para luego pasar a la siguiente fase denominada minería de datos. Por lo que, después de aplicar esta fase de procesamiento previo, los datos resultantes pueden ser observados como una fuente compacta y conveniente de datos con calidad, que sirvan para extraer conocimiento a través de la aplicación de ciertos algoritmos.

Además, Moreno (2014) señala, que la sociedad produce una gran cantidad de datos y muchos de estos a su vez no llegan a ser procesados, ya que los sistemas tradicionales no cuentan con la capacidad computacional para ello, incluyendo también que una gran cantidad de empresas no cuentan con soluciones unificadas que les permita capturar y luego realizar un análisis con ellos.

Además, Escobar y Mercado (2019) en su investigación señala que las organizaciones actualmente están considerando el uso de aplicaciones de Big data en sus procesos productivos y sociales, aumentando su representatividad en el mundo digital, el cual les permitirá mejorar su posicionamiento en el contexto social a nivel mundial.

De lo descrito por estos autores se evidencia que aún son insuficientes los referentes teóricos y prácticos en cuanto a la dinámica del proceso para sistematización, fundamentación teórica, desarrollo de actividades, su apropiación, generalización para el procesamiento de datos.

Por lo que se determina como campo de acción: Dinámica del proceso de procesamiento de datos en la Big data.

Es así que existe una brecha epistémica en donde el estudio del objeto y el campo de acción revelan, que no ha sido lo suficientemente analizadas las técnicas predictivas integradas a procesar grandes volúmenes de datos incluyendo todas las características inmersas en la Big data aplicadas a las instituciones de educación superior.

II. MARCO TEÓRICO

Tomando en consideración los antecedentes que abordan este tema de manera prioritaria se encuentran las siguientes investigaciones:

Vite Cevallos et al. (2020) plantea diversos modelos de Big Data uno de ellos, es un modelo propuesto en la Universidad de Ciencia y Tecnología Huazhong en China que permita a los pequeños productores agrícolas tomar decisiones basados en el denominado Modelo BIG DATA Agrícola, otro desarrollado en China enfocado en la nube computacional denominado Modelo de Procesamiento en Internet de las cosas en agricultura, en donde trabaja con sensores que permiten la lectura de datos, su recopilación, transmisión y almacenamiento haciendo uso de la nube. Y por último un modelo propuesto en la India denominado Modelo Big Data para agricultura inteligente, el cual, genera una arquitectura multidisciplinaria integrando 5 módulos.

Según Russo et al. (2016), la investigación pretende seleccionar, diseñar y desarrollar un modelo que haga uso de algunos algoritmos que permitan la correcta clasificación y predicción de los datos, haciendo uso de conjuntos de datos de entrenamiento necesarios para este fin. Por lo que se enfoca en el tratamiento masivo de datos y su procesamiento mediante sistemas inteligentes.

También, Russo et al. (2016) menciona que en el proceso de descubrimiento de conocimiento se establecen 5 fases: la primera fase corresponde a la integración y

recopilación de los datos; la segunda fase corresponde a la selección de los datos, la limpieza de los mismos y su transformación; la tercera fase le compete a la minería de datos aplicando ciertos algoritmos de acuerdo a sus necesidades; la cuarta fase atañe a la evaluación e interpretación de los resultados y finalmente la quinta fase corresponde a la difusión. Comúnmente tanto minería de datos como proceso de extracción de conocimiento se usan como palabras sinónimas, no deben ser consideradas de esta manera, ya que, la minería de datos es en realidad parte de una etapa de todo el proceso general de extracción de conocimiento. Por lo que se debe considerar a la minería de datos como un mecanismo que permite explorar, analizar y extraer información que resulta de mucho valor.

Según Quinteros et al. (2016) indica que existe una rama denominada Minería de Datos Educativos o EDM, disciplina la cual desarrolla métodos para obtener y extraer información valiosa en base a lo que generan los entornos educativos, con la finalidad de poder mejorarlo continuamente. Además, indica que es factible la aplicación de la minería de datos centrado en todos los datos que administran las universidades, contando con los sistemas de gestión y transaccionales que están implementando las universidades nacionales, que permitan integrar su áreas y procesos institucionales; generalmente orientados a diversos temas y con grandes volúmenes de datos, almacenadas en bases de datos con gran cantidad de dimensiones. Y finalmente indica, que los métodos para minería de datos educativos se dividen en dos grupos; el de verificación y el de descubrimiento, entre los que destacan los métodos de clasificación, agrupamiento, predicción, minería de reglas de asociación, redes neuronales y minería web.

Según Tolosa et al. (2016) en su investigación manifiesta que el uso de motores de búsqueda basados en consultas de acuerdo a las necesidades del usuario permite el acceso a la información en internet. Estas consultas recuperan parte del espacio web, el cual se ha recorrido, reunido y manipulado, por lo que es un factor importante y muy esencial en los procesos que ocurren en la industria o el entretenimiento, entre otros. Por otra parte, indica que las fuentes de información como por ejemplo sensores y redes sociales al igual que su almacenamiento están creciendo

agigantadamente, generando mayor complejidad y obligando a responder en tiempo real. Por lo que la mayoría de los problemas se trataban con el enfoque de la minería de datos, ahora estos problemas pasaron a formar parte de los grandes datos, esto implica, una mayor complejidad, debido al gran incremento del volumen de los datos. Adicionalmente, se presenta otra dificultad, lo referente a las arquitecturas que requieren ser flexibles, incluyendo, el cómputo y el almacenamiento. Por último, indica que las evidencias son la base para que las organizaciones tomen decisiones acertadas, las cuales son soluciones a base de los datos y no de simple intuición. Considerando que el descubrimiento de nuevo conocimiento implica el uso de técnicas que son transversales a todas las disciplinas, por lo que existe una gran cantidad de soluciones de optimización que por el momento no han sido investigadas, relacionadas con motores de búsqueda que se apliquen directamente a estos grandes volúmenes de datos.

Malberti et al. (2016) propone el uso de herramientas como Knime, Weka, R, Rapidminer y en algunos casos el uso de programación con Phyton, para acceder y analizar datos enfocándose en el paradigma de la Ciencia de Datos, cuyo objetivo es el de reconocer, analizar y describir grandes volúmenes de datos que provienen de las redes sociales o áreas como la astronomía, educación, bibliotecología, entre otros.

Britos et al. (2016) investiga sobre los sistemas de recuperación de información específicamente multimedia, estos aspectos como el diseño de nuevos índices aplicados a grandes volúmenes de datos, o distintas consultas sobre estos tipos y su relativa eficiencia al manipular muchos datos.

De Battista et al. (2016) indica que actualmente existe un crecimiento permanente en la cantidad de datos que las aplicaciones están generando y almacenando enfocándose en la complejidad de los atributos, los cuales sirven para describir objetos del fenómeno analizado. Pero, además, manifiesta que ha rebasado la capacidad de poder procesar los datos, conservarlos de manera adecuada, analizar y comprender en base a estos grandes repositorios. Todas las organizaciones que están generando datos de forma masiva, desean poder extraer nuevo conocimiento,

que les permita en base a este análisis tomar decisiones y no solo que sirvan como una protección almacenada en las computadoras y se pierdan sin nunca ser utilizados.

El procesamiento de datos que se aplica hace referente al uso de técnicas que se utilizan en minería de datos. Por lo que no están orientadas a ser eficientes en grandes volúmenes de datos como lo es la Big data, teniendo en consideración que está enfocada en el volumen de datos, velocidad con la que llegan los datos, su variedad estructurada o no estructurada, la confianza por parte de los usuarios para suministrar la información y la protección de los datos. Asumiendo también su veracidad, es decir, la precisión y la confianza de los datos que se manejan.

Además, según Quiroz Martinez et al. (2020), el uso de la Big Data apoya la construcción de sistemas novedosos que permiten tomar decisiones más acertadas, benefician tanto a la población, específicamente al personal administrativo. Por lo que, se ve la necesidad de usar nuevas herramientas tecnológicas que permitan procesar toda esta data que va creciendo de manera exponencial y que pueda servir para mejorar la celeridad con la que estos sistemas permiten tomar decisiones.

Según García et al. (2016) menciona que la calidad de los datos influye notablemente en la calidad del conocimiento que se extrae, pero existen ciertos factores negativos que afectan dicha calidad como por ejemplo el ruido, o los valores perdidos, valores inconsistentes o algunos datos superfluos. Por lo que, la baja calidad que presentan los datos acarrea que generalmente se obtenga también una baja calidad en el conocimiento obtenido.

El procesamiento, almacenamiento y transferencia de grandes volúmenes de datos forman parte del denominado Big data, todos estos factores forman parte determinante para el Cómputo de Alto Rendimiento “CAR”. Los algoritmos usados para el procesamiento de los datos deben permitir agilizar o acelerar el cómputo de los mismos para reducir los tiempos en la toma de decisiones.

Así mismo Duque Méndez et al. (2016) distingue que las tareas de extracción, transformación y carga presentan mucha complejidad cuando se construyen los almacenes de datos, tanto de tiempo y recursos como también de los costos asociados a su consumo. Además, se mostró un modelo de extracción, transformación y carga para datos hidrometeorológicos, consiguiéndose buenos resultados al aplicarlo en el caso de estudio con fuentes de datos reales a un volumen alto de datos.

Como manifiesta Camejo Corona et al. (2019) en su investigación, la presencia de problemas de predicción especialmente cuando se manipulan grandes volúmenes de datos, por ejemplo, la gran cantidad de variables que se manipulan, y esto se ha generado porque las empresas están recolectando más datos de los que pueden procesar, debido a la incursión de sensores que se manipulan actualmente en la industria global. Provocando serios problemas como la gran cantidad de tiempo que requiere para el procesamiento computacional que permita realizar predicciones usando los datos, lo que nos indica que se deben replantear las técnicas clásicas usadas en las predicciones.

Actualmente el uso de los sistemas de información por parte de las organizaciones de cualquier sector, apoyan en gran medida en el procesamiento de la información permitiendo así gestionar adecuadamente los negocios, pero es necesario también controlar la calidad de los datos que se tienen almacenados, de modo que permita fácilmente poder determinar anomalías que se puedan presentar, con la intención de realizar una etapa de corrección, para que las decisiones basadas en estos datos sean las más correctas.

Una de las tareas importantes cuando se manejan datos en las empresas u organizaciones es el tema de la seguridad de los datos, ya que, dependiendo de los datos, si estos son preciados, delicados o muy críticos, la empresa debe considerarlos como activos valiosos para ella, y establecer estrategias que permitan su protección, confidencialidad, disponibilidad e integridad de la información. Así mismo, es importante el uso de protocolos de seguridad, que den un mayor nivel de

protección, dado que se presentan entradas no autorizadas de hackers en sistemas de bancos y compañías de transferencia de dinero, entre otros.

Por otra parte, las estrategias orientadas a la Big data, presentan problemas en cuanto a su volumen y también a la diversidad de sus fuentes, por lo que debemos enfocarnos en todo el proceso desde el inicio con la captura de los datos ya sea por procesos secuenciales o por tiempo real, hasta llegar a la generación de nuevo conocimiento. Todo este proceso implica que los datos sufren cierta manipulación desde la corrección de los mismos, la especificación de los datos que servirán y su almacenamiento.

Continuando con los datos, las empresas deben evaluar que conocimiento le es útil dependiendo de los datos que posee, esto le permitirá a estas organizaciones tener éxito en el futuro, por lo que entender la clasificación de los datos es importante, los estructurados los cuales se almacenan en bases de datos relacionales donde todo su formato está predefinido, haciendo uso de sistemas como los ERP o CRM, los no estructurados, enfocándonos en que no cuentan con estructura definida como por ejemplo los videos, audio y otros archivos y por último los semiestructurados, orientados a documentos como HTML basado en etiquetas y marcadores que permiten su entendimiento, al igual que los XML o SGML.

Peñaloza Báez (2018) en su investigación busca encontrar relaciones causa-efecto o proyecciones usando las probabilidades, basadas en asociaciones, patrones y tendencias en los datos. Por lo tanto, se requiere de técnicas y algoritmos que hagan frente a esta nueva arquitectura de la información. A lo que él considera como la aplicación de Big data, convirtiendo en información útil grandes volúmenes de datos de alta calidad, siendo necesario el uso de herramientas tecnológicas que nos permitan realizar la recopilación, manipulación, almacenamiento y el análisis de los mismos.

Según Hernández-Leal et al. (2017) indica que el enfoque de la Big data y toda su tecnología asociada está generando resultados que permiten vislumbrar beneficios orientados a la optimización de recursos y disminución de los tiempos de ejecución.

Además, la Big data presenta más dimensiones relevantes como son la veracidad, la variedad en los datos y la velocidad con la que debe trabajar, pero el inconveniente que aún se tiene corresponde a su implementación, el cual resulta aún costoso agregando también que la adaptación tecnológica se va incrementando en tiempo.

Téllez Carvajal (2020) aporta en su investigación que la utilización de técnicas de análisis manipulando grandes volúmenes de datos apoyan en generar predicciones para prevenir las posibles violaciones a los derechos humanos, estas predicciones permiten que las entidades tomen decisiones teniendo como base una mayor información, pero también advierte que si se realiza una mala programación para la generación de información su utilización puede generar riesgos en la interpretación de los mismos.

Entre estas teorías podemos mencionar en primer lugar al modelamiento relacional de datos, base para todo procesamiento. El cual se centra en el uso de las relaciones como base fundamental de todo, siendo por lo tanto un conjunto de relaciones lo que representa a una base de datos. Estas relaciones son muy parecidas a lo que conocemos como tablas de valores en donde una fila está compuesta de valores relacionados también conocida como tupla. En un modelo de base de datos relacional, un hecho es representado mediante una tupla individual (Elmasri y Navathe, 2007).

Por lo que, en este modelo, las columnas representan a cada uno de los atributos que se desea modelar del fenómeno o ente y la relación representa el nombre de la tabla. Es necesario dentro de este concepto conocer que los diferentes valores que se pueden registrar en cada atributo corresponden al dominio del mismo, siendo estos atómicos, entonces, no se pueden sub dividir en otros valores. Formalmente se expresa la relación denotándola de la siguiente manera: R (A1, A2, ..., An), en donde, A1, A2, .., An representan los atributos de la relación y R la relación en sí.

Además, dentro del modelo relacional se han definido dos lenguajes formales que permiten expresar consultas sobre la base de datos, estos son el álgebra y cálculo

relacional. En ambos, se incluyen un conjunto de operaciones que permiten poder manipular una base de datos relacional.

El lenguaje más práctico y fácil de entender corresponde al álgebra relacional, el cual está compuesto por diversas operaciones como la intersección similar que la intersección (A ∩ B) de conjuntos en donde se obtienen como resultado todas las tuplas que están en la relación A con la relación B, diferencia (A - B) en la que se obtiene como resultado las tuplas de A que no aparecen en B, la unión de relaciones (A U B) en donde la relación resultante contiene las tuplas tanto de A como de B y que no se repitan, en donde la primera condición es que ambas relaciones contengan los mismos atributos y la segunda condición es que los dominios por atributo sean iguales en ambas relaciones, el producto cartesiano (A*B) en el que obtenemos como relación resultante los atributos de A unidos con los atributos de B y como resultado de tuplas a las combinaciones posibles de A y B, la selección en la que el resultado de tuplas coincide con aquellas que cumplen con la condición especificada, proyección en donde la relación resultante contiene ciertos atributos especificados y otras operaciones adicionales.

Por otro lado, hay tres aspectos relevantes en la gestión de los datos, y estos son el adquirir y almacenar los datos, luego la limpieza y depuración, y finalmente, la preparación de los datos para su análisis final.

Se ha encontrado una gran cantidad de técnicas de procesamiento correspondiente a la gestión de los datos, y su posible clasificación como se aprecia en la Tabla 1.

Tabla 1 Técnicas de procesamiento aplicadas a la gestión de los datos

Tipos de Datos	**Ejemplos de técnicas de procesamiento**
Texto	**Extracción de la información**: en donde se extrae las entidades y relaciones que se pueden encontrar; **Resumen de texto**: El uso del lenguaje natural permitiendo generar resúmenes de los documentos;

Tipos de Datos	Ejemplos de técnicas de procesamiento
	Respuesta a la pregunta: Uso de lenguaje natural generando respuestas a cada una de las preguntas y **Análisis de sentimiento**: Generando respuestas que pueden ser positivas o negativas al analizar textos de opinión.
Audio	Para el enfoque basado en transcripción se puede utilizar la aplicación de analítica de texto; para el enfoque en la fonética a través de la representación fonética de un término, el cual se analiza para buscar secuencias.
Video	Aquí se pueden presentar dos arquitecturas la basada en el borde donde localmente se analiza el video y la basada en el servidor el cual implica un equipo específico y sofisticado denominado servidor que realiza el análisis del video.
Redes sociales	Se pueden presentar dos analíticas, la primera basada en la estructura, en la que se extrae inteligencia y se sintetiza los atributos, utilizando para esto técnicas como el análisis de influencia social, la detección de comunidades y la predicción de enlaces.

Existen diferentes metodologías cuando hablamos de analítica de datos orientadas a productos y servicios basados en diversas implementaciones tecnológicas como se aprecia en la tabla 2.

Tabla 2 Ejemplos de metodologías para modelar y analizar grandes volúmenes de información

Metodología	**Descripción**	**Aplicaciones / Ejemplos**
Análisis espacial	Conjunto de técnicas que permiten resolver problemas en datos sobre propiedades geográficas, geométricas y topológicas.	Entre los ejemplos tenemos a las regresiones espaciales y las simulaciones.
Análisis de redes	Conjunto de técnicas que se centran en la evaluación de nodos dentro de una red o grafo.	Identificación de líderes de opinión para focalizar campañas de marketing. Identificar cuellos de botella en flujos de información de una empresa. Modelamiento de redes de transporte y predicción del tiempo de desplazamiento de un punto a otro.
Aprendizaje automático (Machine Learning)	Subespecialidad de la Ciencia de la Computación (denominada "Inteligencia Artificial") que se ocupa del diseño y desarrollo de	Predicción de fenómenos como crimen, deserción escolar y universitaria,

Metodología	Descripción	Aplicaciones / Ejemplos
	algoritmos que permiten inferir comportamientos basados en datos empíricos.	esperanza de vida post-operatoria, ventas. Sugerencias y recomendaciones de productos basado en el análisis histórico. Procesamiento de lenguaje natural, reconocimiento de patrones y detección de anomalías.
Pruebas A/B	Esta técnica que analizan una variable objetivo manipulando la comparación de varios grupos de prueba sobre un solo grupo de control.	Evaluar la efectividad de un tratamiento médico o de una de campaña de marketing.
Visualización analítica de datos	Forma de presentación visual de los datos que han sido descubiertos para la posterior toma de decisiones.	Análisis visual interactivo de componentes principales.

Por otro lado, el uso del análisis de datos a través de la manipulación de técnicas computacionales que establecen modelos predictivos, los cuales permiten inferir resultados usando las probabilidades, posibilitando así identificar oportunidades de negocio.

Además, los científicos de datos son los designados para tratar los datos y convertirlos en información que expresan valor para la empresa. Esto está generando el desarrollo de la rama del Big data y su aplicación con respecto a datos reales.

El apoyo que brinda el análisis predictivo enfocado en el mundo empresarial cumple un papel fundamental actualmente. Las empresas la utilizan para adecuar las decisiones de negocio y reducir el riesgo, mejorar la información obtenida del cliente y generar una alta capacidad para predecir comportamientos.

En el mundo empresarial cobra un papel importante y fundamental el análisis predictivo, teniendo en cuenta que permite reducir el riesgo, aumento de la capacidad para poder predecir comportamientos de los clientes, entre otros, los cuales hacen uso de algunas técnicas como: los árboles de decisión, que representan modelos basados en algoritmos de aprendizaje supervisado, el cual usa técnicas que extrapolan en dos conjuntos de datos homogéneos, partiendo primero de un nodo raíz, y graficando los nodos posteriormente en lo que se conocen como hojas, dando la apariencia real de un árbol. Los árboles de decisión se usan porque son intuitivos, y fáciles de usar y controlar. Constituyen una opción idónea para resolver problemas, ya que se obtienen datos sobre la ruta óptima. Además, el análisis de regresión nos permite estimar relaciones entre las variables, en ellos se puede apreciar dos modelos: los lineales y los logísticos. Y, por último, el uso de redes neuronales, que nos permiten modelar relaciones complejas.

Caracterización del proceso de procesamiento de datos y su dinámica.

El procesamiento de datos, es el proceso por el cual se recaban datos y se transforman en información que será útil, por lo que se convierten los datos de su forma original a un formato más entendible, dándoles la forma y el contexto necesarios para que puedan ser entendidos por las computadoras y las personas en las organizaciones, es muy importante que este proceso se realice de la manera correcta para no afectar negativamente al producto final o los resultados obtenidos.

El uso de las tarjetas perforadas en el siglo pasado utilizados en el censo de los EE.UU. acarreó la utilización de sistemas mecánicos que permitan procesar las tarjetas de manera rápida. Además, el uso de estas tarjetas ha permitido convertirse en un medio en que el procesamiento de datos estimula el avance de los computadores (Silberschatz et al., 2002).

Según Schab et al. (2018) indica que el procesar y analizar grandes volúmenes de datos producidos actualmente, con la posibilidad de detectar tendencias y patrones, que permanecen ocultos en los datos, impactan directamente en la toma de decisiones de cualquier área de estudios. También se conoce que generan datos a gran velocidad y en grandes cantidades.

Además, indica que la manipulación registro por registro es el más conveniente para realizar el procesamiento de datos, generándose resultados que pueden emplearse en diversos tipos de análisis de muestreo, correlacionales o de agregaciones. Los resultados obtenidos de estos análisis brindan un mayor conocimiento del negocio y de las actividades que puedan tener con sus clientes.

Podemos indicar que el procesamiento de datos se distingue de la conversión de datos, ya que implica el cambio de datos a otros formatos, y no implica ninguna manipulación de datos. Entonces, al momento del procesamiento, los datos sin manipular se usan como entrada para generar información como salida, casi siempre en forma de informes y otras herramientas analíticas.

Además, se pueden especificar etapas en el procesamiento de datos como son la recolección de datos, preparación de los datos, entrada de los datos correctos y el procesamiento, la interpretación de los datos y el almacenamiento de los mismos.

La primera etapa, corresponde a la recolección de datos, en la cual pueden intervenir diferentes fuentes disponibles como por ejemplo hojas de cálculo, archivos de texto, almacenes de datos, entre otros. Un punto a considerar en esta etapa es conseguir que los datos a manipular sean de calidad, por lo que las fuentes de las cuales se extraiga tendrán que ser fiables y de buena calidad.

La segunda etapa, implica la preparación de los datos, partiendo de lo recabado en la etapa anterior, se procede con la limpieza, cuya intención es detectar los datos erróneos, incorrectos o incompletos y proceder a eliminarlos para que no afecten el procesamiento de los mismos. Esta etapa también es llamada pre-procesamiento, en la que el producto final es un conjunto de datos listos para poder trabajar con ellos.

Adicionalmente a esto, Nuñez-Arcia et al. (2016), se enfoca en la evaluación y análisis de posibles errores que se puedan presentar en las fuentes de grandes volúmenes de datos. La presencia de errores es independiente del formato por el cual se presenta los datos.

La tercera etapa, corresponde a la introducción de datos y su tratamiento, para esta etapa ya se manipulan los datos limpios de la etapa anterior agregándolo a un repositorio para su posterior tratamiento, este tratamiento puede demorar significativamente de acuerdo al volumen de datos, además, utiliza algoritmos establecidos y dependen del estudio a realizar con los datos como por ejemplo estudiar patrones, determinar necesidades, entre otros.

La cuarta etapa, consiste en la interpretación de los datos, también conocido como salida de datos. En esta etapa, los datos son legibles y se pueden presentar en diferentes formas como por ejemplo gráficos, videos, imágenes, texto simple, entre otros.

El almacenamiento de datos es la última de las etapas, cuando los datos están procesados, se almacenan para su futuro uso. Generalmente la utilidad de los mismos es a posteriori. Una de las ventajas para los empleados de las organizaciones, es tener los datos bien almacenados, esto permitirá un acceso fácil y rápido cuando se los necesite.

Elementos del procesamiento de datos

Para lograr que la computadora pueda procesar datos se debe primero, evaluar que datos son los que se necesitan y convertirlo a un formato que sea entendible por la

computadora. Teniendo los datos, se puede obtener información importante a través del uso de diversos procedimientos para su aplicación.

Enfocándose en el Procesamiento de datos, éste puede implicar diferentes procesos, entre ellos se destaca lo siguiente:

- **Entrada de datos**: Por una parte, se hace uso de formularios para el ingreso de los mismos de manera automatizada, y en caso el registro se realice manual, es necesario una buena concentración durante un periodo largo de tiempo.

- **Captura de datos**: Múltiples métodos están disponibles para capturar datos de documentos, debe tenerse en cuenta el origen de los documentos que se deben capturar. Algunos métodos son: reconocimiento óptico de caracteres (OCR), reconocimiento inteligente de caracteres (ICR), reconocimiento de código de barras, reconocimiento inteligente de documentos (IDR), captura de voz, entre otros.

- **Depuración de los datos**: Implica la revisión de manera cuidadosa del conjunto de datos y los protocolos asociados con cualquier tecnología de almacenamiento de datos en particular.

- **Integridad de los datos**: Se refiere a la fiabilidad que nos pueden brindar los datos, así como también su exactitud. Se deben considerar los datos completos sin variar el original.

- **Codificación o cifrado de datos**: Es un proceso de seguridad, los cuales utilizan diversos algoritmos que convierten información con la intención de que sea ilegible, y con esto protegen sus datos sensibles como por ejemplo el tema de las tarjetas de crédito.

- **Transformación de los datos**: Proceso que permite convertir de un formato a otro los datos o información que se presentan, basándose en dos fases claves: el mapeo de los datos y la generación de código.

- **Traducción de datos**: es una parte inherente de una solución, además, el XML se está convirtiendo rápidamente en el estándar para intercambiar información entre aplicaciones.

- **Resúmenes de datos**: Son formatos menos complicados, que nos permita manipular los almacenes de datos, también proporcionan la capacidad de dar una visión global de volúmenes dispares de datos.

- **Validación de datos**: Este proceso debe asegurar que los datos estén claros y limpios, además de comprobar la integridad y validez que son ingresados a través de diferentes aplicaciones.

- **Modelado de datos**: Forma de mostrar los datos ordenados y organizados que permitan una utilización fácil por parte de las bases de datos.

- **Análisis de datos**: Técnicas y procesos cuantitativos y cualitativos que se utilizan con la finalidad de incrementar la productividad y por ende la ganancia de las empresas, aplicando diversas técnicas.

- **Visualización de datos**: Permite a través de formatos gráficos mostrar los datos, brindando la posibilidad de detectar correlaciones, tendencias o patrones que podrían no ser detectados en los informes tradicionales que se tienen.

- **Almacenamiento de datos**: Constituye el espacio físico en donde se aloja la base de datos, considerando sus características propias.

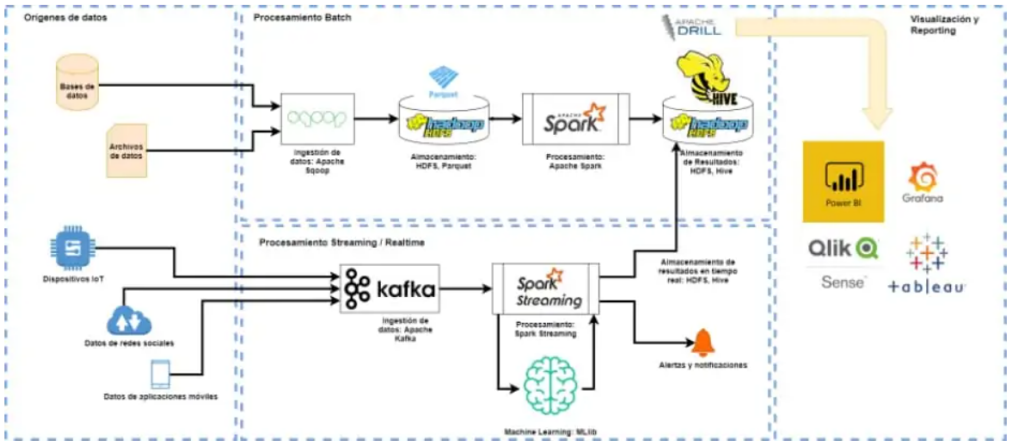
- **Minería de datos**: proceso en el cual se pueden detectar anomalías, patrones o correlaciones, que le permitan a la empresa aumentar los ingresos y reducir costos.

- **Interpretación de datos**: analiza datos reales y llega a una conclusión.

Arquitecturas de procesamiento de datos

En la década del 2000 aparecen herramientas con la posibilidad de procesar gran cantidad de datos, comenzando a investigarse métodos que permitan de manera distribuida procesar y almacenar datos. Cuando indicamos grandes cantidades hacemos referencia a la Big data. La figura 1 muestra la arquitectura orientada a manipular grandes volúmenes de datos para el procesamiento de datos productivos.

Figura 1 Arquitectura de Big Data para Procesamiento de Datos Productivos



Fuente: https://www.lisdatasolutions.com/blog/arquitectura-big-data-para-procesamiento-de-datos-de-produccion/

Las funciones que debe tener una arquitectura de Big data considerando un diseño eficiente dentro de un entorno productivo sería:

- El procesamiento de datos que permite determinar en tiempo real la eficiencia de los procesos industriales.
- El uso de datos para el análisis financiero los cuales son generados en tiempo real a través del uso de procesos por lotes.
- Optimización de procesos que permiten tiempos de respuesta muy cortos especialmente en la reasignación de recursos o la optimización de rutas.

- A través de su operatividad general, permite el análisis de tendencias como también la generación de recomendaciones.

- La utilización de herramientas de gestión que permiten actualizar lo que se muestra de manera inmediata en base a los datos procesados.

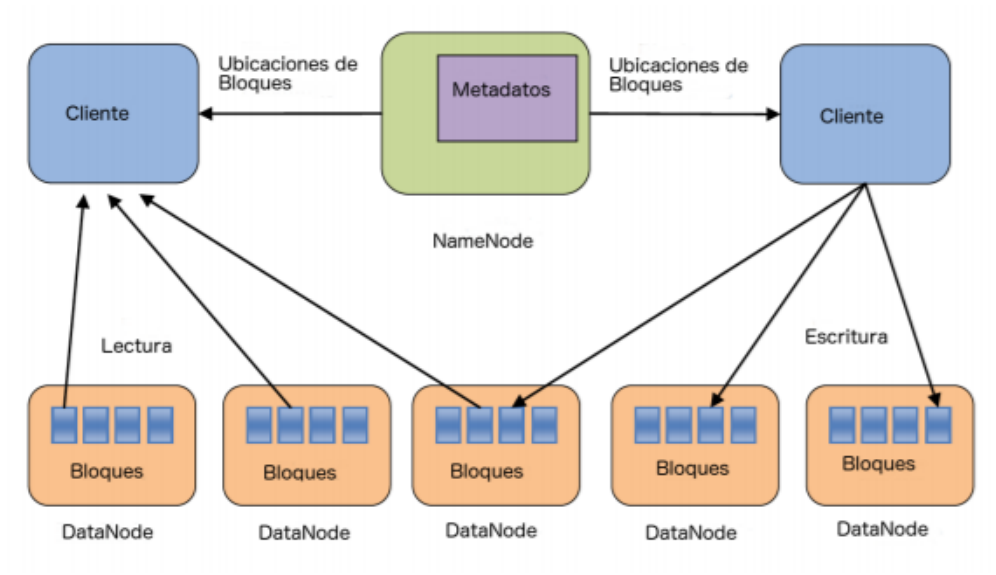
- Definir planes de mantenimiento predictivo basados en el uso de los dispositivos IoT analizando sus incidencias.

Entornos de procesamiento de Datos

Según Guzmán Ponce et al. (2018) plantea y analiza tres diferentes entornos de procesamiento de datos destacando el modelo de programación, los lenguajes de programación y el tipo de fuente de datos. Entre los que menciona a Apache Hadoop, Apache Spark y Apache Flink.

Apache Hadoop: Desde el 2008 esta tecnología viene siendo usada para la manipulación de datos masivos, es de código libre, desarrollado para ser utilizado de forma distribuida y escalable. Se divide en dos componentes: HDFS y el cómputo distribuido con la idea de MapReduce. HDFS este sistema de archivos almacena los archivos a lo largo del cluster, diseñado para obtener un acceso rápido para grandes archivos o conjuntos de datos grandes, es escalable y tolerante a fallas como se aprecia en la figura 2. Por otro lado, MapReduce es un algoritmo de cómputo distribuido, funciona para el procesamiento de grandes volúmenes de datos en paralelo, permitiendo realizar códigos de manera distribuida o paralela.

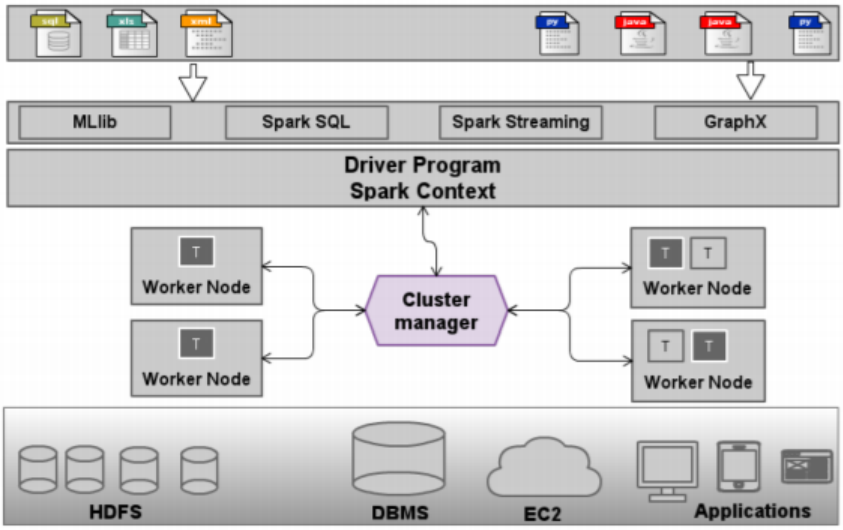
Figura 2 Arquitectura HDFS



Fuente: (Guzmán Ponce et al., 2018)

Apache Spark: Cuenta con diversas ventajas sobre el resto de los entornos de trabajo, hoy en día es utilizado por empresas como Yahoo, Baidu, entre otras. De igual manera que Hadoop, es de código libre, para procesos de manera distribuida, basado en el mejoramiento del rendimiento de memoria. En la figura 3 se observa la arquitectura que plantea Apache Spark.

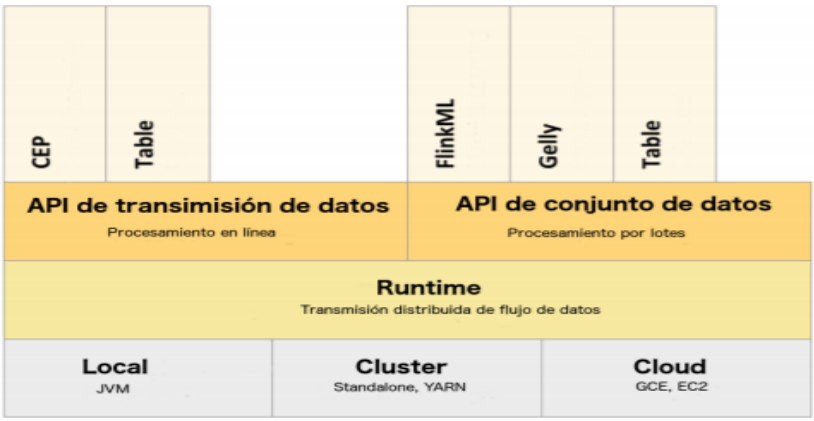
Figura 3 Arquitectura Spark



Fuente: (Guzmán Ponce et al., 2018)

Apache Flink: Flink es un entorno de trabajo de código libre, utilizado en la manipulación de datos tanto en tiempo real como por lotes, capaz de ser aprovechado por las características de ser distribuido, alto rendimiento, alta disponibilidad y preciso. Una ventaja competitiva es que las aplicaciones pueden mantener una agregación o resumen de los datos procesados, asegurando el estado de una aplicación en caso de falla. En la figura 4 se muestra la arquitectura Flink, donde integra el procesamiento en línea con el procesamiento por lotes.

Figura 4 Arquitectura Flink



Fuente: (Guzmán Ponce et al., 2018)

Aprendizaje Automático

Aprendizaje automático es muy frecuentemente utilizado en estos entornos de trabajo de la Big data, ya sean supervisados como por ejemplo el SVM o no supervisados como el KNN como lo indica Sandoval (2018). Por otro lado, algunos pueden servir tanto para clasificación como para regresión, y son:

- **Regresión logística**. Este método estadístico es muy útil aplicado a problemas de clasificación.
- **Árbol de decisión**. Sirven en la toma de decisiones y se pueden procesar grandes volúmenes de datos.
- **Random forest**. Es una combinación de árboles predictores mejorados, se puede utilizar en regresión y clasificación.
- **Perceptron multicapa**. Es considerada un Red neuronal artificial formado por múltiples capas, con la posibilidad de resolver problemas que no son linealmente separables.
- **Support Vector Machine (SVM)**. Permite encontrar la forma óptima de clasificar entre varias clases. Se puede aplicar tanto para clasificación como para regresión.
- **Naive Bayes**. Se basa en el uso de la técnica de clasificación llamada "Teorema de bayes", permite construir fácilmente modelos con un buen comportamiento y a la vez simples.
- **K-means**. Algoritmo de clasificación no supervisada que agrupa objetos en k grupos en base a sus características.
- **Latent Dirichlet allocation (LDA).** Técnica que intenta reducir las dimensiones del conjunto de características.

Por otro lado Cravero et al. (2020) en su artículo utilizan datos y tecnología basadas en las Arquitecturas de Big data enfocadas en el análisis del cambio climático. También, plantea para el sector salud una arquitectura denominada Big Data Lambda, enfocadas en aplicaciones en tiempo real y modelos escalables, también plantea otra arquitectura orientada al análisis del impacto del cambio climático dirigidos hacia la biodiversidad.

De manera que, en la dinámica del procesamiento de datos que se propone, se debe desarrollar un tránsito ordenado entre la comprensión de los datos y la construcción del conocimiento a partir de estos datos, enriqueciendo la fiabilidad de los resultados que se obtengan y su posterior aprendizaje.

III. METODOLOGÍA

3.1. Tipo y diseño de investigación

El tipo de estudio para esta investigación es transversal descriptiva de tipo mixta aplicada según Sampieri et al. (2014) es descriptiva porque se busca medir o recoger información de manera independiente sobre las variables en estudio, y transversal, ya que el fenómeno de estudio se analiza en un periodo corto de tiempo o en un punto exacto en el tiempo. Es por ello, que la función principal del estudio es identificar y describir el fenómeno a través del análisis de datos en un periodo concreto en el tiempo. Es aplicada dado que tiene como propósito dar solución a situaciones o problemas concretos e identificables.

El diseño de contrastación de hipótesis es cuasi experimental según Sampieri et al. (2014) éstos manipulan deliberadamente, al menos, una variable independiente para observar su efecto y relación con una o más variables dependientes. Se realizarán dos experimentos con los datos, el experimento 1 corresponderá a evaluar el tratamiento de los datos académicos de manera tradicional y el experimento 2 se realizará aplicando el modelo predictivo centrado en el uso de grandes volúmenes de datos.

3.2. Variables y Operacionalización

Las variables identificadas se describen a continuación:

VARIABLE INDEPENDIENTE:

Modelo predictivo

Definición conceptual: Es un conjunto de procesos ejercidos a través de técnicas computacionales de análisis de datos que ayudan a inferir la probabilidad de que ocurran determinadas situaciones previas a su consecución y, a su vez, detectar oportunidades de negocio.

VARIABLE DEPENDIENTE:

Tratamiento de datos en la Big data

Definición Conceptual: Operación o conjunto de operaciones, aplicadas a los datos mediante los cuales se obtiene, usa, registra, organiza, conserva, elabora, modifica o consulta.

3.3. Población y Muestra

La población está definida por los datos académicos de todos los estudiantes de la universidad en estudio a partir del año 2005. La base de datos consta de un alto volumen de datos conteniendo información académica de la Universidad. La cantidad de registros académicos es de 2’425,966 registros, por lo que este total supera los 2 millones de registros académicos, éstos están relacionados con las matrículas de los estudiantes en los diferentes semestres académicos y sus detalles de matrícula en donde se registran las asignaturas seleccionadas por el estudiante, al final de cada semestre se registra su nota final.

Tabla 3 Total de registros académicos desde el 2005 al 2020

Semestre	Total registros Académicos
2005-I	77083
2005-II	69797

Semestre	Total registros Académicos
2006-I	77364
2006-II	70261
2007-N	4753
2007-I	75703
2007-II	68116
2008-I	75401
2008-II	70202
2009-I	77719
2009-II	71554
2010-I	78110
2010-II	71032
2011-I	78741
2011-II	10273
2011-N	74093
2012-I	81918
2012-II	10457
2012-N	77961
2013-I	85099
2013-II	9612
2013-N	76673
2014-I	82970
2014-II	76322
2015-I	85397
2015-II	78804
2016-N	5770
2016-I	82666
2016-II	76174
2017-N	4775
2017-I	83394

Semestre	Total registros Académicos
2017-II	75333
2018-N	5417
2018-I	80684
2018-II	71363
2019-N	6192
2019-I	79045
2019-II	67243
2020-N	4358
2020-I	68137

La muestra a ser tratada corresponde a los ciclos académicos regulares del 2016-I al 2020-I, siendo un total de 684 039 registros académicos, como se detalla en la siguiente tabla.

Tabla 4 Muestra de datos académicos de los ciclos 2016-I al 2020-I

Semestre	Total registros Académicos
2016-I	82666
2016-II	76174
2017-I	83394
2017-II	75333
2018-I	80684
2018-II	71363
2019-I	79045
2019-II	67243
2020-I	68137

3.4. Técnicas e instrumentos de recolección de datos.

Las técnicas a utilizar en la presente investigación se describen a continuación:

- La observación: Con frecuencia se usa está técnica para profundizar en el conocimiento del comportamiento de exploración. Es el registro visual de lo que ocurre en una situación real, clasificado y consignando los datos de acuerdo con algún esquema previsto y de acuerdo al problema que se estudia. El instrumento es una guía de observación.

- Análisis Documental: Técnica que permite recolectar datos de libros, boletines, periódicos, revistas, bases de datos científicas y artículos científicos, consideradas fuentes secundarias para extraer datos sobre las variables de estudio, se utiliza frecuentemente como instrumento el uso de la ficha de registro de datos.

- Encuesta: Técnica de recolección de datos mediante la aplicación de un cuestionario establecido con anterioridad. Esta encuesta nos permitirá evaluar el estado actual del procesamiento de datos académicos en la institución.

- Entrevista: Técnica en la que se conversa o intercambia ideas entre dos partes, con el fin de obtener información de valor. Tiene como objetivo evaluar el proceso actual del procesamiento de datos académicos en la institución y está dirigida al director de servicios académicos y al jefe de la oficina de tecnologías de la información.

3.5. Procedimiento

Se utilizará Excel y SPSS como herramientas para el procesamiento y análisis de los datos recolectados a través de los instrumentos, luego se presentará los

resultados a través de tablas y gráficos estadísticos. Se utilizarán las fórmulas de la estadística descriptiva en la presente investigación.

3.6. Aspectos éticos

En esta investigación se toman los siguientes aspectos éticos:

- La Confidencialidad, que es parte del secreto profesional que se debe mantener y permite ser celoso con la investigación, evitando divulgar la información de datos personales que se obtienen producto de esta investigación.
- La Objetividad, mediante la cual se pretende dar a conocer aspectos con veracidad en relación a los estudios realizados en el tratamiento de datos en la Big data dejando a un lado la subjetividad del investigador.
- La Originalidad, mediante la cual se citarán las fuentes bibliográficas de la información mostrada en la presente investigación, a fin de demostrar la inexistencia de plagio intelectual.

IV. ANÁLISIS SITUACIONAL

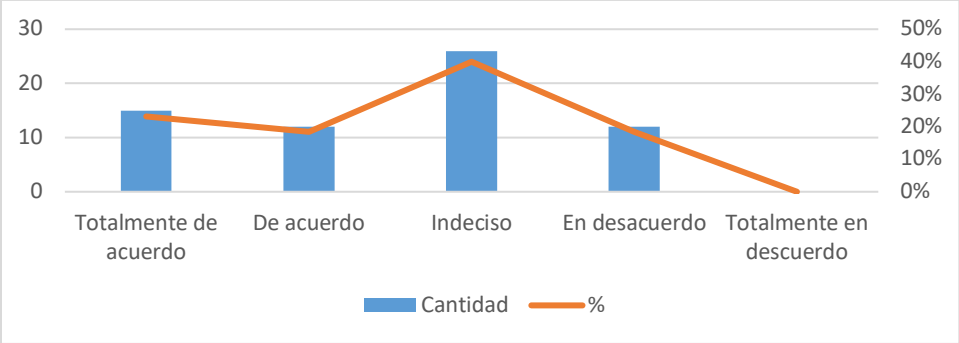
Se esta etapa se aplicó una encuesta al personal de la institución relacionada con los procesos académicos con el objetivo de diagnosticar el estado actual de la dinámica del procesamiento de datos del área académica de esta institución. La aplicación de la encuesta se realizó a través de los correos electrónicos institucionales, ésta fue aplicada entre los meses de setiembre y octubre del presente año.

Dicha encuesta se aplicó a una muestra de 65 personas y luego de procesar la misma se obtuvo los siguientes resultados en cada dimensión consultada.

DIMENSIÓN: RECOLECCIÓN

1. ¿Considera que los datos académicos que se registran son suficientes para la toma de decisiones?

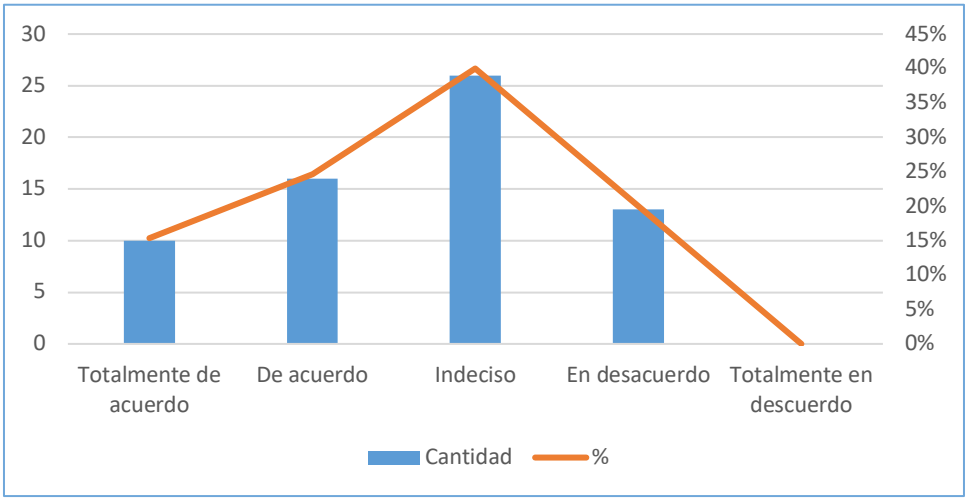
Figura 5 ¿Considera que los datos académicos que se registran son suficientes para la toma de decisiones?



Se observa que un 42% está totalmente de acuerdo o de acuerdo en que los datos académicos que son registrados son suficientes para la toma de decisiones y un 40% que está indeciso en si son suficientes para la toma de decisiones.

2. ¿Considera que los datos académicos obtenidos en los procesos de su área son claros?

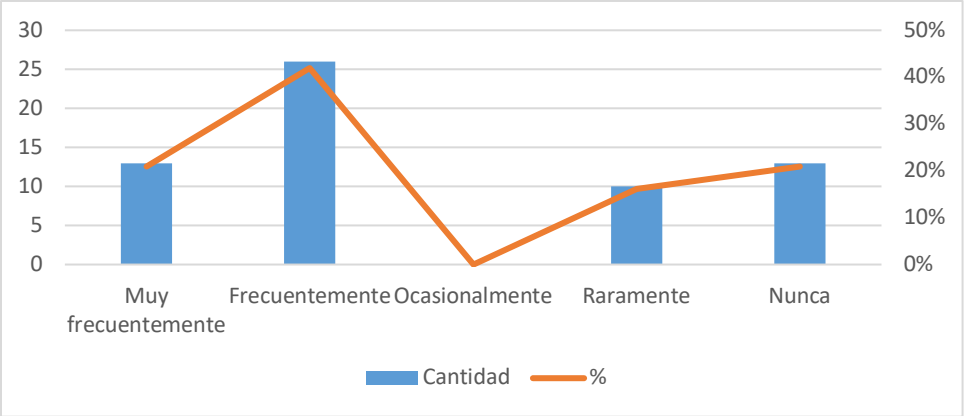
Figura 6 ¿Considera que los datos académicos obtenidos en los procesos de su área son claros?



Se puede apreciar que un 40% está totalmente de acuerdo o de acuerdo en que los datos académicos obtenidos son claros para los procesos de su área, mientras que otro 40% no está ni de acuerdo ni en desacuerdo, pero si hay un 20% que está en desacuerdo en que estos datos sean claros.

3. ¿Con que frecuencia se recopilan los datos académicos?

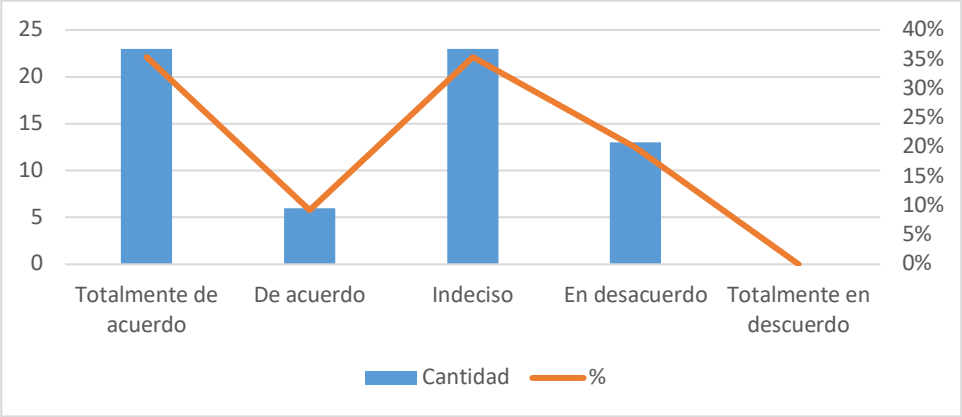
Figura 7 ¿Con que frecuencia se recopilan los datos académicos?



Se evidencia que muy frecuentemente o frecuentemente se recopilan los datos académicos en un 63%, mientras que un 37% manifiesta que es raramente o nunca que se recopilan datos académicos.

4. ¿Considera que los sistemas académicos son intuitivos y fáciles de manipular?

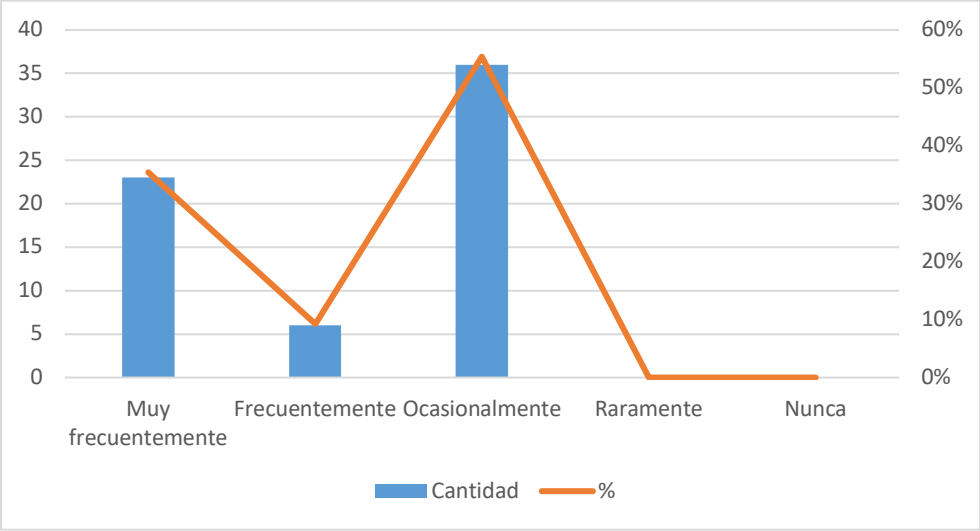
Figura 8 ¿Considera que los sistemas académicos son intuitivos y fáciles de manipular?



Se aprecia que un 45% está totalmente de acuerdo o de acuerdo en que los sistemas académicos son intuitivos y fáciles de manipular, pero un 55% está indeciso y en desacuerdo en que los sistemas sean fáciles e intuitivos.

5. ¿Considera que los datos que se registran en los sistemas académicos se validan?

Figura 9 ¿Considera que los datos que se registran en los sistemas académicos se validan?

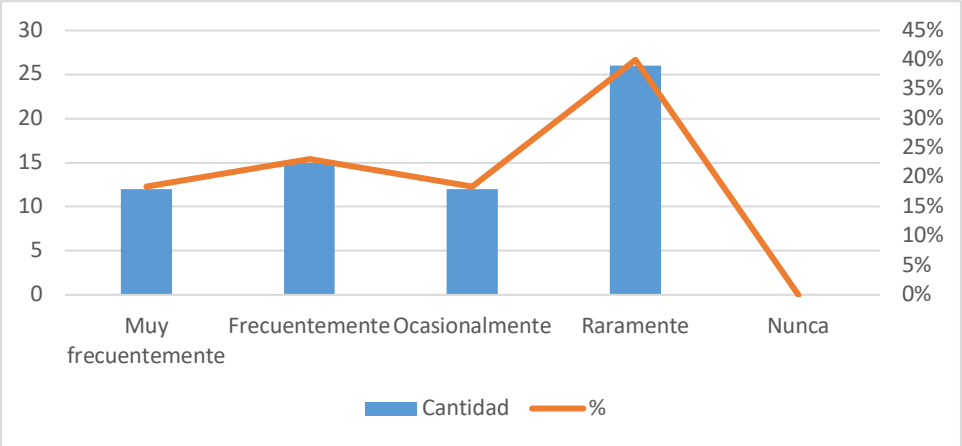


Se observa que un 45% considera que los datos que se registran en los sistemas académicos se validan, mientras que un 55% considera que ocasionalmente los datos se validan al momento de ser registrados en los sistemas académicos.

DIMENSIÓN: MANIPULACIÓN

6. ¿Qué tan frecuente se encuentran los datos disponibles para cuando usted lo necesita?

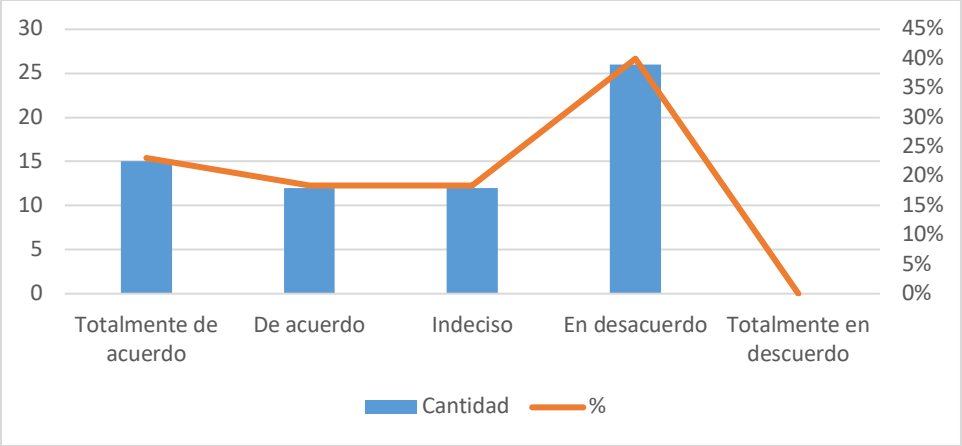
Figura 10 ¿Qué tan frecuente se encuentran los datos disponibles para cuando usted lo necesita?



Se aprecia que un 40% considera que los datos raramente se encuentran disponibles cuando se necesitan y un 42% considera que muy frecuentemente o frecuentemente los datos se encuentran disponibles si son necesitados.

7. ¿Los reportes académicos obtenidos le permite un análisis completo para los requerimientos de su oficina?

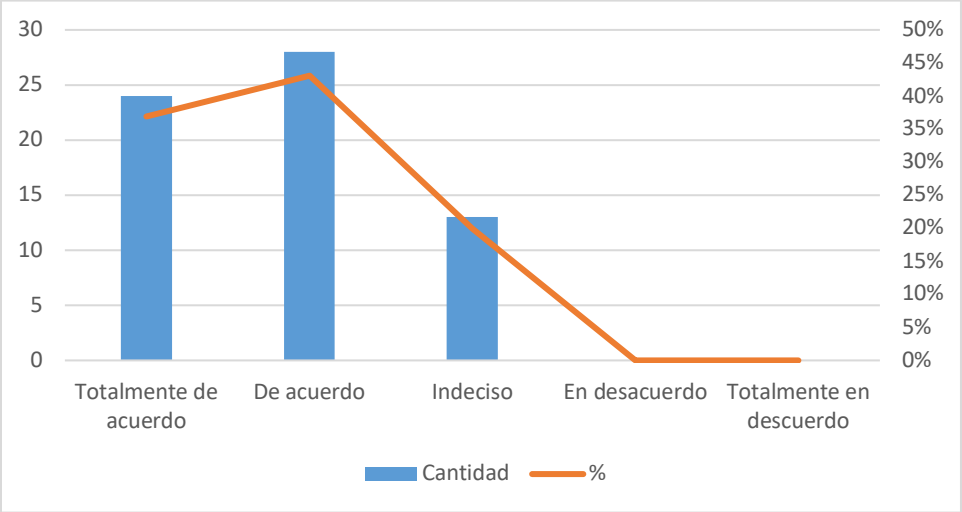
Figura 11 ¿Los reportes académicos obtenidos le permite un análisis completo para los requerimientos de su oficina?



Se observa que un 42% considera que los reportes académicos obtenidos les permiten un análisis completo para los requerimientos de su oficina, mientras que un 40% está en desacuerdo y un 18% se considera indeciso.

8. Cree usted que la falta de herramientas tecnológicas que posea la institución para la extracción y procesamiento de grandes volúmenes de datos influye notablemente en su manipulación.

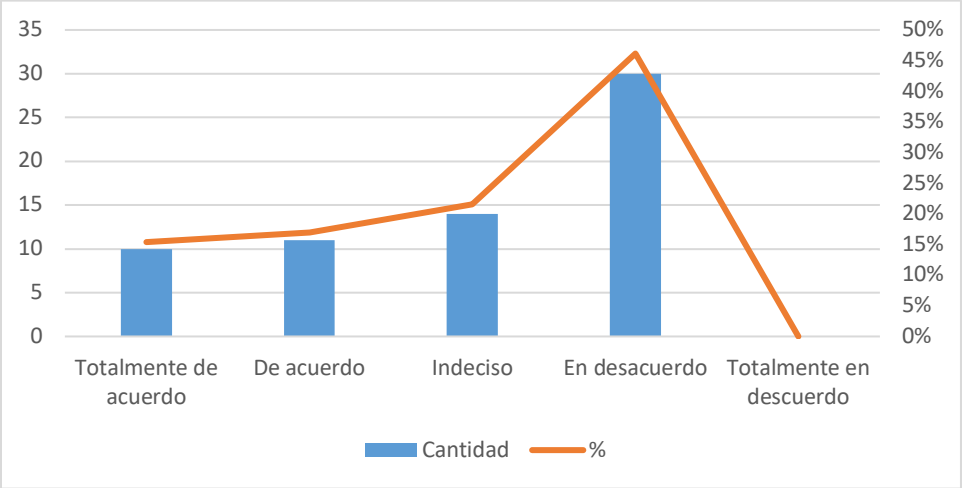
Figura 12 Cree usted que la falta de herramientas tecnológicas que posea la institución para la extracción y procesamiento de grandes volúmenes de datos influye notablemente en su manipulación.



Se evidencia que un 80% de los encuestados cree que la falta de herramientas tecnológicas que posea la institución para la extracción y procesamiento de grandes volúmenes de datos influye notablemente en su manipulación.

9. Los datos académicos que tiene almacenada la institución son de fácil acceso a los usuarios que lo requieran.

Figura 13 Los datos académicos que tiene almacenada la institución son de fácil acceso a los usuarios que lo requieran.

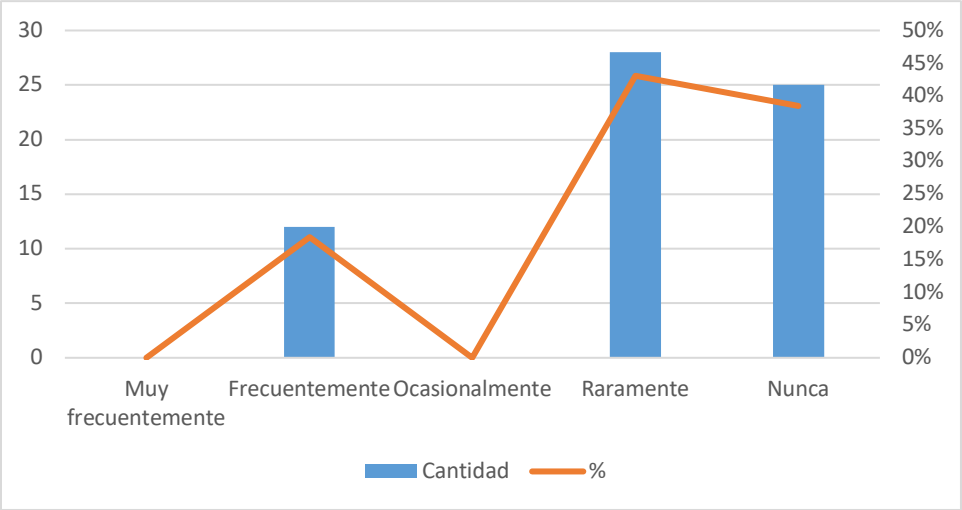


Se puede apreciar que un 46% están en desacuerdo con que los datos académicos que tiene almacenada la institución son de fácil acceso a los usuarios que lo requieran, mientras que solo un 32% manifiesta que están totalmente de acuerdo o de acuerdo en que si son de fácil acceso.

DIMENSIÓN: CALIDAD

10. ¿Qué tan frecuente se detectan datos erróneos en el procesamiento de la información?

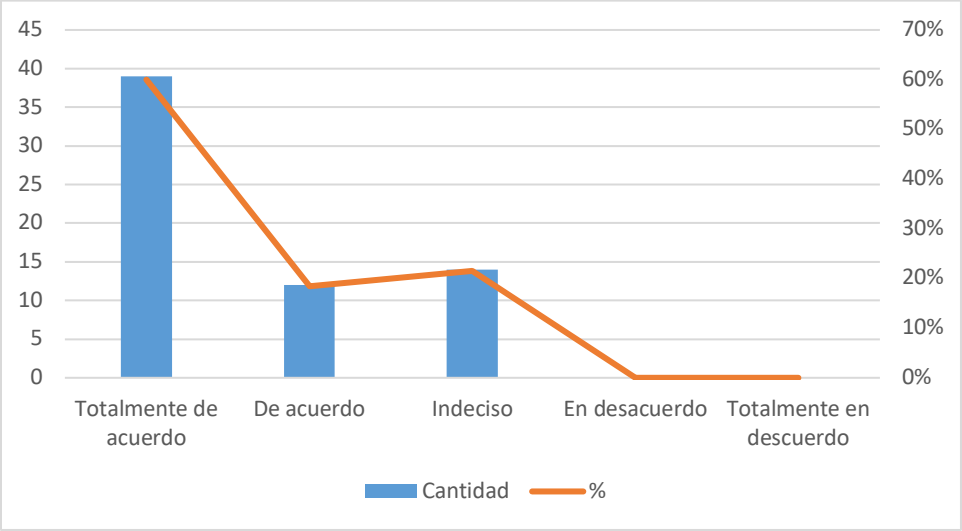
Figura 14 ¿Qué tan frecuente se detectan datos erróneos en el procesamiento de la información?



Se puede apreciar que un 82% indica que raramente o nunca se detectan datos erróneos en el procesamiento de información, mientras que un 18% indica que frecuentemente si se detectan datos erróneos.

11. ¿Está de acuerdo en que la institución aplique algún estándar de calidad para el procesamiento de datos?

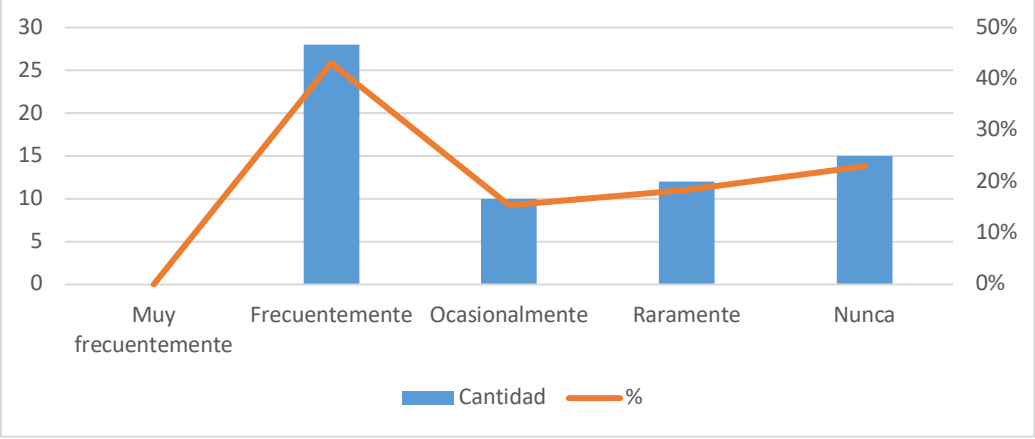
Figura 15 ¿Está de acuerdo en que la institución aplique algún estándar de calidad para el procesamiento de datos?



Se observa que un 78% considera totalmente de acuerdo o de acuerdo que se aplique algún estándar de calidad para el procesamiento de datos en la institución.

12. La institución captura los datos académicos centrado en las necesidades organizacionales.

Figura 16 La institución captura los datos académicos centrado en las necesidades organizacionales

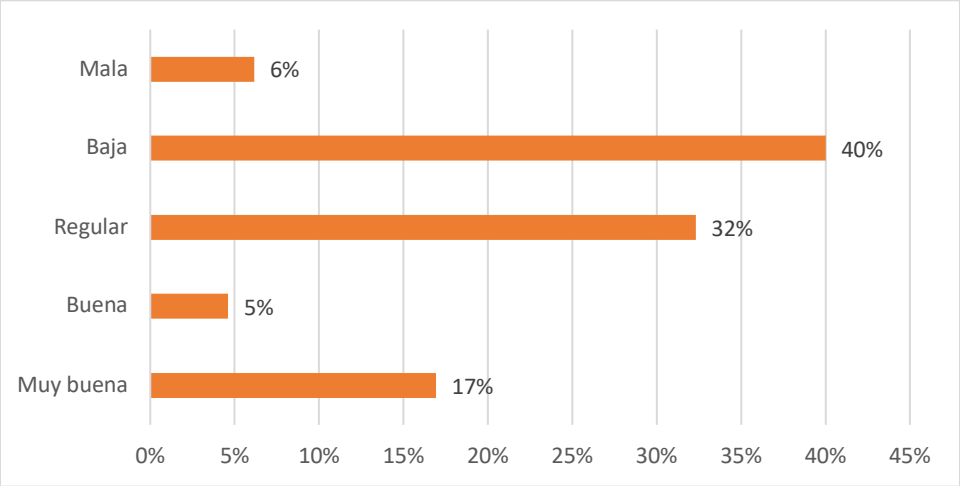


Se puede evidenciar que un 43% considera que frecuentemente la institución captura los datos académicos centrado en las necesidades organizacionales, mientras que un 34% ocasionalmente o raramente se centran en las necesidades organizacionales, y finalmente un 23% indican que nunca se centran en las necesidades organizacionales.

DIMENSIÓN: RENDIMIENTO

13. ¿El tiempo de respuesta de las aplicaciones al solicitar datos es?

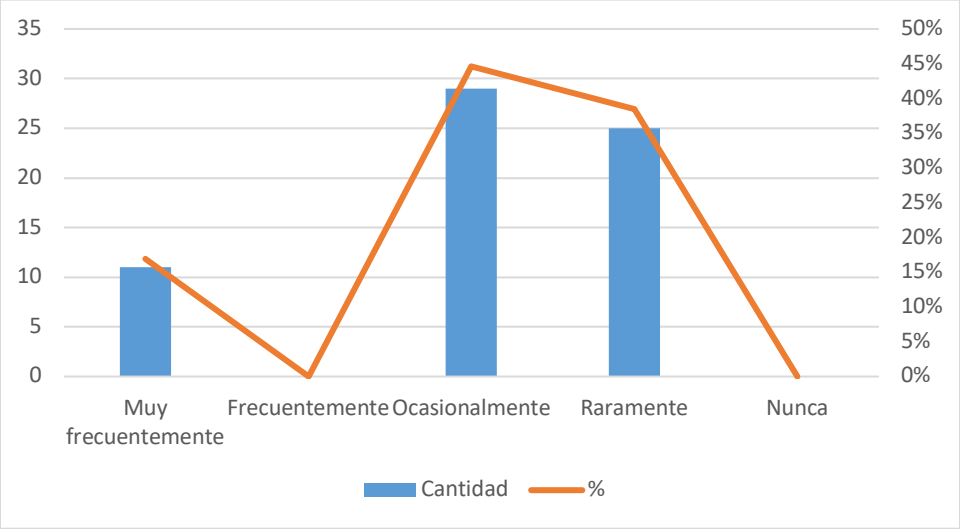
Figura 17 ¿El tiempo de respuesta de las aplicaciones al solicitar datos es?



Se observa que un 46% considera que el tiempo de respuesta de las aplicaciones al solicitar datos es mala o baja, mientras que un 32% lo considera como un tiempo de respuesta regular y un 22% lo considera como buena o muy buena.

14. ¿Los datos académicos se obtienen en tiempo real?

Figura 18 ¿Los datos académicos se obtienen en tiempo real?

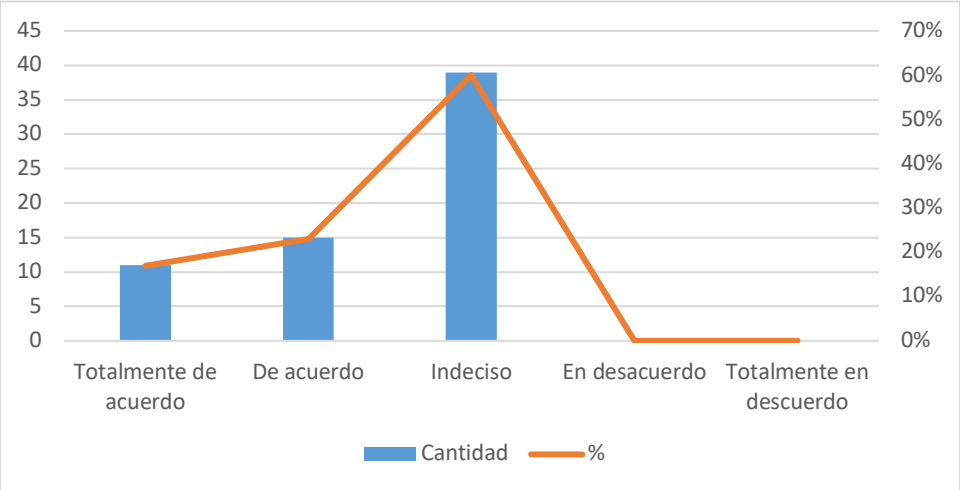


Se observa que un 45% considera que ocasionalmente los datos académicos se obtienen en tiempo real, mientras que un 38% considera que raramente se obtienen en tiempo real.

DIMENSIÓN: SEGURIDAD

15. ¿Considera usted que los datos académicos están totalmente seguros?

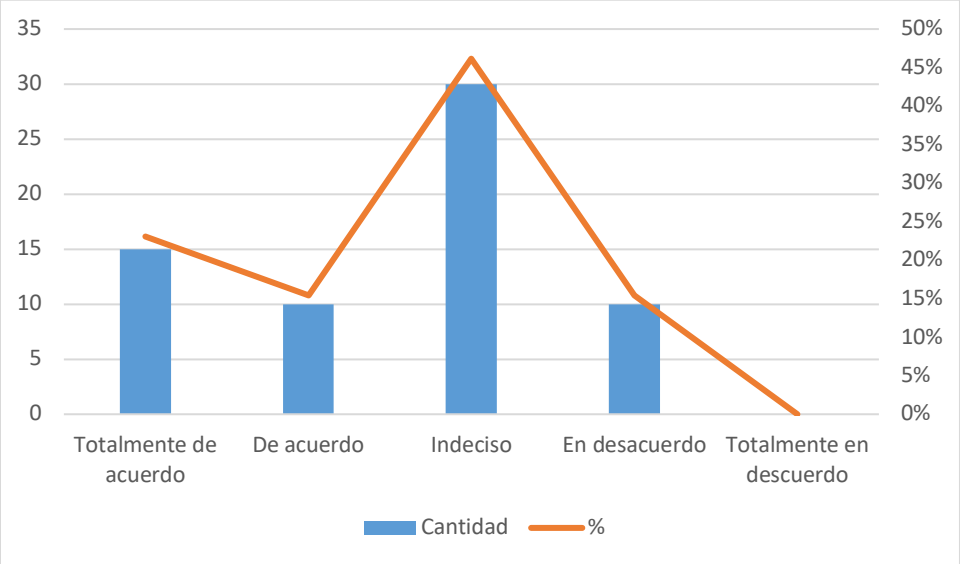
Figura 19 ¿Considera usted que los datos académicos están totalmente seguros?



Se aprecia que solo un 17% considera totalmente de acuerdo en que los datos académicos están totalmente seguros, mientras que un 23% también lo considera de acuerdo, mientras que un 60% no tiene clara una decisión con respecto a la seguridad de los datos.

16. ¿Cree usted que los datos académicos son accesibles sólo por personal autorizado?

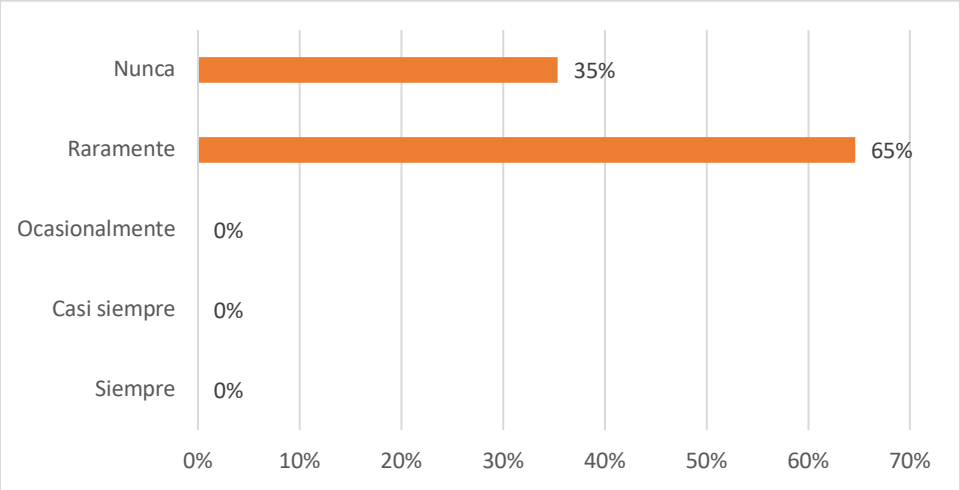
Figura 20 ¿Cree usted que los datos académicos son accesibles sólo por personal autorizado?



Se observa que un 38% considera que está totalmente de acuerdo o de acuerdo en que los datos académicos son accesibles sólo por personal autorizado, mientras que un 15% no son solo accesible por personal autorizado, habiendo un 46% que ni están de acuerdo ni en desacuerdo.

17. ¿La institución periódicamente le solicita que actualice claves para acceder a los sistemas académicos?

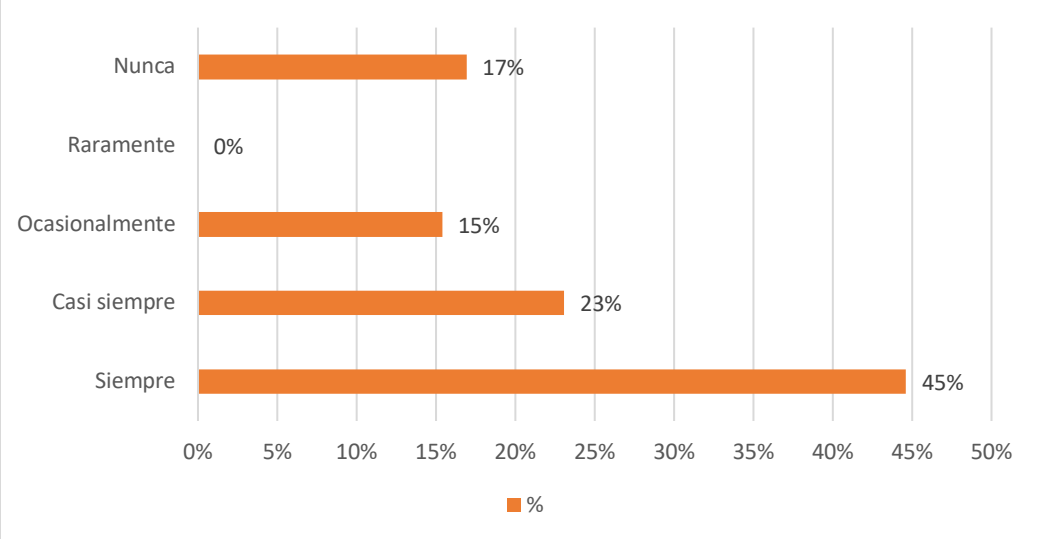
Figura 21 ¿La institución periódicamente le solicita que actualice claves para acceder a los sistemas académicos?



Se observa que un 65% considera que raramente la institución solicita periódicamente que actualicen claves para acceder a los sistemas académicos, mientras que un 35% indica que nunca le solicita actualizar claves en los sistemas académicos.

18. Los datos académicos son solo modificados mediante autorización.

Figura 22 Los datos académicos son solo modificados mediante autorización

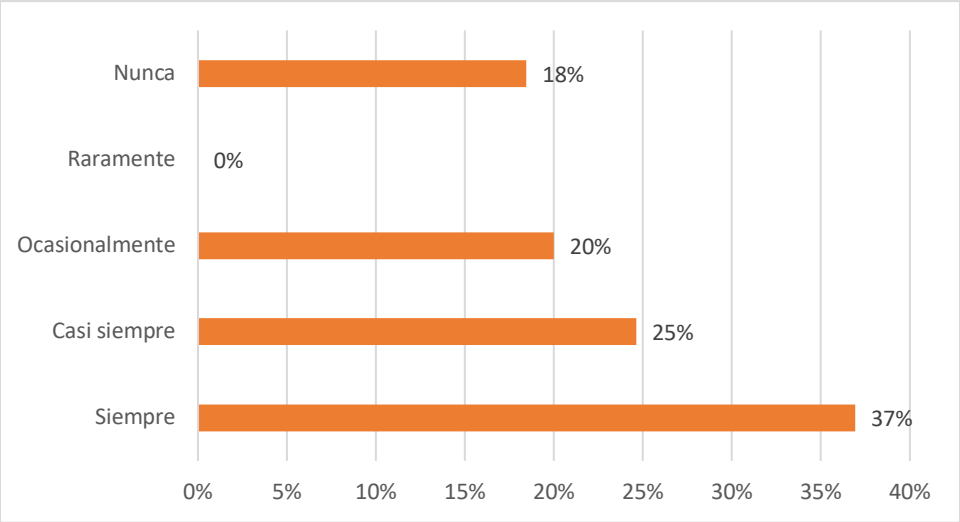


Se aprecia que un 45% considera que siempre los datos académicos son solo modificados mediante autorización y un 23% casi siempre se modifican con autorización.

DIMENSION: SOPORTE

19. Ante un requerimiento nuevo, la atención de la OTI es rápida.

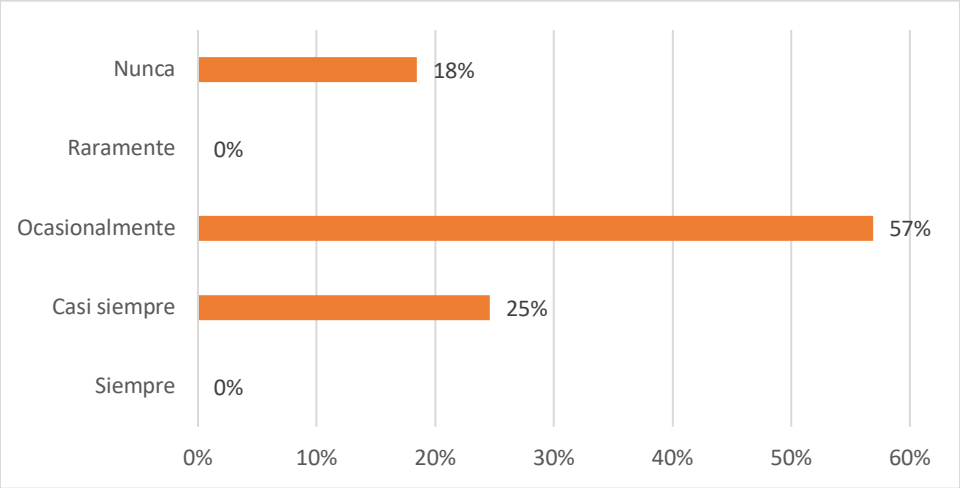
Figura 23 Ante un requerimiento nuevo, la atención de la OTI es rápida.



Se observa que un 62% consideran que siempre o casi siempre ante un requerimiento nuevo, la atención de la Oficina de Tecnologías de Información (OTI) es rápida, mientras que un 18% indica que nunca la oficina responde rápidamente ante un requerimiento nuevo.

20. ¿Ante un incidente relacionado con los sistemas académicos el tiempo de atención es rápido?

Figura 24 ¿Ante un incidente relacionado con los sistemas académicos el tiempo de atención es rápido?



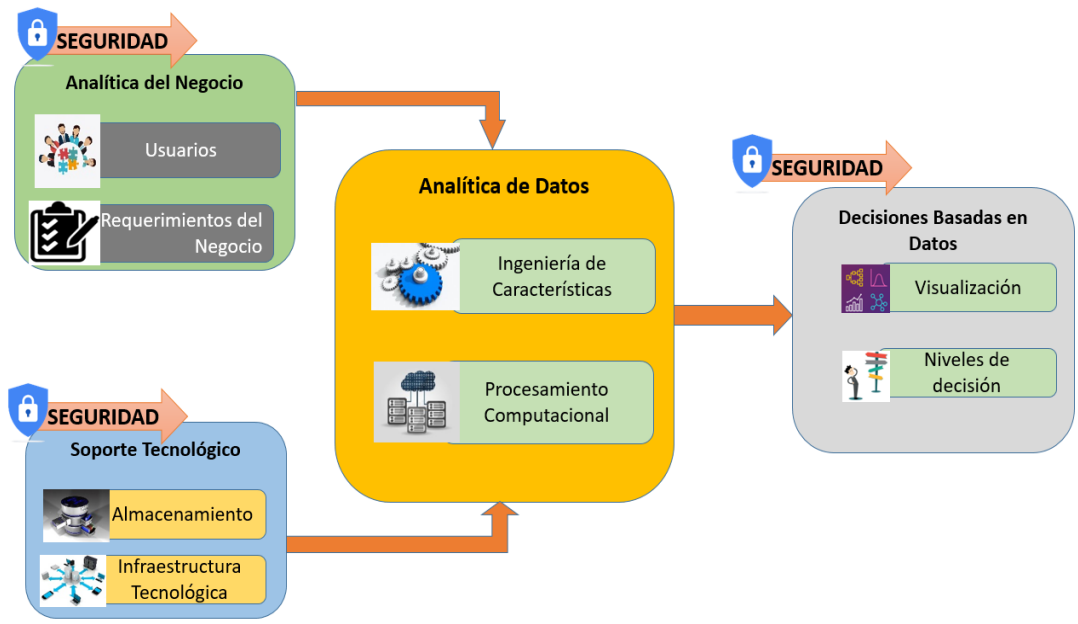
Se aprecia que un 25% casi siempre el tiempo de atención es rápido ante un incidente relacionado con los sistemas académicos, mientras que un 18% indica que nunca es rápida la atención ante un incidente y un 57% considera que ocasionalmente el tiempo de atención para atender un incidente es rápido.

V. MODELO PREDICTIVO PARA EL PROCESAMIENTO DE DATOS ACADÉMICOS EN BIG DATA EN LA EDUCACIÓN SUPERIOR

El Modelo predictivo aplicado a la educación superior, toma en cuenta los elementos del procesamiento de datos, los enfoques arquitectónicos presentes, así como también los entornos de trabajo que permiten orientar el desarrollo del procesamiento, además, tomando en consideración la parte tecnológica, la parte analítica tanto del negocio como de los datos que se requieren y la generación de decisiones producto de los resultados obtenidos del procesamiento que permitan mejorar los procesos académicos de la institución.

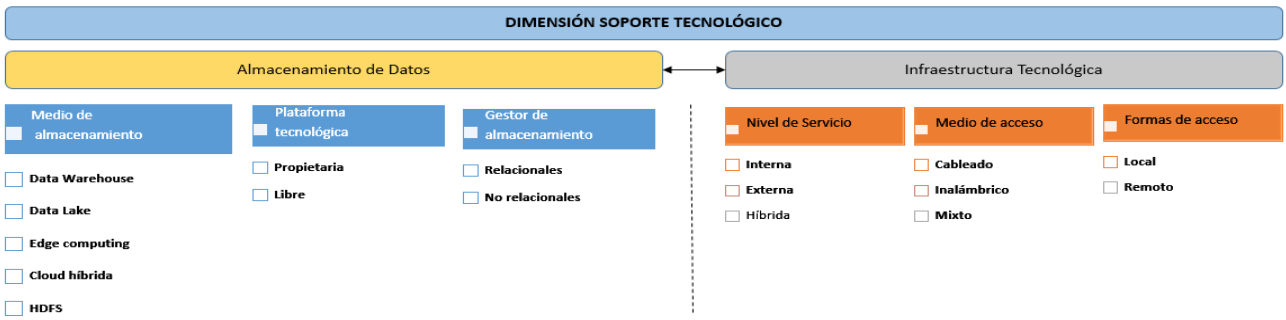
Para la elaboración del modelo propuesto como se muestra en la figura 25, se han considerado **4 dimensiones**: La dimensión Soporte tecnológico, la dimensión Analítica del Negocio, la dimensión Analítica de Datos y la dimensión de Decisiones basadas en datos.

Figura 25 Modelo predictivo de procesamiento de datos académicos en la Big data



La **dimensión Soporte Tecnológico**, contempla el almacenamiento de los datos como también la infraestructura tecnológica que requiera la institución, ambos necesarios para brindar el soporte que permita el procesamiento de grandes volúmenes de información.

Figura 26 Dimensión Soporte Tecnológico



En cuanto al almacenamiento de datos, es la recolección de la información mediante el uso de tecnología desarrollada especialmente para guardar esos datos y mantenerlo lo más accesible posible. A continuación, se detalla cómo se han clasificado para su adecuada gestión y acceso confiable:

Basado en el medio de almacenamiento:

- **Data Warehouse**: Los datos se extraen de sistemas transaccionales.
- **Data Lake**: Guarda los datos independientemente de la fuente y su estructura por eso los datos se mantienen en su forma sin procesar.
- **Edge computing**: El almacenamiento y el procesamiento de la información se lleve a cabo cerca del punto de recolección, permitiendo evitar sobrecargas en la nube.
- **Cloud híbrida**: Se basa en aprovechar las ventajas de combinar el uso de una nube pública y una nube privada, con una configuración a medida para los miembros de la organización.
- **HDFS**: Potente sistema de almacenamiento distribuido de archivos conocido como Hadoop Distributed File System.

Basado en la Plataforma tecnológica, podemos clasificarlos en:

- **Propietaria**: Son plataformas en las que existe un costo económico por la instalación y el mantenimiento.
- **Libre**: Son plataformas en las que la instalación y el mantenimiento no implican un costo económico.

Basado en el Gestor de almacenamiento se puede clasificar en:

- **Relacionales**: Se trabaja con un conjunto de tablas que contienen filas que corresponden a los registros y las columnas que contienen las características almacenadas de los objetos.
- **No relacionales**: La característica de estos gestores es que no utilizan el lenguaje SQL para la definición de consultas, por lo que no utiliza el esquema tabular de filas y columnas como el modelo relacional.

En cuanto a la seguridad en el almacenamiento de los datos, es la protección que se aplica para evitar el acceso no autorizado a los datos, protegiendo también de una posible corrupción de los mismos.

Se puede considerar las siguientes estrategias de seguridad en el almacenamiento:

- **Copias de respaldo**: Considerándose la copia de un objeto y que se pueda conservar en un lugar seguro, se le puede considerar como: Diario, Mensual / Trimestral.
- **Controles de acceso**: Mecanismo que restringe la entrada a un objeto del sistema y puede ser: Alta, Media y Baja.

En cuanto a la infraestructura tecnológica, esta se centra a nivel de los servicios o roles que brindan los servidores para dar soporte al modelo propuesto y los medios de acceder a éstos. Y se puede clasificar en base a los tipos de servicios, medios y formas de acceso.

A nivel de tipos de servicios:

- **Interna**: Corresponde a servicios a nivel de servidores ubicados dentro de la institución.
- **Externa**: Corresponde a servicios alojados en la nube (Cloud).
- **Hibrida**: Corresponde a servicios que integran ambos, tanto interna como externa.

A nivel de medios de acceso al sistema analítico:

- **Cableado**: Utilizando como medio físico par trenzado o fibra óptica principalmente.
- **Inalámbrico**: A través de frecuencia de radio como principal medio utilizado actualmente.
- **Mixto**: Integra ambos tipos de acceso.

A nivel de la forma de acceso:

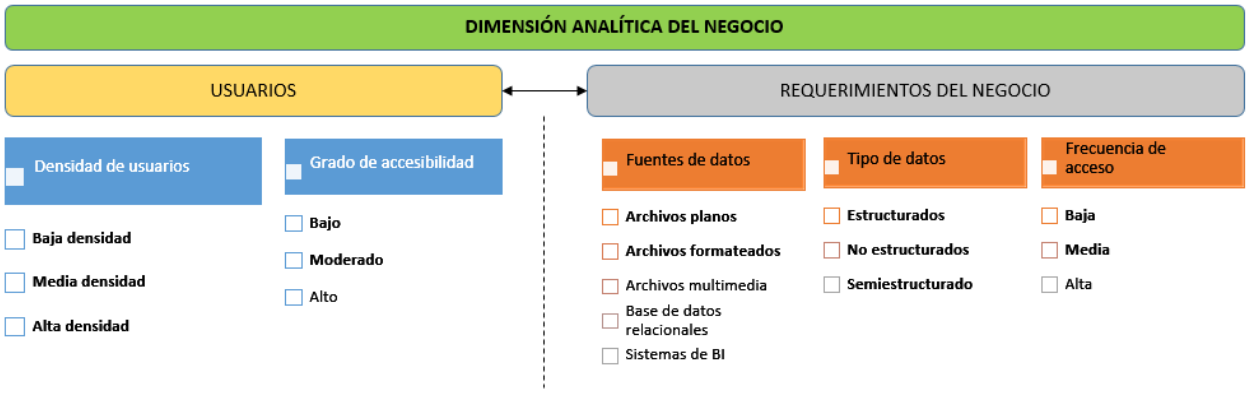
- **Local**: Acceso a los datos a través de los equipos dentro de la misma institución.
- **Remoto**: acceso a los datos desde ubicaciones remotas (por ejemplo, el trabajo remoto)

En cuanto al componente de la Seguridad en infraestructura tecnológica, debemos tener en cuenta las siguientes estrategias, cuyo propósito es establecer los mecanismos orientados a proteger físicamente cualquier recurso del sistema, en los 4 niveles establecidos a continuación:

- **Seguridad del perímetro**: A través de Firewall e IDS (Sistemas de Detección de Intrusos)
- **Seguridad de las instalaciones**: La vigilancia en los interiores, sistemas de identificación y métodos de verificación.
- **Seguridad de la sala de ordenadores**: Restringir el acceso a través de múltiples métodos de verificación, monitorear los accesos autorizados y contar con redundancia energética y de comunicaciones.
- **Seguridad a nivel de racks**: Por ejemplo, a través de bloqueo electrónico para racks de servidores, sistemas biométricos para acceso a los racks y video vigilancia IP.

En la **dimensión Analítica del Negocio**, contempla tanto a los usuarios como los requerimientos del negocio, enfocándose en las necesidades y en quienes las necesitan.

Figura 27 Dimensión Analítica del Negocio



En cuanto a los Usuarios, entendiéndose como la persona que interactúa con la definición de los requerimientos de análisis para el procesamiento de datos basado en la extracción de características relevantes del problema a tratar, se logra identificar 3 perfiles:

- **Administrador**: Es aquel que tiene la responsabilidad de administrar el soporte tecnológico permitiendo una alta disponibilidad y calidad del servicio.
- **Académico**: Involucra a los jefes de departamento, directores de escuela y director de asuntos académicos de la institución, siendo los que toman decisiones ejecutivas y a nivel de las oficinas de Alta Dirección para la toma de decisiones estratégicas.
- **Docente**: Persona que se dedica a la enseñanza en una determinada área de conocimiento y aquel que monitorea resultados del proceso de enseñanza.

Además, se ha considerado la siguiente clasificación de los Usuarios:

- **Densidad de usuarios**: Cantidad de usuarios que harán uso del servicio. Se ha clasificado en:
    - **Baja densidad**: Menos de 400 usuarios conectados.
    - **Media densidad**: Entre 400 a 800 usuarios conectados.
    - **Alta densidad**: Más de 800 usuarios conectados.

- **Grado de accesibilidad**: Define los privilegios que los usuarios tendrán con respecto a los datos almacenados. Considerándose los siguientes grados:
    - **Bajo**: El usuario solo tendrá acceso a reportes del sistema.
    - **Moderado**: El usuario solo podrá especificar sus requerimientos, procesar modelos y acceder a los reportes.
    - **Alto**: El usuario podrá especificar requerimientos, procesar/crear modelos y generar reportes.

Por otra parte, los Requerimientos del Negocio, permitirá definir claramente lo que se requiere evaluar centrándose en las tareas que determinan las necesidades o condiciones que satisfacen el producto final. Además de identificar las diversas fuentes que serán requeridas para la extracción de datos necesarios y la selección de las variables de estudio.

En este eje se puede apreciar tres aspectos claves:

- **Fuentes de datos**: Son aquellas que nos proveen información relevante en cuanto a la investigación a realizar. Clasificándose en:
    - **Archivos planos**: Está compuesto por archivos donde solo lo que se almacena son textos.
    - **Archivos formateados**: Estos archivos tiene un formato específico como por ejemplo los archivos de Excel.
    - **Archivos multimedia**: Estos archivos pueden contener imágenes, sonido, videos, animaciones, etc. y pueden ser de gran tamaño.
    - **Bases de datos relacionales**: Utilizan el modelo relacional basado en tablas con filas y columnas además de tener relaciones entre dichas tablas.
    - **Sistemas de Business Intelligence (BI):** Sistemas que utilizan metodologías, aplicaciones, prácticas y capacidades para la creación y administración de datos, información y conocimiento, que permiten a los gestores y usuarios tomar mejores decisiones.

Figura 28 Archivos según fuente de datos

- **Tipos de datos:**
  - **Estructurados**: Son los que tradicionalmente se han utilizado en el tratamiento de datos, siendo sus características principales que se pueden almacenar en tablas y tienen una clara definición de longitud y formato.
  - **No Estructurado**: Se trata de datos en su forma original, tal y como fueron recogidos: No poseen un formato específico que permita almacenarlos de forma tradicional, pues no se puede desglosar la información que facilitan a tipos de datos definidos en longitud y formato, estos datos plantean múltiples desafíos para el procesamiento.
  - **Semiestructurados**: Siguen una especie de estructura, pero esta no es lo suficientemente regular como para gestionarla como datos estructurados. Posee ciertos patrones comunes que los describen y dan información sobre sus relaciones entre los mismos.

Figura 29 Clasificación según tipo de dato

- **Frecuencia de acceso**: Tiempo con el que se realizan solicitudes de datos y accesos al sistema. Se clasifican en:
    - **Baja**: 4 accesos al mes
    - **Media**: 12 accesos al mes
    - **Alta**: 20 accesos al mes

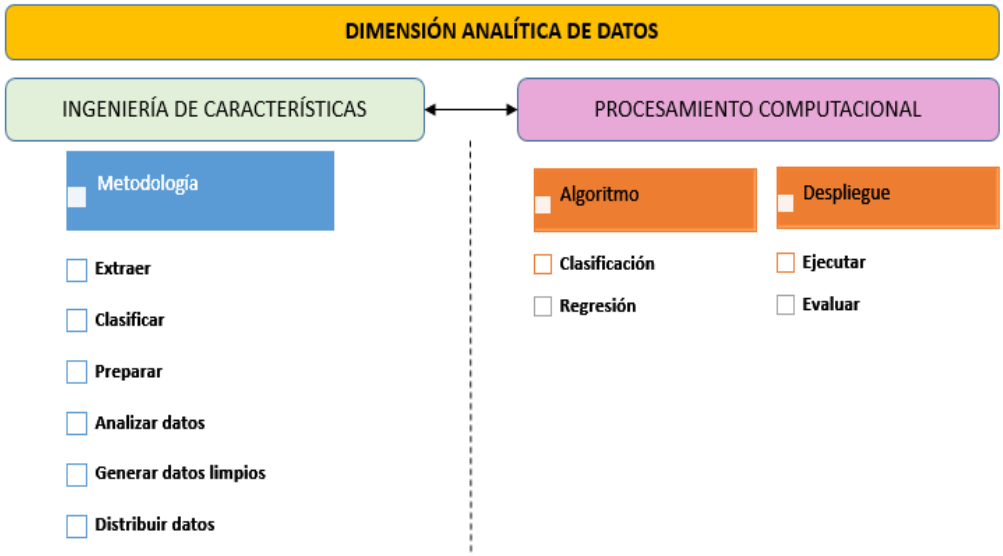
Considerando la seguridad en la dimensión Analítica del negocio, se debe tomar en cuenta las siguientes estrategias:

- El uso de roles o perfiles para así restringir el acceso al sistema.
- Monitoreo de la calidad de los datos.
- Encriptación y cifrado de los datos.

La **dimensión Analítica de datos**, se centra en dos ejes principales, el primero la ingeniería de requerimientos, el cual tiene como propósito la extracción de los requerimientos necesarios para abordar el problema planteado, proporcionando mecanismos para entender lo que el cliente desea, analizar las necesidades, evaluar la factibilidad de las mismas planteando una solución final y el segundo eje corresponde al procesamiento computacional en donde se debe seleccionar el

algoritmo o los algoritmos que permitirán evaluar los datos previamente obtenidos y posteriormente su despliegue funcional.

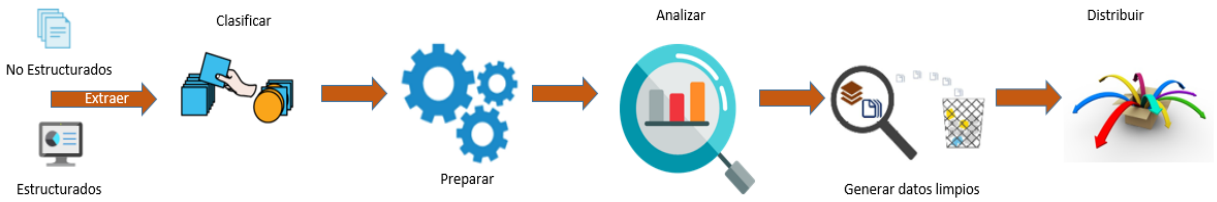
Figura 30 Dimensión Analítica de Datos



Teniendo en consideración el primer eje correspondiendo a la ingeniería de características, se propone una metodología de seis pasos cuyo objetivo es identificar las características relevantes para su extracción optimizando tiempos y recursos previos a su procesamiento:

- **Extraer**: Identificación de los datos sean estructurados o no estructurados y las fuentes donde se encuentren.
- **Clasificar**: Seleccionar los atributos más relevantes sobre el análisis a realizar.
- **Preparar**: Integrar los atributos seleccionados en la fase de clasificación.
- **Analizar datos**: Evaluar los datos para detectar valores fuera de rango, valores vacíos, etc.
- **Generar datos limpios**: Proceso de eliminar datos no necesarios y quedarse con los que serán utilizados en el análisis.
- **Distribuir datos**: Separar bloques de datos para su tratamiento de manera independiente.

Figura 31 Ingeniería de Características

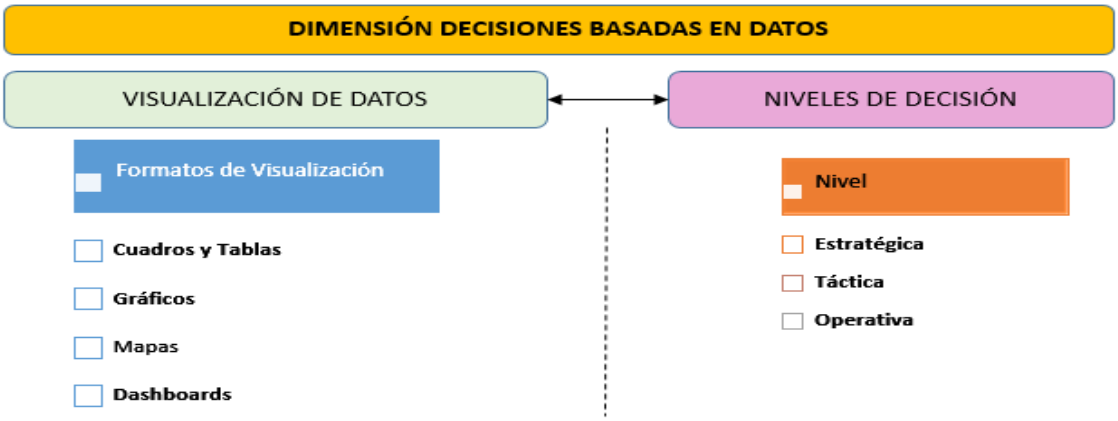


Teniendo en consideración el segundo eje sobre procesamiento computacional, se enfoca primero en:

- **Selección del algoritmo**: Se pueden encontrar diversos algoritmos que permiten apoyar en el procesamiento de datos. Se han clasificado de la siguiente manera:
    - **Clasificación:** Cuando se manipula una clase y en base a ella se predice la pertenencia o no del dato.
    - **Regresión:** Predicen, pero un valor como resultado final.
- **Despliegue y ejecución**: En este punto todo lo visto anteriormente se utiliza para realizar la ejecución y el procesamiento de datos, generando así los resultados analíticos deseados en base a la obtención de modelos, los cuales deben ser evaluados, posteriormente pasarán por el proceso de toma de decisiones.

La **Dimensión Decisiones basadas en datos**, se centra en dos ejes principales: la visualización de datos y los niveles de decisión.

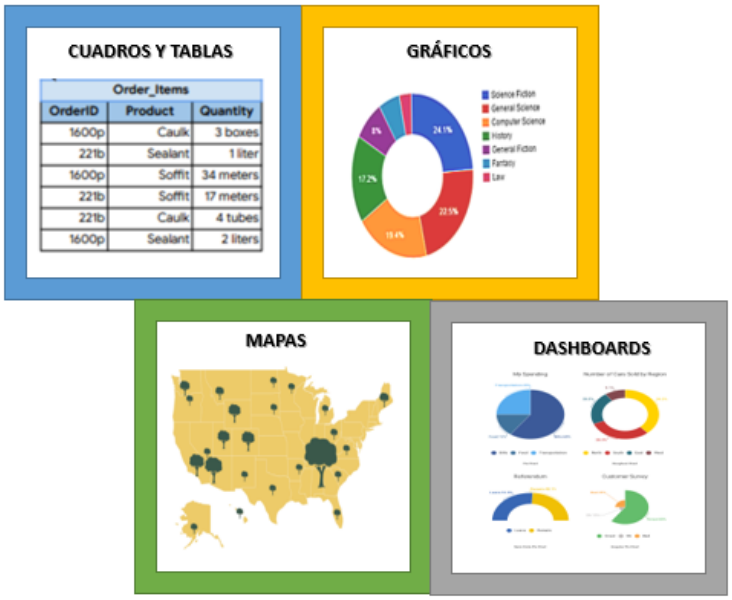
Figura 32 Dimensión Decisiones Basadas en Datos



Enfocándonos en el primer eje, el de la visualización de datos, entendiéndola como la capa que permite al usuario observar una representación de información a través de un formato visual para su mejor comprensión. Estos formatos de visualización se pueden clasificar en:

- **Cuadros y tablas:** Arreglo sistemático y ordenado de datos en columnas y filas, según ciertos criterios, con el objetivo de resumir y ordenar.
- **Gráficos:** Dibujo que permite observar las tendencias de un fenómeno permitiendo su análisis.
- **Mapas:** Se utilizan mapas que incluyen el uso de una variedad de colores logrando que se observe diferentes grados de intensidad.
- **Dashboards:** Herramienta de gestión de la información, que monitoriza, analiza y muestra de manera visual datos fundamentales que benefician a la institución.

Figura 33 Formatos de visualización de datos



Por otro lado, los niveles de decisión se pueden clasificar de la siguiente manera:

- **Decisiones estratégicas o de planificación**: Su objetivo es la de mejorar las prestaciones institucionales, son tomadas por altos directivos.
- **Decisiones tácticas o de pilotaje**: suelen ser tomadas por los directivos de nivel intermedio y supone la puesta en marcha de decisiones estratégicas.
- **Decisiones operativas o de regulación**: Orientadas a las actividades funcionales y rutinarias de la organización.

En lo referente a la seguridad en la Dimensión decisiones basada en datos, se debe considerar las siguientes estrategias:

- Definir controles de acceso a los datos.
- Definir los riesgos en la toma de decisiones.

VI. DISCUSIÓN

El diagnóstico situacional del estado actual de la dinámica del procesamiento de datos del área académica de esta institución, se obtuvo de los indicadores de la variable dependiente, a través de la aplicación de encuestas al personal de la institución que participan en los procesos académicos como directores de escuela, jefes de departamento, jefes de oficinas de asuntos académicos de cada facultad, una entrevista al director de asuntos académicos, una entrevista al jefe de la oficina de tecnologías de la información y análisis documental, posteriormente fueron procesadas y trianguladas para destacar los hallazgos más importantes del contexto analizado.

En cuanto, a la recolección de los datos en la institución se puede percibir que menos de la mitad (42%) consideran que los datos académicos almacenados son suficientes y necesarios para que en la institución se puedan tomar decisiones que permitan apoyar la parte académica institucional, por otro lado, un 40% de los encuestados está de acuerdo o totalmente de acuerdo con que los datos académicos obtenidos son claros dentro de su área de trabajo, pero hay un 60% que está en desacuerdo o simplemente están indecisos en su apreciación.

Por otro lado, la frecuencia con la que se recopilan datos académicos pueden influir en el conocimiento que se tiene de este proceso a nivel de los datos de matrículas, docentes y asignación de cargas lectivas, entre otros, se obtuvo un 63% que consideran que es muy frecuente o frecuentemente la recopilación de datos académicos en la institución universitaria.

Considerando los sistemas académicos con los que cuenta la institución, un 55% considera que NO son intuitivos ni fáciles de manipular, por lo que se puede deducir que el sistema fue desarrollado sin los correctos requerimientos de usuario o por la avanzada edad de las personas que lo utilizan. Mientras que un 45% consideran que los datos se validan al momento de registrarse en los sistemas.

En referencia a la dimensión de Manipulación de los datos, un 40% de usuarios considera que raramente los datos están disponibles en el momento en que se necesitan para sus actividades diarias, mientras que un 42% manifiesta que están frecuentemente o muy frecuentemente disponibles los datos. Así mismo un 42% indica que los reportes que se obtienen en base a los datos si les permiten realizar análisis de acuerdo a las necesidades que presenta su área.

Por otro lado, sólo un 20% cree que las herramientas tecnológicas que posee la institución influyen en la extracción y procesamiento de los datos, mientras que un 80% cree que faltan herramientas tecnológicas que la institución necesita.

Además, solo un 32% manifiesta que existe facilidad de acceso a los datos académicos que tiene almacenada la institución, y un 46% plantea que no existe facilidad de acceso a dichos datos.

Considerando la dimensión de Calidad, más del 80% indican que raramente o nunca se detectan datos erróneos, mientras que un reducido 18% considera que si se detectan datos erróneos.

Un 78% considera que se debe implementar estándares de calidad para el procesamiento de datos en la institución, que brinden la posibilidad de que los datos que se registren o procesen sean buenos.

Por otro lado, un 43% de los encuestados indican que frecuentemente la captura de los datos está centrada en las necesidades organizacionales y de cada área en particular.

Teniendo en consideración la dimensión Rendimiento, solo un 22% considera que es buena o muy buena el tiempo de respuesta de las aplicaciones al solicitar datos, mientras que un 78% considera que es mala, baja o regular.

También se aprecia que solo un 17% considera que los datos académicos se obtienen en tiempo real.

Considerando la dimensión Seguridad, un 60% no está de acuerdo ni en desacuerdo sobre la seguridad de los datos académicos con los que cuenta la institución. Además, un 61% están indecisos o en desacuerdo que la accesibilidad se presenta solo a las personas autorizadas. Sin embargo, un 65% considera que raramente se solicita que actualicen claves de acceso a los diferentes sistemas académicos, a la vez que un 35% considera que nunca se solicita actualizaciones de claves. Por otro lado, las modificaciones de datos si se realiza mediante autorización, tal como lo manifiesta un 68% de encuestados.

Por último, la dimensión Soporte presenta un 37% de usuarios que consideran que los requerimientos son atendidos de manera rápida por la oficina de tecnologías de la información, y un 25% casi siempre. Por otro lado, un 25% de los usuarios consideran que la atención de algún incidente relacionado con los sistemas académicos fue casi siempre rápida.

VII. CONCLUSIONES

- Se caracterizó el procesamiento de datos de acuerdo a los planteamientos conceptuales establecidos en su manipulación, captura, integridad, transformación, análisis y visualización.
- Se diagnosticó el estado actual del procesamiento de datos en la Universidad, así como también se realizó un análisis de estudios previos relacionados al tema en donde se determinó que el rendimiento promedio es de 84.95%, también, se observó que hay un 89.09% de rendimiento académico clasificado correctamente.
- Se elaboró un modelo predictivo para el procesamiento de datos académicos utilizando características del rendimiento académico de los estudiantes. El modelo se compone de cuatro dimensiones: soporte tecnológico, analítica del negocio, analítica de datos y las decisiones basadas en datos.

VIII. RECOMENDACIONES

- Ampliar la recolección de datos de los estudiantes en el ámbito familiar y económico que permita tener más características que puedan ser incluidas en el procesamiento y que influyan dentro de los modelos analizados.
- Establecer los requerimientos correctamente, que permita aplicar el modelo que se ajuste a las necesidades, y evaluar temas como deserción estudiantil, influencia de herramientas tecnológicas en la virtualidad, influencia del aula virtual en los cursos de los estudiantes, etc.

REFERENCIAS

Álvarez Valle, J. (2013, noviembre 10). *El Big Data: Una gran oportunidad*. La Nueva España. https://www.lne.es/opinion/2013/11/10/big-data-gran-oportunidad-20536839.html

Britos, L., Di Gennaro, M. E., Gil Costa, G. V., Kasián, F., Lobos, J., Ludueña, V., Molina, R., Printista, A. M., Reyes, N. S., Roggero, P., & Trabes, G. (2016, mayo 19). *Búsquedas en grandes volúmenes de datos*. XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). http://sedici.unlp.edu.ar/handle/10915/52900

Camejo Corona, J., Gonzalez, H., & Morell, C. (2019). Los principales algoritmos para regresión con salidas múltiples. Una revisión para Big Data. *Revista Cubana de Ciencias Informáticas*, *13*(4), 118-150.

Cano, I. C. M. (2016). *Reflexiones epistemológicas sobre Big Data*. 24.

Cravero, A., Sepúlveda, S., & Muñoz, L. (2020). Big Data Architectures for the Climate Change Analysis: A Systematic Mapping Study. *IEEE Latin America Transactions*, *18*(10), 1793-1806. https://doi.org/10.1109/TLA.2020.9387671

De Battista, A., Cristaldo, P., Ramos, L., Nuñez, J. P., Retama, S., Bouzenard, D., & Herrera, N. E. (2016). *Minería de datos aplicada a datos masivos*. https://core.ac.uk/reader/301069340

Duque Méndez, N. D., Hernández Leal, E. J., Pérez Zapata, Á. M., Arroyave Tabares, A. F., & Espinosa Gómez, D. A. (2016). Modelo para el proceso de extracción, transformación y carga en bodegas de datos. Una aplicación con datos ambientales. *Ciencia e Ingeniería Neogranadina*, *26*(2), 95-109. https://doi.org/10.18359/rcin.1799

Elmasri, R., & Navathe, S. B. (2007). *Fundamentos de Sistemas de Bases de Datos* (5 ed.). https://www.ingebook.com/ib/NPcd/IB_BooksVis?cod_primaria=1000187&codigo_libro=2886

Escobar Borja, M., & Mercado Pérez, M. (2019). Big data: Un análisis documental de su uso y aplicación en el contexto de la era digital. *Revista La Propiedad Inmaterial*, *28*, 273-293. https://doi.org/10.18601/16571959.n28.10

García, S., Ramírez-Gallego, S., Luengo, J., & Herrera, F. (2016). *Big Data: Preprocesamiento y calidad de datos*. novática; 2133_Nv237-Digital-sramirez.pdf. https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf

Guzmán Ponce, A., Valdovinos Rosas, R. M., Marcial Romero, J. R., & Alejo Eleuterio, R. (2018). Entornos de trabajo para procesamiento de datos masivos y aprendizaje automático. *Research in Computing Science*, *147*(5), 225-237. https://doi.org/10.13053/rcs-147-5-17

Hernández-Leal, E. J., Duque-Méndez, N. D., & Moreno-Cadavid, J. (2017). Big Data: Una exploración de investigaciones, tecnologías y casos de aplicación. *TecnoLógicas*, *20*(39), 15-38. https://doi.org/10.22430/22565337.685

Malberti, A., Klenzi, R. O., & Beguerí, G. (2016, mayo 17). *Análisis, interpretación y toma de decisiones estratégicas en la Ciencia de Datos*. XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). http://sedici.unlp.edu.ar/handle/10915/52853

Moreno, J. P. (2014). Una aproximación a Big Data. *Revista de Derecho de la UNED (RDUNED)*, *14*, 471-506. https://doi.org/10.5944/rduned.14.2014.13303

Nuñez-Arcia, Y., Díaz-de-la-Paz, L., & García-Mendoza, J. L. (2016). *Algoritmo para corregir anomalías a nivel de instancia en grandes volúmenes de datos utilizando MapReduce*. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992016000300008

Peñaloza Báez, M. J. (2018). Big data y analítica del aprendizaje en aplicaciones de salud y educación médica. *Investigación en educación médica*, *7*(25), 61-66. https://doi.org/10.1016/j.riem.2017.11.003

Quinteros, O. E., Funes, A., & Ahumada, H. C. (2016, mayo 17). *Extracción de conocimiento en el cursado del ciclo común de articulación de carreras de Ingeniería*. XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). http://sedici.unlp.edu.ar/handle/10915/52850

Quiroz Martinez, M., Aguilar Duarte, R. A., & Intriago Cedeño, D. B. (2020). *Proceso de diseño de una arquitectura Big Data para el análisis de grandes volúmenes de datos e información | Opuntia Brava*. https://opuntiabrava.ult.edu.cu/index.php/opuntiabrava/article/view/968/1280

Russo, C. C., Ramón, H. D., Alonso, N., Cicerchia, B., Esnaola, L., & Tessore, J. P. (2016, mayo 17). *Tratamiento masivo de datos utilizando técnicas de machine learning*. XVIII Workshop de Investigadores en Ciencias de la

Computación (WICC 2016, Entre Ríos, Argentina). http://sedici.unlp.edu.ar/handle/10915/52838

Sampieri, R. H., Collado, C. F., & Lucio, P. B. (2014). *Metodología de la investigación*. McGraw Hill. https://dialnet.unirioja.es/servlet/libro?codigo=775008

Sandoval, L. J. (2018). *Algoritmos de aprendizaje automático para análisis y predicción de datos.* 5.

Santos Grueiro Igor. (2015, enero 19). *El tamaño sí es importante | Revista Ingeniería*. https://revistaingenieria.deusto.es/el-tamano-si-es-importante/

Schab, E., Rivera, R., Bracco, L., Coto, F., Cristaldo, P., Ramos, L., Rapesta, N., Núñez, J. P., Retamar, S., Casanova, C., Battista, A. D., & Herrera, N. E. (2018). *Minería de Datos y Visualización de Información*. 5.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2002). *Database system concepts* (4th ed). McGraw-Hill.

Téllez Carvajal, E. (2020). Análisis documental sobre el tema del big data y su impacto en los derechos humanos. *Derecho PUCP*, *84*, 155-188. https://doi.org/10.18800/derechopucp.202001.006

Tolosa, G. H., Banchero, S., Ríssola, E. A., Delvechio, T., & Feuerstein, E. (2016, mayo 17). *Grandes datos y algoritmos eficientes para búsquedas de escala web*. XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). http://sedici.unlp.edu.ar/handle/10915/52852

Vite Cevallos, H., Townsend Valencia, J., Carvajal Romero, H., Vite Cevallos, H., Townsend Valencia, J., & Carvajal Romero, H. (2020). Big Data e internet

de las cosas en la producción de banano orgánico. *Revista Universidad y Sociedad*, *12*(4), 192-200.

Vivas, H. L., Cambarieri, M. G., Petroff, M., García Martínez, N., Formia, S., & Muñoz Abbate, H. (2015). *Tratamiento de Grandes Volúmenes de Datos en Ciudades Inteligentes Una Propuesta de Big Data con NoSQL*. http://rid.unrn.edu.ar/handle/20.500.12049/150

Roger Ernesto Alarcón García
Universidad Nacional Pedro Ruiz Gallo –Lambayeque
https://orcid.org/0000-0003-1178-0519
ralarcong@unprg.edu.pe
Doctor en Ciencias de la Computación y Sistemas. Maestro en Ingeniería de Sistemas. Ingeniero en Computación e Informática. Docente nombrado en la Universidad Nacional Pedro Ruiz Gallo de Lambayeque – Perú, adscrito al Departamento académico de Computación y Electrónica. Asesor de tesis de pregrado y posgrado con más de 20 años en la docencia universitaria.

Jessie Leila Bravo Jaico
Universidad Nacional Pedro Ruiz Gallo – Lambayeque
https://orcid.org/0000-0001-6841-2536
jbravo@unprg.edu.pe
Doctora en Ciencias de la Computación y Sistemas. Magister en Informática y Multimedia en la Universidad de Los Lagos - Chile. Magister en Administración de empresas con mención en Gerencia Empresarial. Coordinadora del Grupo de Investigación en Transformación Digital - UNPRG. Directora de la carrera prof. de Ing. Computación en Informática - UNPRG. Docente nombrada en la Universidad Nacional Pedro Ruiz Gallo de Lambayeque – Perú, adscrita al Departamento académico de Computación y Electrónica. Conferencista en eventos nacionales e internacionales.

Janet del Rosario Aquino Lalupú
Universidad Nacional Pedro Ruiz Gallo – Lambayeque
https://orcid.org/0000-0003-0536-3882
jaquino@unprg.edu.pe
Maestra en Administración de Empresas con mención en Gerencia Empresarial en la Universidad Nacional Pedro Ruíz Gallo. Ingeniero en Computación y Sistemas, egresada de la Universidad Privada Antenor Orrego de Trujillo. Doctorante en Educación. Docente nombrada en la Universidad Nacional Pedro Ruiz Gallo de Lambayeque – Perú, adscrita al Departamento académico de Computación y Electrónica. Asesora de tesis de pregrado y posgrado con más de 25 años en la docencia universitaria.

Carlos Alberto Valdivia Salazar
Universidad Nacional Pedro Ruiz Gallo – Lambayeque
https://orcid.org/0000-0002-2895-9120
cvaldivias@unprg.edu.pe
Magister en Ingeniería de Sistemas. Ingeniero en Computación e Informática. Estudios concluidos de doctorado en Ciencias de la Computación y Sistemas. Docente nombrado en la Universidad Nacional Pedro Ruiz Gallo de Lambayeque – Perú, adscrito al Departamento Académico de Computación y Electrónica. Investigador con publicaciones en revistas indexadas. Especialista en tecnologías de la información. Desarrollador de software empresarial. Desarrollador de investigaciones en Transformación Digital. Docente en educación superior y asesor de tesis de pregrado y posgrado, con más de 20 años de experiencia.

Nilton César Germán Reyes
Universidad Nacional Pedro Ruiz Gallo – Lambayeque
https://orcid.org/0000-0003-0232-2129
ngerman@unprg.edu.pe
Doctor en Educación. Maestro en Administración de empresas con mención en Gerencia empresarial. Ingeniero en Computación y sistemas. Estudios concluidos de Doctorado en Ciencias con mención en Sistemas. Docente nombrado en la Universidad Nacional Pedro Ruiz Gallo de Lambayeque – Perú, adscrito al Departamento académico de Computación y Electrónica. Asesor de tesis de pregrado y posgrado con más de 25 años en la docencia universitaria.

ISBN: 978-9942-603-58-6