UNIVERSIDAD SEÑOR DE SIPÁN

ESCUELA DE POSGRADO

TESIS

SISTEMA ANALÍTICO BASADO EN UN MODELO PREDICTIVO DE PROCESAMIENTO DE DATOS EN LA BIG DATA EN LA EDUCACIÓN SUPERIOR

PARA OPTAR EL GRADO ACADÉMICO

DE DOCTOR EN CIENCIAS DE LA COMPUTACIÓN Y SISTEMAS

Autor:

Mg. Alarcón García, Roger Ernesto

ORCID: 0000-0003-1178-0519

Asesora:

Dra. Bravo Jaico, Jessie Leila

ORCID: 0000-0001-6841-2536

Línea de Investigación:

Gestión de Aplicaciones con innovación Tecnológica

Pimentel – Perú

2021

UNIVERSIDAD SEÑOR DE SIPÁN

ESCUELA DE POSGRADO

DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN Y SISTEMAS

"SISTEMA ANALÍTICO BASADO EN UN MODELO PREDICTIVO DE PROCESAMIENTO DE DATOS EN LA BIG DATA EN LA EDUCACIÓN SUPERIOR."

AUTOR

Mg. ROGER ERNESTO, ALARCÓN GARCÍA

PIMENTEL – PERÚ

2021

SISTEMA ANALÍTICO BASADO EN UN MODELO PREDICTIVO DE PROCESAMIENTO DE DATOS EN LA BIG DATA EN LA EDUCACIÓN SUPERIOR

APROBACIÓN DE LA TESIS

Dr. Bustamante Quintana, Pepe Humberto

Presidente del jurado de tesis

Dr. Callejas Torres, Juan Carlos

Secretario del jurado de tesis

Dra. Bravo Jaico, Jessie Leila

Vocal del jurado de tesis

Dedicatoria

A mi esposa Jessie por ser la pieza fundamental en mi vida y su constancia de superación profesional y personal, me impulsan a seguir adelante.

A mis hijas Vanessa y Luciana, por ser la razón de vivir, por darme tanto amor, cariño y alegrías, que me permiten seguir mejorando cada día.

Mg. Roger Ernesto Alarcón García

Agradecimientos

A Dios, por darme las fuerzas necesarias
y protegerme durante todo este camino.

Al Dr. Juan Carlos Callejas Torres, asesor metodólogo, por su paciencia y dedicación, en el proceso de culminación de la tesis. Y también a mi asesora de tesis Dr. Jessie Bravo por ser la guía en el desarrollo de la tesis.

Al Jefe de Oficina de Servicios Académicos, al Jefe de la Oficina de tecnologías de la Información y al Jefe de la Oficina de Admisión de la Universidad Nacional Pedro Ruiz Gallo, por las facilidades brindadas al proporcionar la información relevante de su área para la realización de este trabajo de investigación.

Y a todas aquellas personas que, de una u otra forma, colaboraron o participaron en la realización de esta investigación, hago extensivo mi agradecimiento.

Resumen

Las organizaciones en la actualidad están generando un alto volumen de datos a través de sus procesos internos, los mismo que permiten satisfacer sus necesidades, pero acarrea dificultades dadas las limitantes capacidades humanas que han sido superadas para lograr analizar los datos y generar conocimiento útil que le permita a la organización tomar las mejores decisiones, la presente investigación tiene como problema el inadecuado procesamiento de datos que limitan el tratamiento de los datos académicos. Las causas detectadas sugieren profundizar en el procesamiento de datos en la Big data, por lo que se plantea como objetivo Aplicar un Sistema Analítico basado en un modelo predictivo que tenga en cuenta la relación entre las técnicas predictivas integradas y los grandes volúmenes de datos para el procesamiento de los datos académicos.

Se propone un modelo predictivo de procesamiento de datos en la Big data diseñado en base a 4 dimensiones: Soporte tecnológico, analítica del negocio, analítica de datos y decisiones basadas en datos; este modelo se materializa en un Sistema Analítico que se implementa en 8 etapas: Caracterización del usuario, requerimiento del negocio, almacenamiento de datos, infraestructura tecnológica, ingeniería de características, procesamiento computacional, visualización y decisiones.

Finalmente, el sistema analítico se implementa en forma práctica, utilizando 5479 datos de estudiantes de los semestres del 2016-I al 2020-I, de los cuales se utilizaron 1096 para la evaluación del rendimiento del sistema, obteniéndose un 92.97% en la detección correcta de forma global el cual fue superior al 84.95% obtenido en el sistema base.

Palabras Clave:
Big data, modelo predictivo, procesamiento de datos, sistema analítico.

Abstract

Organizations today are generating a high volume of data through their internal processes, the same ones that allow satisfying their needs, but entails difficulties given the limiting human capacities that have been overcome to be able to analyze the data and generate useful knowledge that Allowing the organization to make the best decisions, the present investigation has as a problem the data processing that limits the treatment of academic data. The causes detected suggest deepening the processing of data in big data, so the objective is to apply an Analytical System based on a predictive model that takes into account the relationship between integrated predictive techniques and large volumes of data for the processing of academic data.

A predictive data processing model in big data designed based on 4 dimensions is proposed: technological support, business analytics, data analytics and data-based decisions; This model is materialized in an Analytical System that is implemented in 8 stages: User characterization, business requirement, data storage, technological infrastructure, characteristics engineering, computational processing, visualization and decisions.

Finally, the analytical system is implemented in a practical way, using 5479 student data from the semesters of 2016-I to 2020-I, of which 1096 were used to evaluate the performance of the system, obtaining 92.97% in the correct detection. globally which was higher than the 84.95% obtained in the base system.

Keywords:

Big data, predictive model, data processing, analytical system.

Índice

Lista de Tablas

Listado de Figuras

I. INTRODUCCIÓN

1.1. Realidad Problemática.

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 las que es imposible acceder utilizando las actuales técnicas de recuperación y generación de información.

Las actuales organizaciones están generando un alto volumen de datos los mismo que permiten satisfacer sus necesidades, 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 datos.

Por lo tanto, actualmente se observa el requerimiento de gestionar grandes volúmenes de información las cuales puede estar en formato estructurado o no estructurado producidas por el internet o por diversas organizaciones públicas o privadas, las cuales están 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. Peter Sondergaard, vicepresidente de la reputada consultora Gartner, dijo: "La información es la gasolina del siglo 21 y el análisis de datos el motor de combustión". Según Santos, el uso de nuevas y robustas tecnologías de aprendizaje automático a partir de datos, está permitiendo aplicar métodos que permiten traducir texto, predecir trayectorias o inclusive recomendar búsquedas. (Santos Grueiro, 2015)

(Alvarez 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, logrando.

La Universidad Nacional Pedro Ruiz Gallo ubicada en la Ciudad de Lambayeque, 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 dato que tiene registrada por años a través de los diferencias 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, la gran cantidad de datos con la que cuentan las organizaciones los llevan a poder, 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. (Vivas, y otros, 2015)

Desde el punto de vista de Maestro, 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. (Maestro Cano, 2016)

También, Maestro 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 (Maestro Cano, 2016); 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, encuentra que las capacidades de procesamiento en los sistemas clásicos que utilizan la minería de datos ha sido superado 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. (García, Ramírez Gallego, Luengo, & Herrera, 2016)

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, Puyol 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. (Puyol Moreno, 2014)

Además, Escobar 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. (Escobar Borja & Mercado Pérez, 2019)

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 públicas.

1.2. Trabajos Previos

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

(Vite Cevallos, Townsend Valencia, & Carvajal Romero, 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 el denominado Modelo BIG DATA Agrícola, otro también 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, y otros, 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, y otros, 2016) menciona que el proceso 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 dato 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, Funes, & Ahumada, 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 en su investigación manifiesta que el uso de motores de búsqueda basadas 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 esto, 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 intuiciones. 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 aplique directamente a estos grandes volúmenes de datos. (Tolosa, Banchero, Ríssola, Delvechio, & Feuerstein, 2016)

(Malberti, Klenzi, & Beguerí, 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, y otros, 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, 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 almacenado en las computadoras y se pierdan sin nunca ser utilizados. (De Battista, y otros, 2016)

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 la Big Data está enfocada en el volumen de datos, velocidad con la que llegan los datos, variedad de estos datos siendo estructurada o no estructurada, la confianza por parte de los usuarios para suministrar la información y la protección de los datos. Asumiendo también la veracidad de los datos, es decir, la precisión y la confianza de los datos que se manejan.

Además, según Quiroz, el uso de la Big Data apoya la construcción de sistemas novedosos que permiten tomar decisiones más acertadas, benefician tanto a la población como específicamente al personal administrativo. Por lo que, se ve la necesidad 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, y que también estas sean precisas. (Quiroz Martinez, Aguilar Duarte, & Intriago Cedeño, 2020)

Según (García, Ramírez Gallego, Luengo, & Herrera, 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 de los tiempos en la toma de decisiones.

Así mismo (Duque Méndez, Hernández Leal, Pérez Zapata, Arroyave Tabares, & Espinosa Gómez, 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. En este artículo, 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, González Diez, & Morell, 2019) en su investigación, la presencia de problemas de predicción especialmente cuando se manipulan grandes volúmenes de datos, dado 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 den la industria global. Provocando serios problemas como la gran cantidad de tiempo que requiere para el procesamiento computacional que permitan 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 también 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 es 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 cuentas 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, 2017) 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 mismo.

Según (Hernández Leal, Duque Méndez, & Moreno Cadavid, 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.

1.3. Teorías relacionadas al tema.

1.3.1. Modelamiento de datos.

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 & Navathe, 2007)

Por lo que, en este modelo, las columnas representas 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 (A_1, A_2, ..., A_n), en donde, A_1, $A_{2,..,}$ A_n representan los atributos de la relación y R la relación en sí.

Además, dentro del modelo relación 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 mismo atributos y la segunda condición es 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 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; 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 algoritmos que permiten inferir comportamientos basados en datos empíricos.	Predicción de fenómenos como crimen, deserción escolar y universitaria, 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 permite establecer **modelos predictivos**, los cuales permiten

inferir resultados usando las probabilidades, posibilitando así identificar oportunidades de negocio.

Además, los científicos de datos son los designado 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 conocería 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 ya será útil, por lo que se convierte los datos de su forma original a un formato más entendible, dándoles la forma y el contexto necesarios para que pueda 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 a partir de los datos.

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 puede indicar que el procesamiento de datos estimula el avance de los computadores. (Silberschatz, Korth, & Sudarshan, 2014)

Según (Schab, y otros, 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, impactando directamente en la toma de decisiones de cualquier área de estudios. También se conoce que generan datos a gran velocidad y en grandes cantidades.

Según (Schab, y otros, 2018) 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 entonces 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 recogida 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 datos de las cuales se extraiga tendrán que ser fiables y de buena calidad.

La segunda etapa implica la preparación de los datos, partiendo de todos los datos recabados en la etapa anterior, se procede con la *limpieza* de los datos, 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, Díaz de la Paz, & García Mendoza, 2016), se enfoca en la evaluación y análisis de posible errores que se puedan presentar en las fuentes de datos en 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 presentan 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 necesiten y convertirlo a un formato que sea entendible por la computadora. Teniendo lo datos se puede obtener información importante a través del uso de diversos procedimientos para que 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 de datos para el ingreso de los mismos de manera automatizada, y en el caso sea el registro

de los datos se realice manual, es necesarios una buena concentración durante un periodo largo de tiempo.

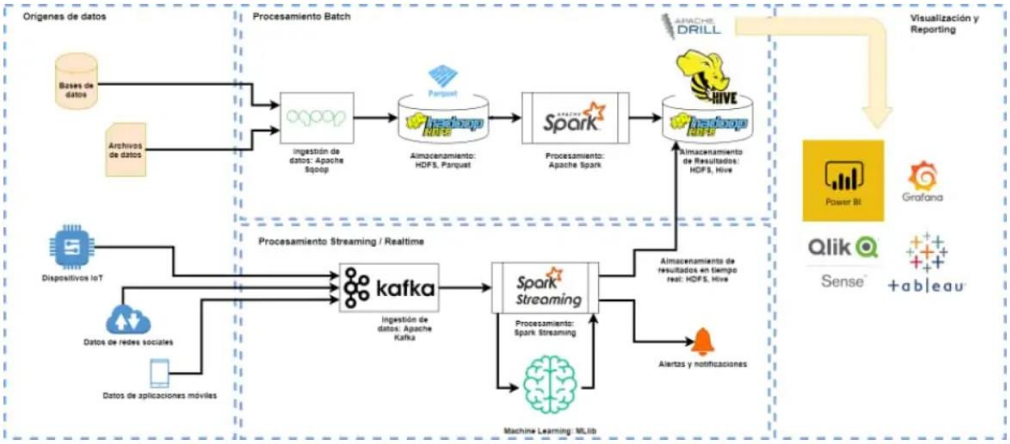
- Captura de datos: Múltiples métodos están disponibles para capturar datos de documentos, debe tenerse debidamente en cuenta el origen de los documentos que se deben capturar. Algunos métodos son: reconocimiento óptico de caracteres (OCR), reconocimiento inteligente de caracteres (ICR), reconocimiento 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 intensión de que sea ilegible, y con esto pueda 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 de los datos 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 también cualitativos que se utilizan con la finalidad de incrementar la productividad y por ende la ganancia de las empresas, aplicando diversas técnicas que pueden variar.
- 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.

1.3.2. 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 también procesar y almacenar datos. Cuando indicamos grandes cantidades nos 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



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

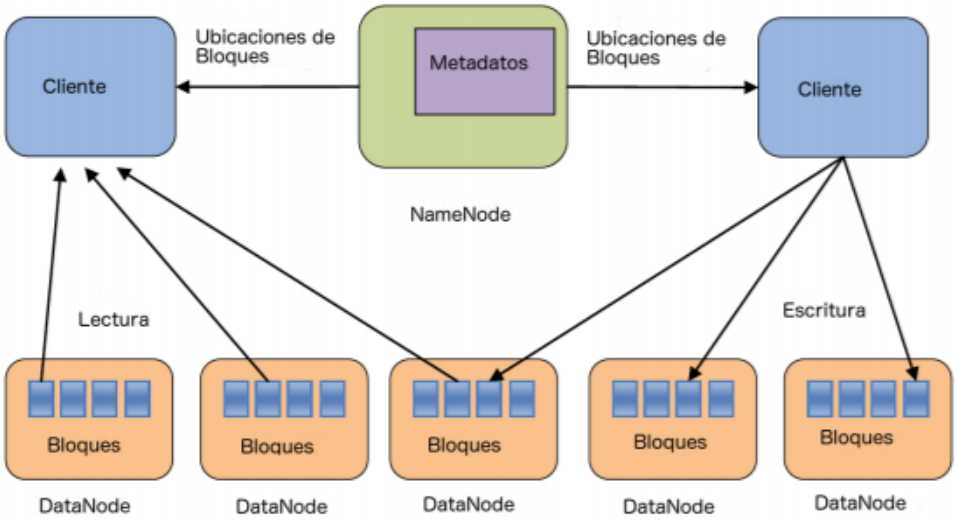
- El procesamiento de datos que permite determinar en tiempo real la eficiencia de los procesos industriales.
- El uso de datos para el análisis financiero los cuales son generados en tiempo real a través del uso de procesos por lotes.
- Optimización de procesos que permiten tiempos de respuesta muy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

1.3.3. Entornos de procesamiento de Datos

Según (Guzman Ponce, Valdovinos Rosas, Marcial Romero, & Alejo Eleuterio, 2018) plantea y analiza tres diferentes entornos de procesamiento de datos destacando propiedad clave tales como 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.

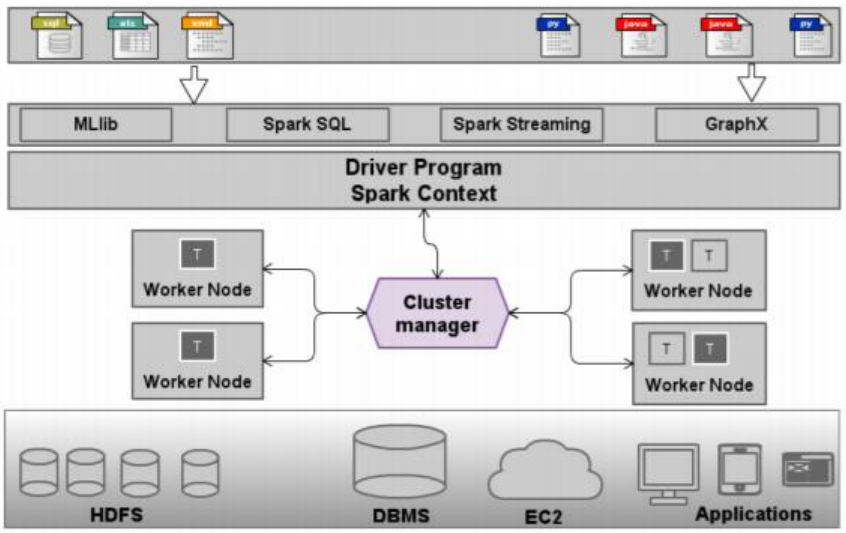
Apache Hadoop: Desde el 2008 esta tecnología viene siendo usada para la manipulación de datos masivos, es de código libre, desarrollado para ser utilizado de forma distribuida y escalable. Se divide en dos componentes: HDFS y el cómputo distribuido con la idea de MapReduce. HDFS este sistema de archivos 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



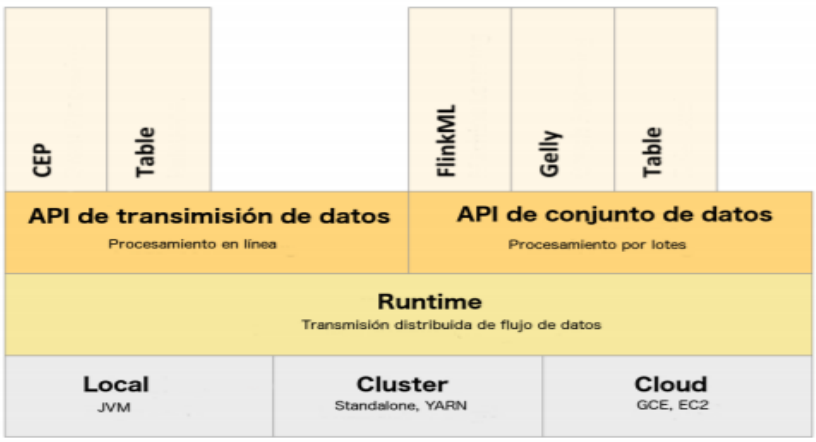
Apache Spark: Cuenta con diversas ventajas sobre el resto de los entornos de trabajo, hoy en día es utilizado por empresas como Yahoo, Baidu, entre otras. De igual manera que Hadoop, es de código libre, para procesos de manera distribuida, basado en el mejoramiento del rendimiento de memoria. En la figura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

Figura 3 Arquitectura Spark



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



1.3.4. 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.
- 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 onjetos en k grupos en base a sus características.
- LDA (Latent Dirichlet allocation). Técnica que intenta reducir las dimensiones del conjunto de características.

Por otro lado (Cravero, Sepúlveda, & Muñoz, 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.

1.3.5. Tendencias históricas del proceso de procesamiento de datos

Desde que aparece las primeras computadoras, los usuarios requieren de la manipulación de datos cada en mayor escala, la velocidad de acceso a los mismos y la capacidad de almacenamiento que se incrementa constantemente, está generando necesidades mayores en los usuarios, y más aún al manipular volúmenes de datos organizacionales que les permitan tomar decisiones eficientemente.

Se puede indicar la evolución histórica del procesamiento de datos teniendo en cuenta los siguientes **indicadores de análisis**: recolección, calidad, seguridad, tratamiento, rendimiento y soporte.

Época Antigua - desde la Antigüedad hasta el año 1945

En esta primera etapa se presenta una **recolección** de datos de manera primitiva utilizando para esto pequeñas máquinas que operaban con sumas y restas.

No se tenía la idea de **calidad** en los datos recopilados, ni mucho menos de **seguridad** en los mismos.

En cuanto al **tratamiento** estaba centrado básicamente en operaciones básicas de sumas y restas, para esto Von Leibniz apoyó con la construcción de una máquina capaz de realizar las cuatro operaciones básicas.

En cuanto al **rendimiento,** era muy pobre dada la poca cantidad de datos que se trataban. No se tenía una clara idea del **soporte** que se le brinde a las aplicaciones que manipulaban datos debido a que eran aplicaciones simples.

Época de Archivos - desde 1946 hasta el año 1960

Entre los años 1946 a 1955 aparece la primera generación de computadoras que fueron construidas utilizando válvulas al vacío y relés electromagnéticos. Además de usar un procesador secuencial, la **recolección** de los datos se realizaba usando las tarjetas perforadas. Solo trabajaba con 20 números de diez dígitos.

Ya en el año 1950 se conservan los datos después de que el proceso que los creó deja de existir. Aparecen los ficheros que son un tipo de archivos que permitía

almacenar información en una unidad como, por ejemplo, un disco duro. En esta etapa muy similar a la etapa anterior no se profundiza en el tema de **calidad** ni **seguridad** de los datos. En cuanto al **tratamiento** de los datos, se realizaba de forma secuencial, esto implicaba tiempo de procesamiento y por lo tanto generaba lentitud al obtener resultados finales. Por otro lado, el **rendimiento** era aún muy bajo y el procesamiento costoso. En esta etapa el **soporte** para la manipulación de datos en los diferentes sistemas estaba en una etapa incipiente.

Época de Modelos de archivos – Del Año 1960 al año 1970

En el año 1960, aparecen los modelos de datos jerárquicos y de red, estos permitían tener un mejor control de los datos. La **recolección** y captura de los datos se realizaba a través de pequeñas aplicaciones. La forma de almacenamiento dependía del modelo utilizado. Por lo que se conocían dos modelos: el de red en donde se organizaban como colecciones de grafos arbitrarios y el modelo jerárquico que apareció a mediados de los años 60 y domino el mercado hasta mediados de los 80. Es un modelo de datos orientado a registro por lo que están organizados como colecciones de árboles.

La **calidad** de los datos se considera importante, pero se manipulaba a través de los programas construidos para servir en el llenado de los datos que la dependencia requería. Su **tratamiento** lo realizaba también haciendo uso de las aplicaciones construidas, y en cuanto al **rendimiento** este solo se enfocaba en los métodos tradicionales de acceso a los datos de forma secuencial, lo que generaba problemas si había demasiados datos que tratar. En esta etapa el **soporte** para la manipulación de datos en los diferentes sistemas estaba en una etapa básica.

Época de Base de Datos – Del Año 1970 al año 2010

Entre los años 1970 a 1980 aparecen las bases de datos relacionales, que permiten almacenar una colección de datos conectados unos con otros, sin que se duplique la información; con el objetivo de apoyar a todas las aplicaciones que se construyan, por lo que los datos son totalmente independientes a las aplicaciones que usen dichos datos. Estas bases de datos permiten registrar nuevos datos, modificarlos y extraerlo de una manera fácil y práctica.

La base de datos se crea en forma separada de los programas que acceden a los datos. Los datos se consideran como un recurso compartido e independiente de las aplicaciones que las utilicen.

Para la **recolección** aparece el uso de un sistema de administración de base de datos (SGBD). Los efectos del uso de este enfoque permiten la mejora de la **calidad** de los datos, soluciona el problema de múltiples usuarios y los datos se caracterizan por ser finitos, mientras que las aplicaciones son infinitas. Ejemplos de este tipo de ambientes son los gestores de base de datos Microsoft Access, Microsoft SQL Server, y ORACLE. El **tratamiento** se realiza a través de herramientas de software como los sistemas gestores de base de datos, los cuales hacen uso de instrucciones en lenguaje denominado T-SQL. En esta etapa se puede considerar que se tiene ya consideración de la influencia de los datos y de su **seguridad,** se establece la posibilidad de trabajar usando roles y usuarios, permitiendo mejorar la seguridad en los datos. El **rendimiento** mejora notablemente, debido al uso de herramientas tecnológicas conocidos como los gestores de base de datos los cuales procesan los datos en minutos y segundos. En esta etapa el **soporte** para la manipulación de datos en los diferentes sistemas se consideraba relevante, ya que estos sistemas eran grandes y complejos.

Época de la Big Data – Desde el año 2010 hasta la actualidad

Desde el año 2010 que incursiona el término de Big Data debido al incremento de la cantidad de datos que se están manipulando, la **recolección** de los datos sigue utilizando los sistemas transaccionales y los contenedores de datos conocidos además también hay una influencia del uso de la IoT y los múltiples dispositivos permitiendo se siga incrementando la captura de datos día a día. La **calidad** de los datos sigue siendo uno de los problemas que se presentan en este enfoque, al igual que, controlar la **seguridad** de los mismos, los cuales permitan que las organizaciones logren tener almacenado tanta cantidad de datos, que el **tratamiento** de los mismo se ha convertido en otro dolor de cabeza para los analistas de datos, los cuales utilizar técnica que se aplican en minería de datos, y no directamente enfocados en grandes volúmenes de datos, incrementado este problema la variedad de datos existentes en la big data. En esta etapa el **soporte** para la manipulación de

datos en los diferentes sistemas estaba en una etapa donde su uso es muy relevante debido a la posibilidad de cambio de requerimientos.

No obstante, a la evolución tendencial que ha tenido el proceso de procesamiento de datos aún son insuficientes los referentes teóricos y prácticos a la dinámica del mismo para su sistematización, diagnóstico, fundamentación teórica, desarrollo de actividades, su apropiación y generalización para el tratamiento de datos.

Tabla 3 Evolución histórica del proceso de procesamiento de datos y su dinámica

Indicador de análisis	Época Antigua Hasta 1945	Época de archivos (1946-1960)	Época de modelos de archivos (1960-1970)	Base de datos (1970-2010)	Big data (2010 - …)
Recolección	Datos primitivos	Uso de tarjetas perforadas	Uso de archivos	Base de datos	Altos volúmenes de datos
Calidad	No se tenía ni idea	No se evaluaba	A través de aplicativos	Centrada en los datos	Alta centrada en los datos
Seguridad	No existía	No existía	Mediante aplicativos	Manejo de niveles de seguridad	Manejo de niveles de seguridad
Tratamiento	Solo operaciones básicas	De forma secuencial	Usando aplicativos	Sistemas gestores y lenguajes T-SQL	Variado
Rendimiento	Pobre	Muy bajo	Intermedio	Muy bueno	Alto
Soporte	No	Incipiente	bueno	Muy bueno	Alto

Fuente: Elaboración propia

1.3.6. Marco Conceptual.

a) **Análisis Predictivo**. Es el uso de métodos estadísticos aplicando algoritmos y técnicas de aprendizaje automático con el objetivo de encontrar comportamientos y resultados futuros en base a la data histórica con la que se cuente.

b) **Arquitectura de datos**. Se basa en el uso de modelos, reglas o políticas que nos guían al momento de almacenar, ordenar y unir datos que son recolectados por las empresas que les permitan utilizarlos y aprovecharlos.

c) **Base de Datos**. Se le considera como un almacén de datos, y en él se registran de manera ordenada, con la finalidad de ser fácil de encontrar y utilizar por las organizaciones.

d) **Big Data**. Son enormes cantidades de datos provenientes de diversos tipos, que incluyen datos con una estructura fija, otros que no tienen estructura fija y los semi estructurados que se basan en el uso de etiquetas. Por lo general el uso de métodos y técnicas convencionales no permite un adecuado procesamiento de volúmenes tan grandes de datos.

e) **Calidad de datos**. Es el proceso de acondicionamiento de los datos para que satisfagan las necesidades concretas de los usuarios corporativos.

f) **Conocimiento**. Adquisición de múltiples datos relaciones a través de la experiencia o usando el aprendizaje a priori o a posteriori.

g) **Dato**. Es considerada como la unidad mínima de información, que por sí sola no ayuda mucho en la comprensión de un fenómeno. Puede ser un valor numérico, alfanumérico, espacial entre otros.

h) **Extracción de conocimiento**. Proceso original para el descubrimiento de información y nuevo conocimiento basado en los datos contenidos en repositorios que permitirá su uso en la toma de decisiones.

i) **Información**. Son los datos procesados, nos dan una mejor idea de lo que se está manipulando a diferencia de los datos simples, aunque depende mucho de las personas que lo utilizan.

j) **Minería de datos**. Aplicación de métodos que permiten obtener patrones desconocidos en los datos, y que son potencialmente útiles para comprender el fenómeno estudiado.

k) **Modelo**. Representaciones abstractas o formales de objetos tanto reales o propios del software que conforman un dominio específico.

l) **Modelo de datos**. Diagrama que representa el flujo de los datos basado en textos y símbolos, que permiten su entendimiento y comprensión.

m) **Modelo predictivo**. Permiten realizar un análisis para identificar la correlación entre un conjunto de variables de datos de entrada y una variable de respuesta o

destino, buscando siempre tener unas salidas deseadas; pero también existen modelos que solo presentan datos de entrada y en este tipo de modelos lo que se busca es encontrar la relación de unos datos con otros.

n) **Lenguaje Estructurado de Consultas**. es un lenguaje similar a la de los lenguajes de programación, pero basado solo en la manipulación de datos, para su inserción, actualización o eliminación a través de los sistemas de información.

o) **Seguridad de los datos**. Evitar el acceso a los datos ubicados en servidores, bases de datos y otros espacios, por personas que no tienen autorización de su uso.

p) **Sistema**. Aplicación de normas y procedimientos de manera ordenada que permitan llevar un control en el funcionamiento de algo.

q) **Sistema analítico**. Sistemas en la que las organizaciones utilizan para estudiar y observar todos sus datos históricos y en tiempo real que permita detectar patrones y producir conocimientos que apoye la toma de decisiones inteligentes.

r) **Sistema informático**. Aquel sistema que integra no solo la parte física o hardware sino también la parte lógica o software que unidos permiten apoyar a mejorar los procesos organizacionales.

s) **Sistema Gestor de Base de Datos**. Programa o conjunto de programas interconectados que permiten registrar, modificar y extraer información que se presenta en las bases de datos.

t) **Procesamiento de datos**. Es el trabajo sobre los datos a través de la recolección, el almacenaje, empleo, movimiento o supresión. Cualquier operación o conjunto de operaciones sobre datos, tales como la recolección, almacenamiento, uso, circulación o supresión.

u) **Toma de decisiones**. Capacidad que pueda tener una persona o conjunto de personas de poder elegir entre varias opciones, basándose en ciertos criterios establecidos que apoyen tal acción.

v) **Técnicas predictivas**. Procedimientos o recursos que permiten entrenar a un modelo o método a través del uso de diferentes datos con la intención de generar una variable partiendo de ellos.

w) **Visualización de datos**. Presentación de la información en un formato simple y práctico, útil para cualquier persona, basada en el uso de mapas, cuadros o gráficas intuitivas. Es muy común el uso de herramientas de visualización.

1.4. Formulación del Problema.

El inadecuado procesamiento de los datos usando técnicas predictivas aplicados a grandes volúmenes de datos y el análisis de la big data limita el tratamiento de los datos académicos.

1.5. Justificación e importancia del estudio.

La justificación de este trabajo consiste en mejorar el procesamiento de los datos en una institución pública superior. El problema de la investigación se define como: La deficiente gestión de los datos y su integración en la Big Data que limitan su procesamiento.

Actualmente la gran cantidad de datos que se obtienen ya sea a través de los diferentes sistemas informáticos o de diversos dispositivos, así como también el uso de IoT, está generando una alta concentración de datos en las organizaciones, pero sin utilizar ni procesar, lo que conlleva a una pérdida en la generación de conocimiento para una mejor toma de decisiones.

Esto trae consigo una serie de problemas a resolver que afectan el procesamiento de los datos orientados a grandes volúmenes, en términos de:

Valor

Si analizamos el valor marginal de los datos, observaremos que a medida que aumenta su volumen y complejidad de los datos, su valor marginal disminuye, debido a su *dificultad de explotación*.

Calidad de los datos

No existe ningún umbral para referirnos a la calidad de los datos, pero si debemos tener presente que cualquier análisis realizado a los datos sin que estos tengan un

cierto grado de calidad implica un resultado errado en su procesamiento. Por lo que debemos revisar desde la misma recopilación de datos, el ingreso de los datos y la depuración de los mismo como parte importante para poder realizar algún procesamiento de datos.

Diversidad de datos

Por otra parte, la existencia de una variedad de tipos de datos como son los estructurados y semiestructurados, así como también las diferentes fuentes de obtención como los textos, imágenes, datos de la web, redes sociales, audio, video, entre otros, generan una dificultad adicional para procesarlos a gran escala.

Volumen de datos

La existencia de múltiples dispositivos conectado actualmente está generando que el volumen de datos se multiplique cada vez más rápido. He ahí la importancia de una buena priorización y gestión de los datos (lo que equivale a decir "gestión de fuentes de datos") pasará a ser imprescindible.

Seguridad de datos

La manipulación y procesamiento de datos ayuda a las organizaciones a ser más eficientes y tomar mejores decisiones, pero también las hace muy vulnerables a múltiples ataques de hackers que intentan acceder a estos datos, por lo que uno de los problemas existentes actualmente es la seguridad de los datos al momento de almacenarlos, transferirlos o tratarlos.

Todos estos problemas generan que no se procese correctamente los datos académicos por lo que retoma importancia el desarrollo de esta investigación que aporta un modelo predictivo para la manipulación de datos académicos haciendo uso de un sistema analítico.

Teniendo como **aporte teórico** el modelo predictivo que permitirá analizar datos académicos en diversos contextos, teniendo en cuenta diferentes técnicas de análisis y volúmenes altos de datos.

El **aporte práctico** se evidencia a través de un Sistema Analítico que permitirá realizar análisis de datos académicos utilizando técnicas de predicción.

1.6. Hipótesis y operacionalización de variables

1.6.1. Hipótesis

La hipótesis planteada en la presente investigación que se sujetará a la contrastación está definida como:

Si se aplica un **Sistema Analítico** basado en un **modelo predictivo** que tenga en cuenta la relación entre las **técnicas predictivas** integradas y los grandes volúmenes de datos, **entonces** se contribuye al procesamiento de los datos en la big data.

La **novedad científica** será determinar técnicas predictivas que apoyen en el análisis de datos académicos con la finalidad de que se tomen mejores decisiones.

La **significancia práctica**, se establece en el impacto social y académico en poder implementar un sistema analítico para el procesamiento de los datos académicos.

1.6.2. Variables.

VARIABLE INDEPENDIENTE:

Sistema analítico basado en un modelo predictivo

Definición Conceptual: son herramientas que proporcionan una fácil y rápida visualización de datos en tiempo real, con el objetivo de analizar situaciones, determinar tendencias, elaborar un plan de acción y más. Todo esto es posible a través de cuadros de mando e informes inteligentes.

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.

1.7. Objetivos

1.7.1. Objetivos General

Aplicar un Sistema Analítico basado en un modelo predictivo que tenga en cuenta la relación entre las técnicas predictivas integradas y los grandes volúmenes de datos para el procesamiento de los datos en la big data.

1.7.2. Objetivos Específicos

- Caracterizar epistemológicamente el proceso de procesamiento de datos y su dinámica.
- Determinar las tendencias históricas del proceso de procesamiento de datos y su dinámica.
- Diagnosticar el estado actual del procesamiento de datos en la Universidad Nacional Pedro Ruiz Gallo.
- Elaborar un modelo predictivo para el procesamiento de datos académicos.
- Elaborar un sistema analítico basado en un modelo predictivo para el tratamiento de datos académicos.
- Validar los resultados de la investigación.

II. MATERIAL Y MÉTODO

2.1. Tipo y Diseño de Investigación.

El tipo de estudio para esta investigación es transversal descriptiva de tipo mixta aplicada según (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014) es descriptiva porque se busca medir o recoger información de manera independiente sobre las variables en estudio, y transversal, ya que se refiere a 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 en 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 (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 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 utilizando el sistema analítico propuesto el cual se basa en un modelo predictivo centrado en el uso de grandes volúmenes de datos.

2.2. Población y muestra.

La población está definida por los datos académicos de todos los estudiantes de la universidad 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 4 Total de registros académicos desde el 2005 al 2020

Semestre	Total registros Académicos
2005-I	77083

2005-II	69797
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
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 del 2016-I al 2020-I, que corresponde a los 5 últimos años, siendo un total de 684 039 registros académicos de los semestres 2016-I hasta el 2020-I, como se detalla en la siguiente tabla.

Tabla 5 Total de registros académicos 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

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

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.
-

2.4. Procedimientos de análisis de datos.

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.

2.5. Criterios éticos

En el presente proyecto de investigación se tomarán en cuenta los siguientes principios 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 sobre el proceso de seguridad informática para redes privadas virtuales 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.

2.6. Criterios de Rigor científico.

En el presente trabajo de investigación se han aplicado un conjunto de procedimientos que han permitido definir el aporte tanto el teórico como el práctico en base a los siguientes criterios de rigor científico.

- **Transferibilidad**: los resultados del presente trabajo de investigación, pueden ser aplicables a otros contextos.
- **Confirmabilidad**: para la presente investigación se han manipulado datos sin realizar ningún tipo de sesgo o intervención de índole personal.
- **Credibilidad**: En el trabajo de investigación se ha valorado el procesamiento y tratamiento de los datos en el contexto de la manipulación de grandes volúmenes de datos, los resultados obtenidos han permitido desarrollar la construcción del aporte teórico y práctico.

III. RESULTADOS

3.1. Resultados del diagnóstico del estado actual de la dinámica del procesamiento de datos del área académica en una institución universitaria

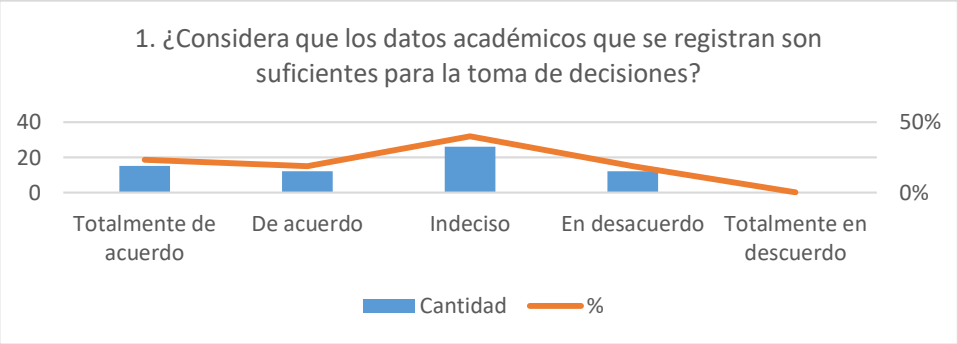
Se esta etapa se aplicó una encuesta al personal de la institución relacionada con los procesos académicos con el objetivo de diagnosticar el estado actual de la dinámica del procesamiento de datos del área académica de esta institución. La aplicación de la encuesta se realizó a través de los correos 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?

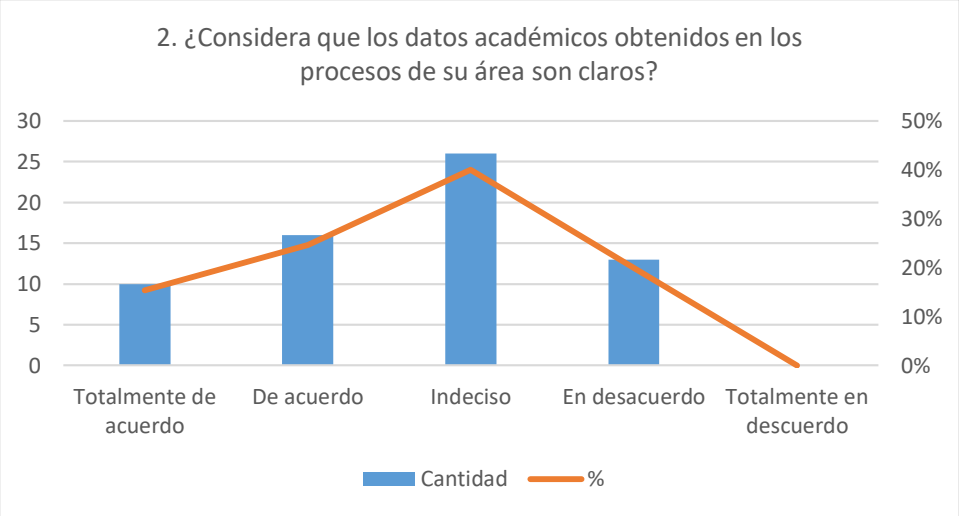


Fuente: Elaboración propia

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?

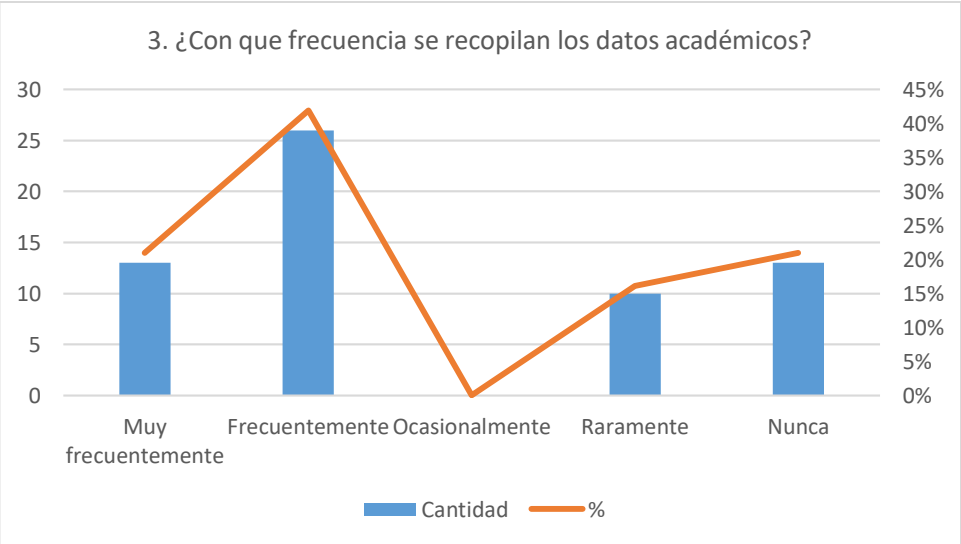


Fuente: Elaboración propia

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?

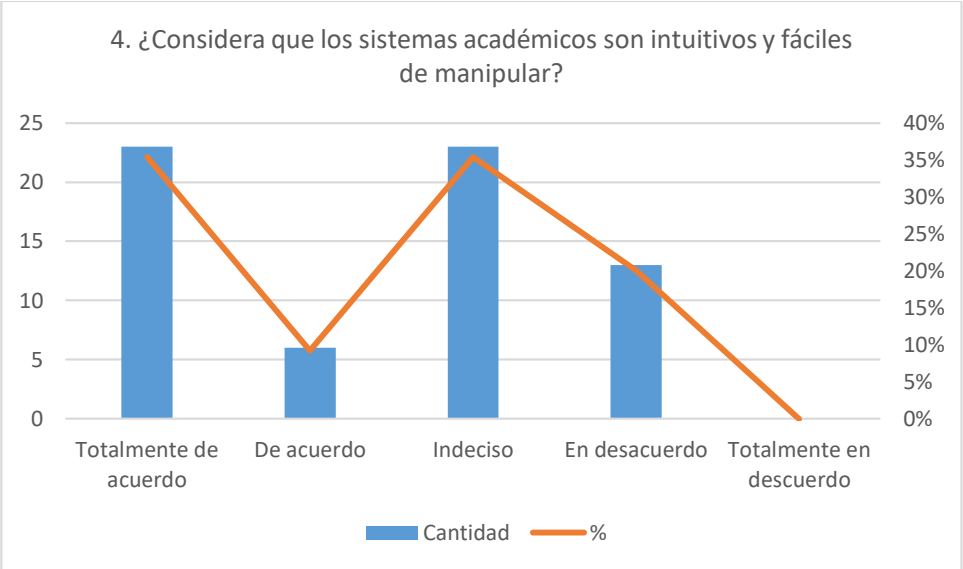


Fuente: Elaboración propia

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?

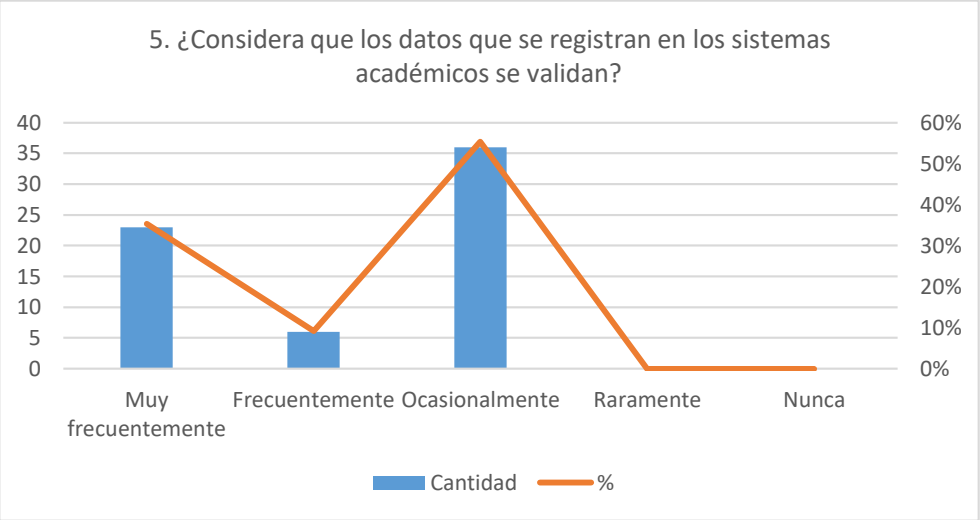


Fuente: Elaboración propia

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?



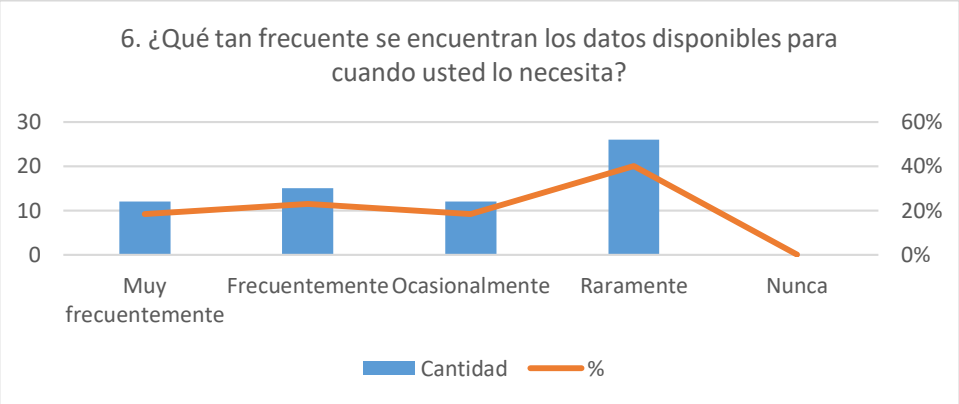
Fuente: Elaboración propia

Se observa que un 45% considera que los datos que se registran en los sistemas académicos se validad, 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?

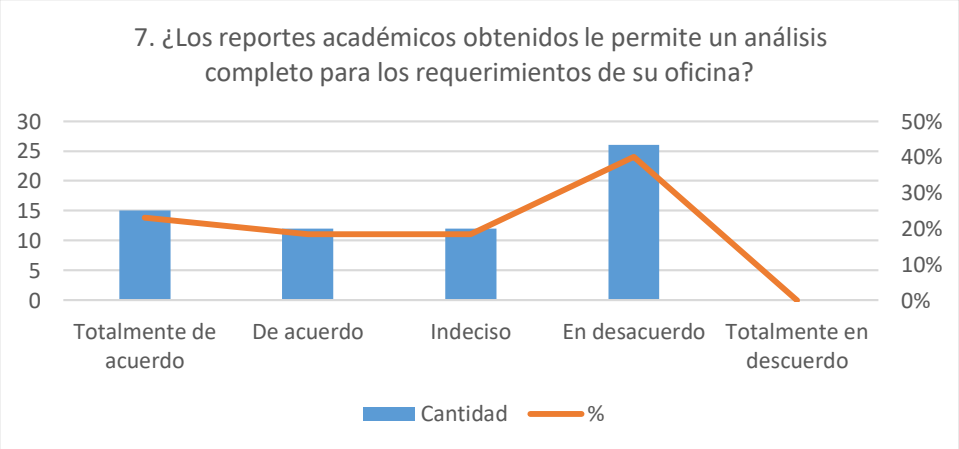


Fuente: Elaboración propia

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?

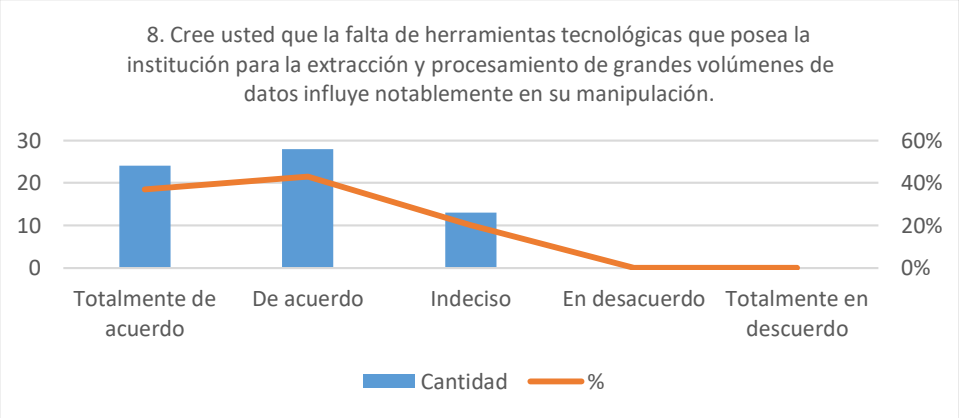


Fuente: Elaboración propia

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.

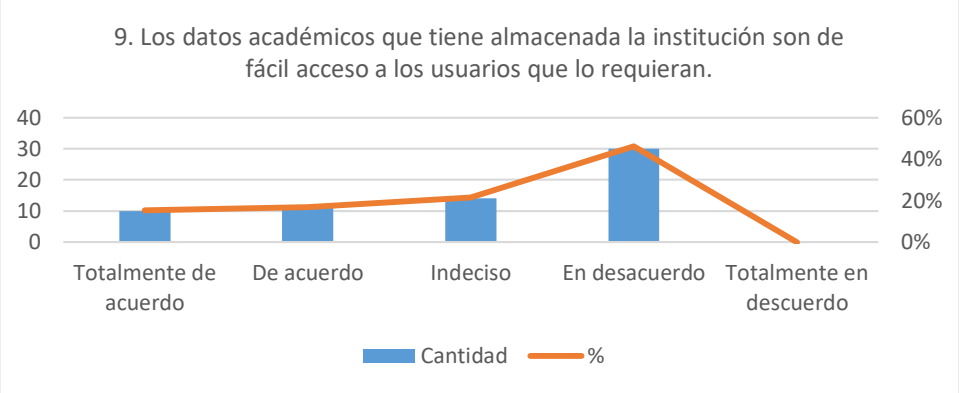


Fuente: Elaboración propia

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.



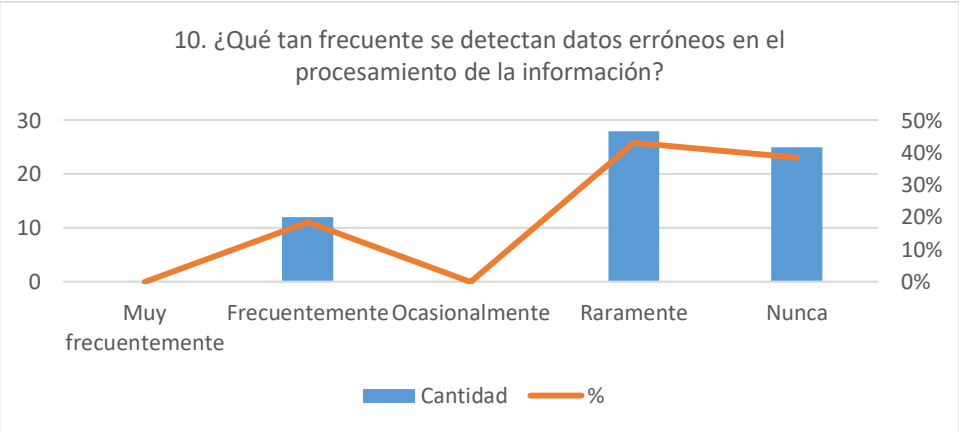
Fuente: Elaboración propia

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?

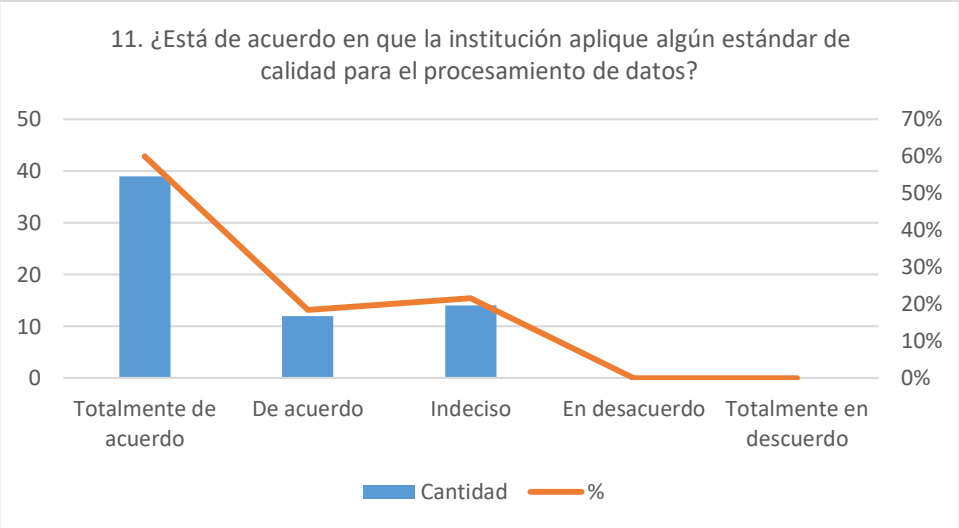


Fuente: Elaboración propia

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?

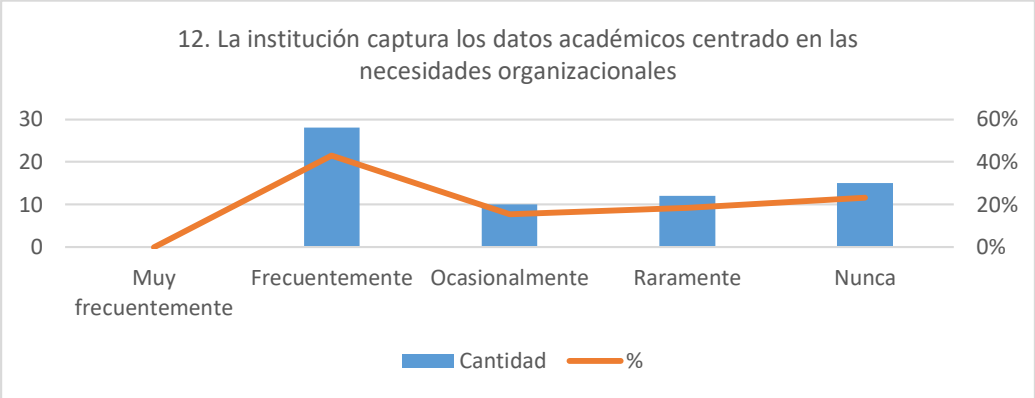


Fuente: Elaboración propia

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



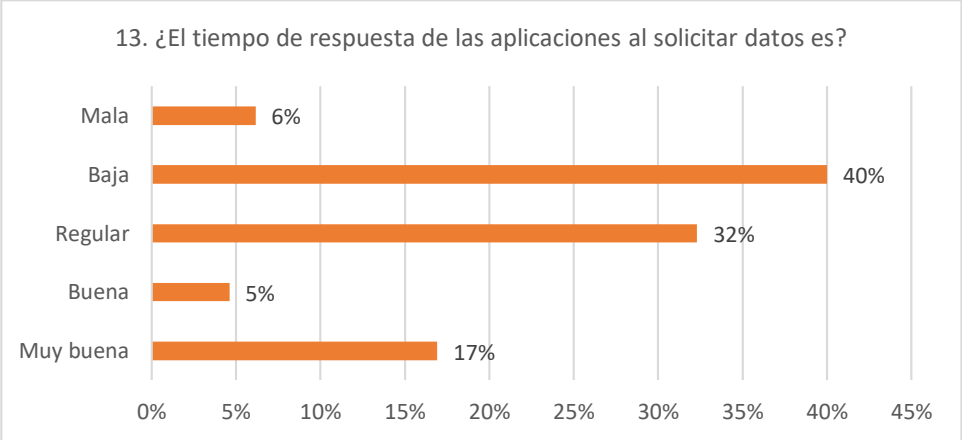
Fuente: Elaboración propia

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 organizaciones, 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?

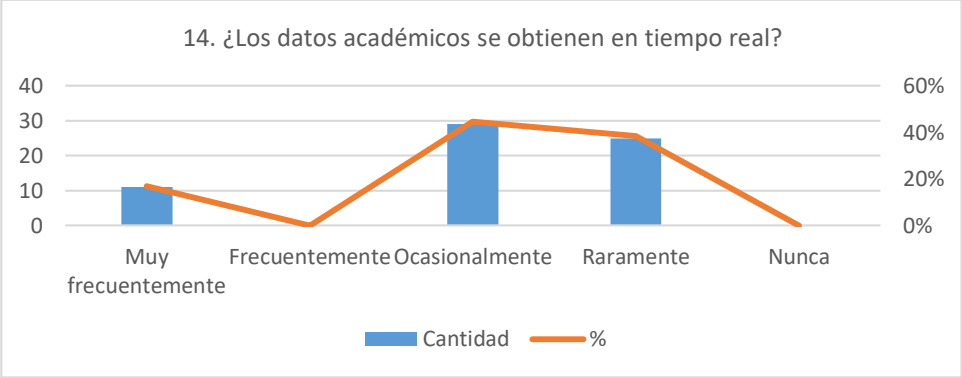


Fuente: Elaboración propia

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?



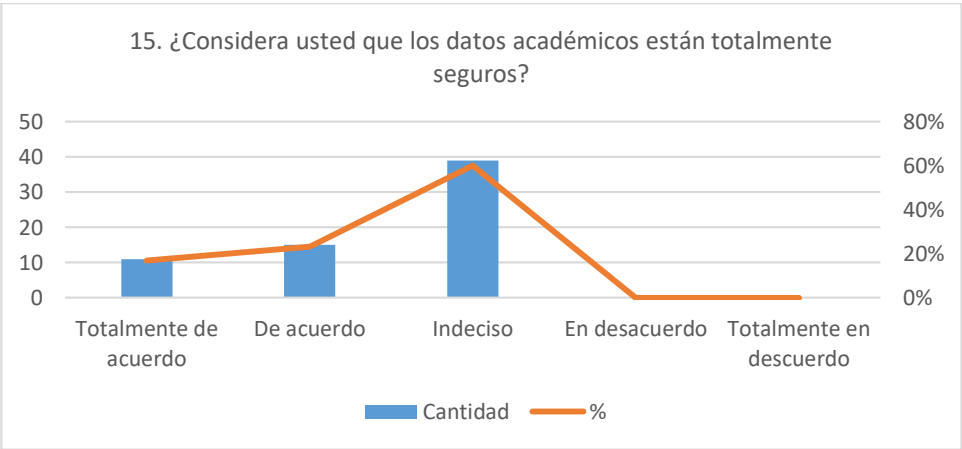
Fuente: Elaboración propia

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?

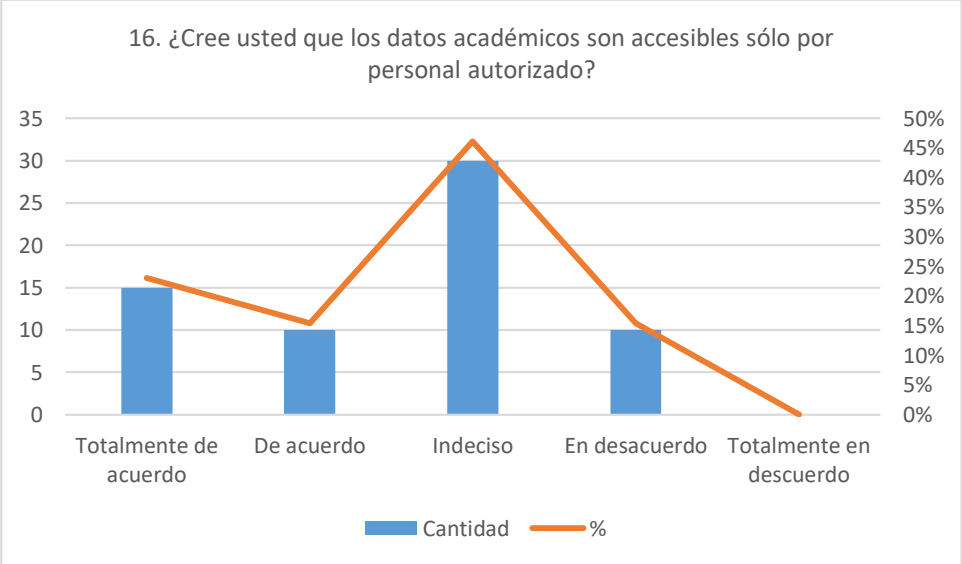


Fuente: Elaboración propia

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?

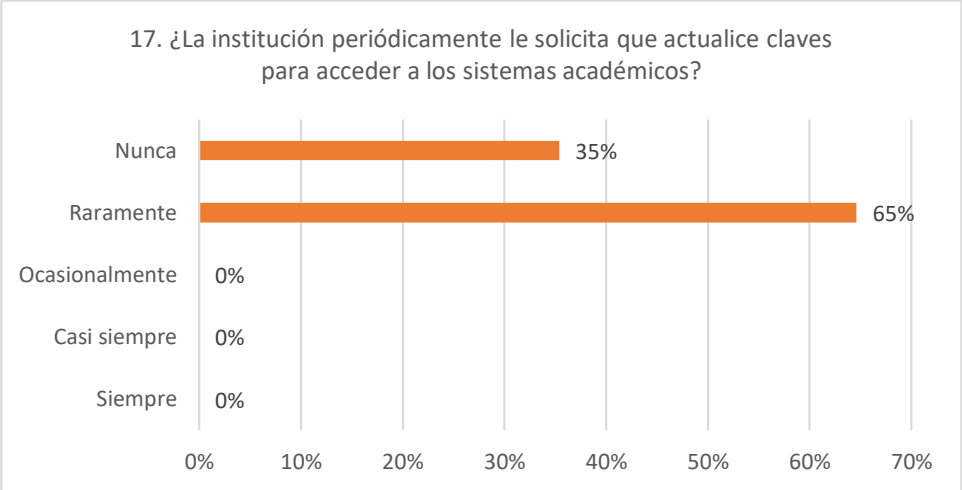


Fuente: Elaboración propia

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?

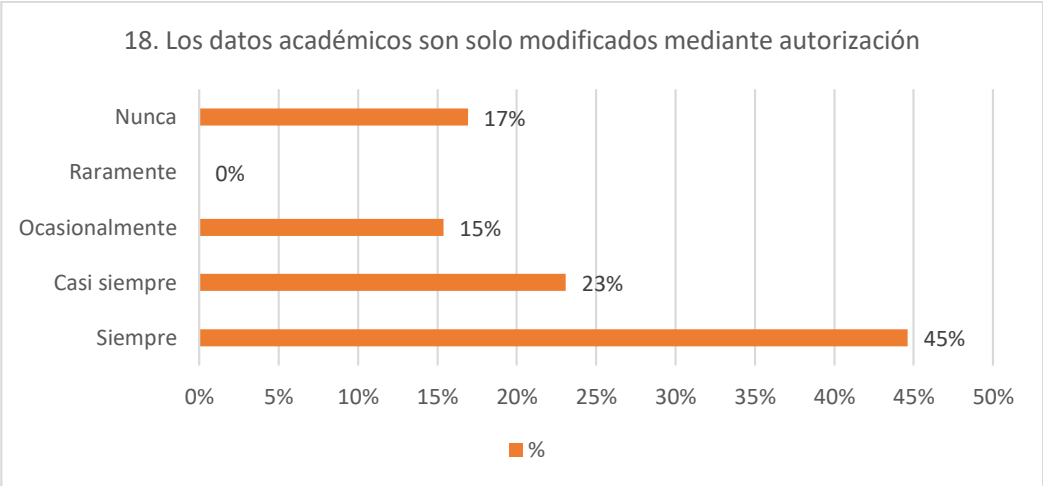


Fuente: Elaboración propia

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



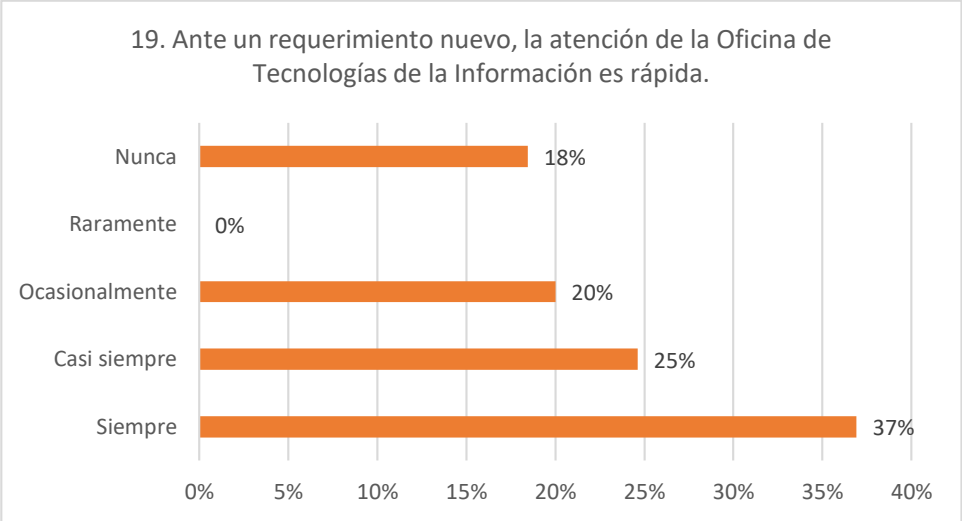
Fuente: Elaboración propia

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.

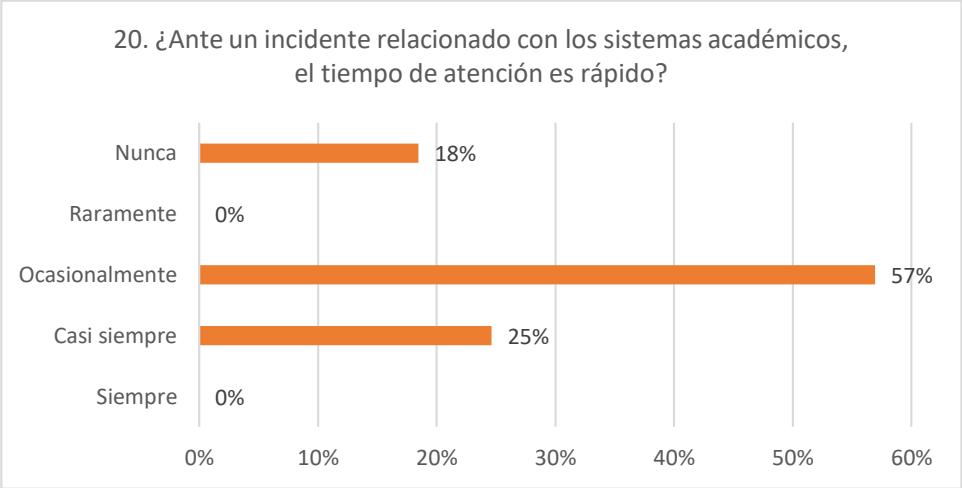


Fuente: Elaboración propia

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?



Fuente: Elaboración propia

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.

3.2. Resultados del Análisis previo

Se analizaron los resultados del rendimiento en la predicción con respecto al rendimiento académico universitario como se muestra en la Tabla 6 especificando el algoritmo utilizado y la precisión.

Tabla 6 Rendimiento académico en estudios previos

Autores	**Algoritmo de análisis de datos**	**Precisión**
(Contreras, Fuentes, & Rodríguez, 2020)	SVC	66.24 %
	KNN	55.00 %
	PERCEPTRÓN	66.00 %

	Árbol de decisión	55.00 %
(Ayala Franco, López Martínez, & Menéndez Dominguez, 2021)	J48	69.87 %
	RandomForest	68.91 %
	LMT	71.08 %
	Logistic	70.84 %
	MultilayerPerceptron	65.54 %
(Orihuela Maita, 2019)	Regresión Logística	75.00 %
	Random Forest	76.00 %
(Pojon, 2017)	Regresión lineal	93.50 %
	Árboles de decisión	91.70 %
	Clasificador Naive Bayes	95.60 %
(Menacho Chiok, 2017)	Regresión Logística	68.40 %
	Árboles de decisión J48	68.30 %
	Redes Neuronales	67.90 %
	Naive Bayes	71.00 %

En la Tabla 7, se observa el rendimiento promedio de 71.99% en la predicción basados en el rendimiento académico, una desviación estándar de 0.11, un mínimo de 55% y un máximo de 95.6%.

Tabla 7 Resumen de rendimiento académico en estudios previos

Métrica	Media	Desviación	Máximo	Mínimo
Precisión	71.99%	0.11	95.60%	55.00%

Se utilizaron los 1096 datos para la ejecución del sistema base dando como resultado lo que se muestra en la tabla 8.

Tabla 8 Matriz de confusión del Sistema Base

	Clases	Resultados del Sistema Base		
		Buen Rendimiento	Mal Rendimiento	Total
Rendimiento académico	Buen Rendimiento	449 (VP)	55 (FN)	504
	Mal Rendimiento	110 (FP)	482 (VN)	592

Por lo que, los resultados del rendimiento del sistema base visible en la tabla 9 indican que el 84.95% es la proporción de clasificación correcta del sistema en forma global, un 15.05% de error, un 80.32 correctamente clasificados y un 89.09% es la precisión.

Tabla 9 Resultados del rendimiento del sistema base

Métrica Evaluada	**Fórmula**	**Resultado**
Accuracy	Accuracy = (VP+VN) / (VP+VN+FP+FN)	84.95%
Classification Error	Classification Error = (FP+FN)/(VP+VN+FP+FN)	15.05%
Recall (TVP)	recall = VP / (VP+FN)	80.32%
Specificity (TVN)	Specificity = VN / (VN+FP)	89.76%
Precision	Precision = VP / (VP + FP)	89.09%

3.3. Discusión de resultados

Los resultados de los estudios previos indican una baja precisión en la evaluación del rendimiento académico que sirva como predictivo para otros casos de estudiantes.

Adicionalmente, en esta sección se presenta el diagnóstico situacional del estado actual de la dinámica del procesamiento de datos del área académica de esta institución, esta información 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 lo referente a la recolección de los datos en la institución se puede percibir que menos de la mitad (42%) consideran que los datos académicos que son

almacenados son lo 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 considerando 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, hay 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 consideran que están frecuentemente o muy frecuentemente disponibles los datos. Mientras que 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 hay una falta de herramientas tecnológicas con las que la institución debería tener.

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 solicita 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, hay un 61% está indecisos o en desacuerdo que la accesibilidad se presenta solo a las personas autorizadas. Sin embargo, un 65% considera que raramente se solicitan que actualicen claves de acceso a los diferentes sistemas académicos, a la vez que un 35% considera que nunca se solicitan actualizaciones de claves. Por otro lado, las modificaciones de datos si se realiza mediante autorización con un 68% de encuestados.

Por último, la dimensión Soporte presenta un 37% de que los usuarios 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.

3.4. Construcción del Aporte teórico

En este capítulo se explica la construcción epistemológica del modelo predictivo de procesamiento de datos en la big data contextualizado al sector académico,

mediante el uso de un enfoque holístico basándose en 4 dimensiones, de soporte tecnológico, la de analítica del negocio, la de analítica de datos y la de decisiones analíticas.

3.4.1. Fundamentación del aporte teórico

El aporte teórico corresponde a un modelo predictivo para el procesamiento de datos en la big data, por lo que se toma como base primero los elementos necesarios para el procesamiento de datos como son la entrada de los datos, la captura, la depuración, la integridad, la transformación, traducción, resumen, validación modelado, análisis, visualización e interpretación, por otro lado, el enfoque de la metodología MAMBO centrado en meditar sobre el negocio, adquirir los datos, manejar los datos, buscar en los datos y ordenar y visualizar planteado por (Vargas Neira, 2021) dio ideas adicionales para el planteamiento de modelo propuesto.

Además, este modelo se basa en las Arquitecturas de Big data productivas en donde su estructura consta de orígenes de datos, correspondiente a las fuentes de donde surgen los datos, el procesamiento en batch y el procesamiento realtime, y finalmente la visualización y el reporte de los resultados, puntos importantes dentro de todo procesamiento de datos.

También como lo manifiesta (Vargas Guzmán, Moreno Cadena, Oñate Escalante, & Sanabria Hivon, 2020) el almacenamiento de datos es prioritario por lo que es necesaria una arquitectura e infraestructura de almacenamiento de los datos.

Por otro lado los **entornos** de manipulación de datos como lo menciona (Guzman Ponce, Valdovinos Rosas, Marcial Romero, & Alejo Eleuterio, 2018) quien establece 3 diferentes entornos enfocados en el tipo de fuente de dato, el modelo de programación y los diferentes lenguajes de programación útiles para el procesamiento, como son: Apache Hadoop, Apache Spark y Apache Flink al igual como lo menciona (Gómez Degraves, 2021) se utilizan estos entornos para datos distribuidos, por lo que el modelo planteado extrae lo mejor de estos entornos.

El modelo también utiliza el aprendizaje automático basado en un conjunto de algoritmos orientados a problemas de clasificación entre los que destacan Random Forest, Árboles de decisión, Redes neuronales, Regresión logística, Máquina de vectores de soporte (SVM), clasificadores de KN vecinos, como lo manifiesta

(Sandoval, 2018), esto permite tener una amplia variedad de algoritmos a utilizar de acuerdo a la problemática que se intenta solucionar. De igual como lo manifiestan diversos autores como (Luna Perejón, 2020) (Espinoza Montalvo, 2019) (Tepepa Cantero, Pérez Meana, & Nakano Miyatake, 2018) (Brusil Cruz, 2020) se pueden clasificar estos algoritmos como:

- Redes Neuronales. Abordan problemas multiclase, es decir son capaces de modelarse y devolver un resultado por cada clase evaluada dentro de los datos.
- SVM. Siendo un clasificador multiclase, busca optimizar el proceso de clasificación y/o regresión a través de la separación de las clases maximizando el margen los puntos más cercanos al hiperplano y la línea.
- Random Forest. Se utilizan gran cantidad de árboles de decisión, pero éstos operan como un solo conjunto. Cada árbol devuelve una predicción de clase, por lo que la clase que tiene más votos se convierte en la clase predictora del modelo.
- Regresión Logística. Para esta técnica el objetivo consiste en modelar como influyen las variables regresoras en la probabilidad de ocurrencia de un suceso particular.
- Árboles de decisión. Es un método de aproximación de una función objetivo de valores discretos en el cual la función objetivo es representada mediante un árbol de decisión.

El presente trabajo se basa en las metodologías de procesamiento de datos, los entornos y arquitecturas de la big data, integrando el uso del aprendizaje automático. Todo esto integrando al usuario y la infraestructura tecnológica.

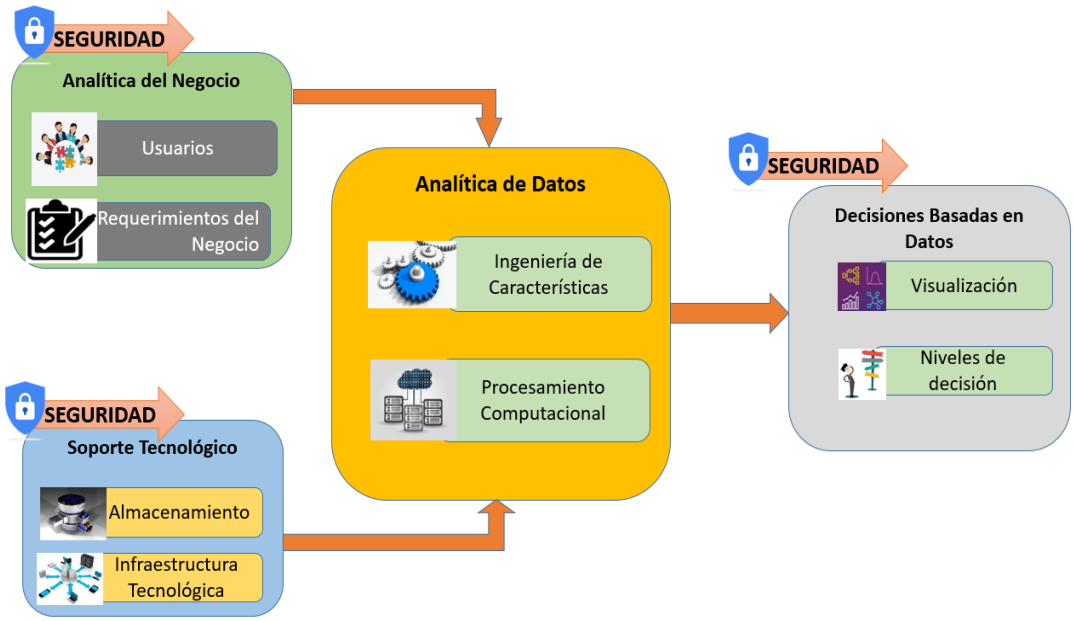
3.4.2. Descripción argumentativa del aporte teórico

En el presente trabajo se desarrolló un Modelo predictivo aplicado al sector académico, que 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.

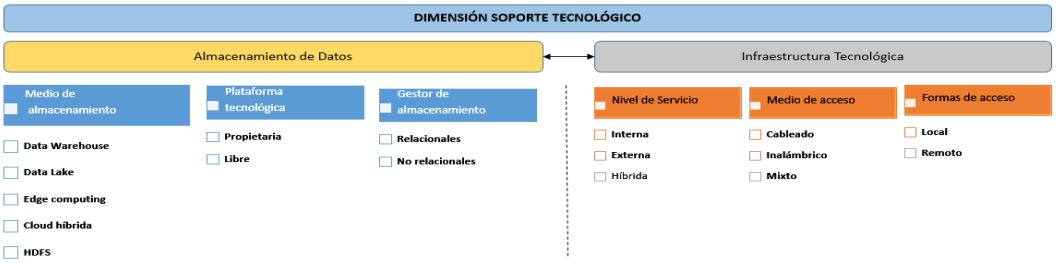
Figura 25 Modelo predictivo de procesamiento de datos en la big data



Fuente: Elaboración propia

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 tecnológico que permita el procesamiento de grandes volúmenes de información.

Figura 26 Dimensión Soporte Tecnológico



Fuente: Elaboración propia

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, acceso confiable y adecuado:

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 corresponden a las características que se almacenan 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 entra 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 sistema analítico propuesto y los medios de acceder a éstos. Y se puede clasificar en bases 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:**

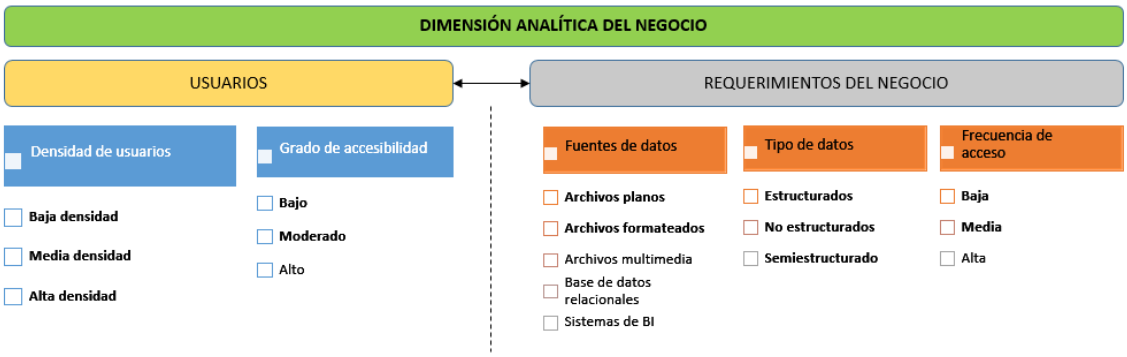
- **Local**: Acceso a los datos a través de los equipos dentro de la misma institución.
- **Remoto**: acceso a los datos desde ubicaciones remotas (por ejemplo, el trabajo remoto)

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



Fuente: Elaboración propia

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 modelos, acceder a los 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 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

Fuente: Elaboración propia

- **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

Fuente: Elaboración propia

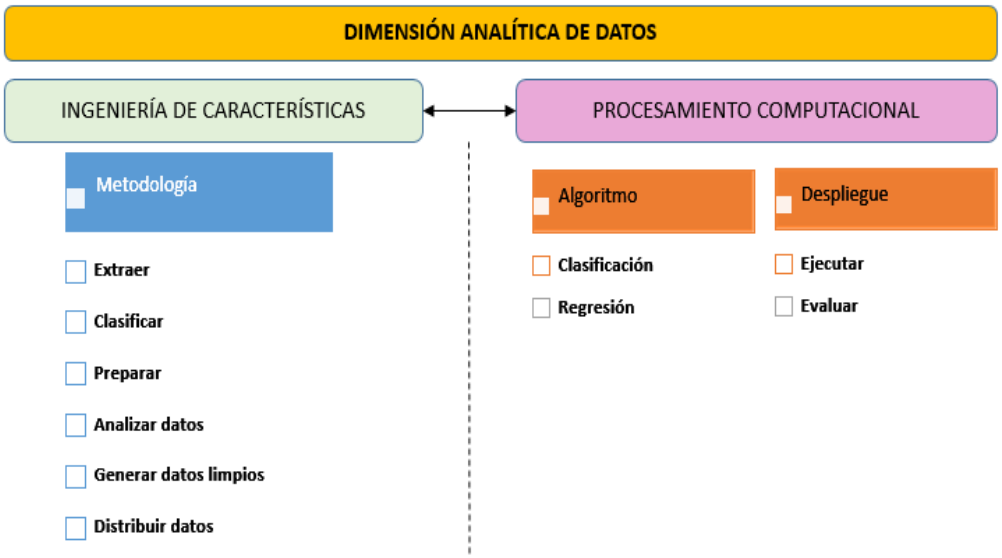
- **Frecuencia de acceso**: Tiempo con la que se realizan solicitudes de datos y accesos al sistema. Se clasifican en:
    - **Baja**: 4 accesos al mes
    - **Media**: 12 accesos al mes
    - **Alta**: 20 accesos al mes

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



Fuente: Elaboración propia

Teniendo en consideración el primer eje correspondiendo a la **ingeniería de características** se propone una metodología de seis pasos cuyo objetivo identificar las características relevantes para su extracción optimizando tiempo 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

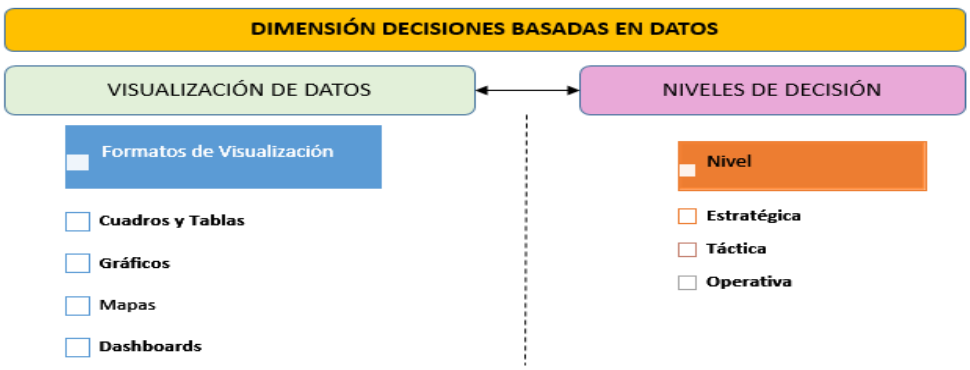
Fuente: Elaboración propia

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

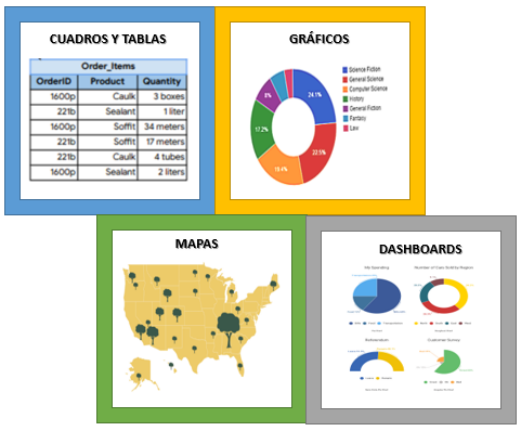


Fuente: Elaboración propia

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. Éstos **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



Fuente: Elaboración propia

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.

3.5. Aporte práctico

El aporte práctico de la presente investigación es un Sistema Analítico basado en el uso de un Modelo predictivo que tenga en cuenta las técnicas predictivas integradas y los grandes volúmenes de datos con la finalidad de mejorar el procesamiento de los datos en gran escala.

3.5.1. Fundamentación del aporte práctico

El **sistema analítico** propuesto toma en cuenta la teoría de las arquitecturas de procesamiento conocidas en la big data en donde se centran primero en los orígenes de datos, luego el procesamiento en batch o procesamiento en Streaming, al final la visualización y el reporting.

Los entornos de procesamiento de datos han evolucionado, entre los que resaltamos a la tecnología de código abierto Apache Hadoop, el cual puede ser escalable y distribuido, también tenemos a Apache Spark basado en el mejoramiento del rendimiento de memoria y Apache Flink de código abierto, manipula datos en tiempo real como también por lotes.

Además, se ha tomado en cuenta, el conocimiento de las técnicas de procesamiento aplicadas a la gestión de los datos, como la manipulación de texto, audio, video u redes sociales.

Por otro lado, el sistema analítico propuesto, se basa en el modelamiento y análisis de grandes volúmenes de datos, lo que conlleva a la utilización de metodologías de análisis espacial, análisis de redes, aprendizaje automático, pruebas A/B y visualización analítica de datos.

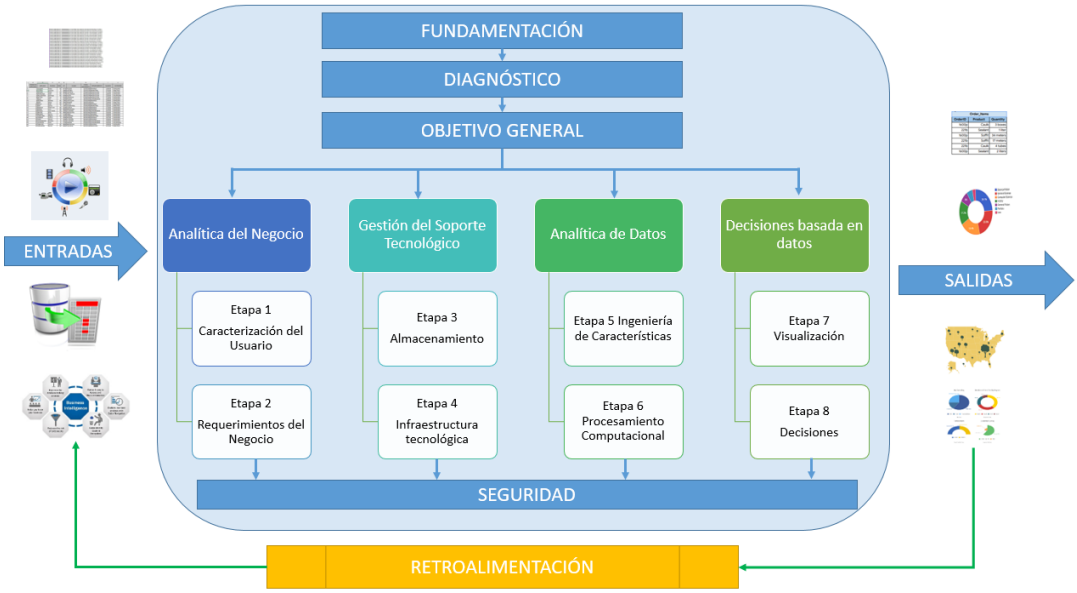
Cabe mencionar, que todo **sistema** se compone de entradas, procesos y salidas, por lo que la propuesta desarrollada se basa en ese enfoque, además de tomar en cuenta el modelo predictivo desarrollado en el aporte teórico de la presente tesis.

3.5.2. Construcción del aporte práctico

El **sistema analítico** se define como sistemas informáticos que las instituciones usan para observar y estudiar datos históricos, generando nuevo conocimiento en base a la detección de patrones y otros comportamientos.

Este sistema analítico usa el modelo predictivo desarrollado en el aporte teórico, el cual se divide en 4 dimensiones: la de soporte tecnológico, analítica del negocio, analítica de datos y las decisiones basadas en datos. Está centrado en el análisis de datos académicos pudiéndose adaptar en el futuro a otros contextos.

Figura 34 Sistema Analítico basado en un Modelo Predictivo de procesamiento de datos



Fuente: Elaboración propia

El desarrollo del sistema analítico tiene como fundamento el uso del procesamiento de datos los cuales han ido evolucionando con el paso del tiempo, pasando de procesos unipersonales a procesamiento en servidores y en múltiples equipos compartidos, además, la gestión de los datos es primordial en todo sistema, éste empieza con la adquisición de los datos, limpieza, procesamiento, y el resultado final a través de la presentación de manera ordenada y fácil de entender para los usuarios.

En el diagnóstico de la problemática se detectó que el volumen de datos se va incrementando constantemente cada año sin que exista un uso claro de los datos almacenados que de alguna manera apoyen a la mejora de los procesos académicos, además, la frecuencia de recopilación de los datos se mantiene constante semestre por semestre. La falta de datos o tener datos erróneos influyen en la calidad de los mismo e impiden un correcto tratamiento de los datos a futuro. Tener un alto porcentaje de exactitud con respecto a los datos procesados teniendo una cercanía al valor experimental obtenido. Además, la sensibilidad de los datos nos brinda una capacidad para detectar datos correctamente, la precisión de los datos permite generar mejores modelos predictivos. Por otro lado, el tiempo de respuesta y la rapidez de acceso nos permiten evaluar el rendimiento de los datos.

Tiene como **objetivo general** mejorar el procesamiento de datos basado en el uso de técnicas predictivas y grandes volúmenes de datos.

El sistema analítico requiere como **entrada** aquellos datos que se encuentran en formatos múltiples como archivos planos, archivos formateados, Base de datos, Sistemas de SI, los cuales se procesan a través de 8 etapas y genera como **salida** la visualización de los datos procesados haciendo uso de tablas, gráficos, mapas, dashboard, entre otros.

Además, podemos mencionar aquí que la **seguridad** es considerada como un elemento transversal a este proceso y que forma parte de cada una de las etapas.

El sistema analítico se **retroalimenta** en base a los resultados obtenidos de tal manera que permitirá mejorar la predicción de los datos en el contexto académico.

Etapas consideradas en el Sistema Analítico:

Etapa 1: Caracterización del usuario

Esta etapa tiene como objetivo realizar la identificación de los perfiles de usuario que permitan establecer claramente los requerimientos de información. Uno de los puntos clave en todo sistema es que el usuario conozca lo que quiere procesar. Dentro de esta etapa se debe establecer la densidad de los mismos, teniendo 3 posibilidades: densidad baja, media y alta. Por otro lado, se debe determinar el grado de accesibilidad que tendrá el usuario al sistema, el cual lo podemos catalogar como bajo, moderado o alto. Además, siendo la seguridad transversal en este modelo, es

que se debe definir los criterios de seguridad basado en los perfiles de usuarios previamente definidos.

Etapa 2: Requerimientos del Negocio

Esta etapa tiene como objetivo identificar las fuentes de donde provienen los datos y establecer los requerimientos del negocio de manera precisa, el cual permitirá posteriormente determinar las características relevantes a considerar en el sistema. Dentro de esta etapa se identifican las fuentes de datos, que pueden ser múltiples como archivos planos, formateado, videos, bases de datos u otros y están almacenados en diversos formatos. Además, se deben establecer los tipos de datos a analizar de acuerdo a la problemática planteada inicialmente y la frecuencia de acceso a los datos.

En esta etapa es muy importante reconocer los tipos de datos que se van a manipular contemplados entre estructurados, no estructurados o semi estructurados. Y también, la frecuencia de acceso pudiendo ser baja, media o alta.

Continuando con el enfoque transversal de la seguridad, aquí se debe considerar el uso de roles o perfiles, realizar un monitoreo de la calidad de los datos y también establecer el cifrado de los mismos.

Etapa 3: Almacenamiento

El objetivo de esta etapa es el alojamiento de los datos para que puedan ser utilizados posteriormente, se relaciona con dos procesos específicos, el primero la escritura de los datos comúnmente conocido como registro de datos y la lectura de los mismos para su futuro procesamiento. Aquí primero debemos establecer el medio de almacenamiento, luego definir la plataforma tecnológica y finalmente seleccionar el gestor de almacenamiento.

Continuando con el enfoque transversal de la seguridad, aquí debemos evitar el acceso no autorizado a los datos, estableciendo controles de acceso y copias de respaldo de los datos almacenados.

Etapa 4: Infraestructura tecnológica

El objetivo de esta etapa es brindar el soporte tecnológico a la institución permitiendo contar con la infraestructura tecnológica necesaria para el sistema analítico y éste

brinde las facilidades de comunicación y el intercambio de información orientados al logro de los objetivos organizacionales. Aquí, primero empezamos con definir los tipos de servicios pudiendo ser: interna, externa e híbrida. Los medios de acceso como el cableado, inalámbrico o mixto. Y la forma de acceso, el cual puede ser local o remoto.

Desde el enfoque transversal de la seguridad, aquí debemos considerar los niveles como la seguridad del perímetro, seguridad de las instalaciones, seguridad de la sala de ordenadores, seguridad a nivel de racks.

Etapa 5: Ingeniería de características

El objetivo de esta etapa es establecer la metodología que permita a través de una serie de pasos obtener los datos de acuerdo con los requerimientos establecidos inicialmente y que estos queden preparados para su procesamiento posterior.

Los pasos son:

1) Generar la extracción de datos necesarios, este paso es prioritario, permite realizar una identificación de los datos que son requeridos, y se debe de establecer qué tipo de datos son, así como también cuales son las fuentes donde están almacenadas para su recolección y futuro tratamiento.

2) Clasificar, corresponde a la selección más relevante de las características de los datos que servirán para procesarla posteriormente a través de un análisis futuro.

3) Preparar, mediante la transformación de los datos de diferentes fuentes y tipos a un formato estándar para el tratamiento.

4) Analizar datos, evaluación previa de los datos obtenidos de la etapa anterior, que permita identificar posibles valores fuera de rangos o vacíos que impidan realizar un buen análisis.

5) Generar datos limpios, después de analizar los datos extraídos, es necesario sólo quedarse con los datos correctos, por lo que se debe aplicar técnicas que permitan cumplir con este objetivo.

6) Distribuir datos, separar bloques de datos que puedan ser asignados a diversos procesadores en paralelo o distribuido que permitan realizar el procesamiento mucho más rápido.

Etapa 6: Procesamiento computacional

Esta etapa tiene como objetivo convertir los datos de entrada, en información útil para la toma de decisiones, que apoyen los procesos académicos de la institución. Esta información de salida puede ser representada de manera visual, ya que esto permite entender mejor los datos que se muestran. Entre las actividades que se ven en esta etapa son: la selección del algoritmo predictivo a utilizar con los datos limpios generados en la etapa anterior y la ejecución de dicho algoritmo sobre los datos, que permita al sistema generar resultados confiables que puedan servir para la toma de decisiones futuras.

Etapa 7: Visualización

Esta etapa tiene como objetivo comunicar información de una forma clara, eficiente y precisa, permitiendo a los usuarios realizar un proceso de análisis y razonamiento sobre los datos. Para esto se debe hacer uso de un formato visual, ya que permite observar datos fáciles de entender para los usuarios y mostrar la información según los perfiles de usuario definidos en la primera etapa.

Desde el enfoque transversal de la seguridad, aquí debemos considerar los controles de acceso a los datos.

Etapa 8: Toma de Decisiones

Esta última etapa tiene como objetivo establecer el nivel de decisión sobre los resultados obtenidos de todo el proceso, estos niveles pueden ser operativos, tácticos o estratégicos, que permita a la institución tomar acciones proactivas en beneficio de los estudiantes y la comunidad universitaria.

Desde el enfoque transversal de la seguridad, aquí debemos considerar definir los riesgos en la toma de decisiones.

3.6. Implementación del Sistema Analítico basado en el Modelo predictivo

La implementación del Sistema Analítico se desarrolla en base a las etapas descritas en el aporte práctico. Para su implementación se emplea el lenguaje de Programación Python el cual permite desarrollar aplicaciones de cualquier tipo, siendo un lenguaje multiplataforma y de código abierto nos permite trabajar sin restricciones. Además, se programará en un cuaderno creado a través de la aplicación web de código abierto

Jupyter Notebook para la creación y compartición de documentos, el cual es muy utilizado en la ciencia de datos. La instalación se realizó utilizando la distribución de Anaconda.

Etapa 1: Caracterización del Usuario

El perfil del usuario correspondería a los docentes de la universidad, por lo que se consideraría una densidad media ya que se cuenta con un promedio de 700 docentes y un grado de accesibilidad moderado, los cuales podrían utilizar los resultados de este estudio en beneficio de sus actividades académicas.

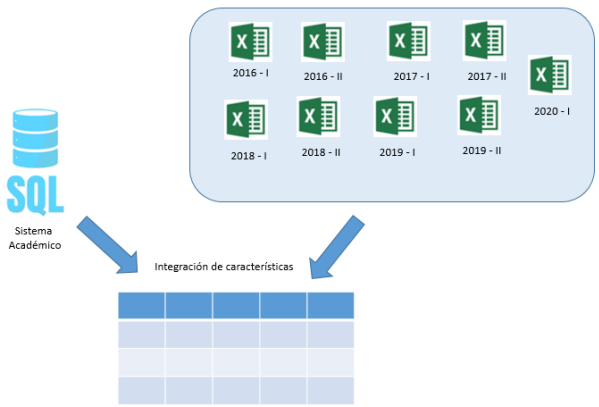
Etapa 2: Requerimiento del Negocio

En esta etapa se establece como requerimiento el análisis del rendimiento académico de los estudiantes universitarios. Por lo que se tienen que identificar las fuentes de datos.

Fuentes de Datos: Base de datos académica implementada en SQL Server 2000 proveída por la Oficina de Servicios Académicos, que contiene 68 tablas siendo las principales las correspondientes a ***Alumno, datosalumno, matricula, detallematricula, escuela, facultad, curso, promedioponderado***. De las cuales se ha extraído parte de la información a procesar a través de sentencias SQL programadas.

Además, se ha manipulado la información de ingresantes a través de 15 archivos en Excel de los semestres 2016-I al 2020-I proveído por la Oficina de Admisión de la UNPRG. Se han integrado en un solo archivo como se aprecia en la figura 35.

Figura 35 Datos de diversas fuentes



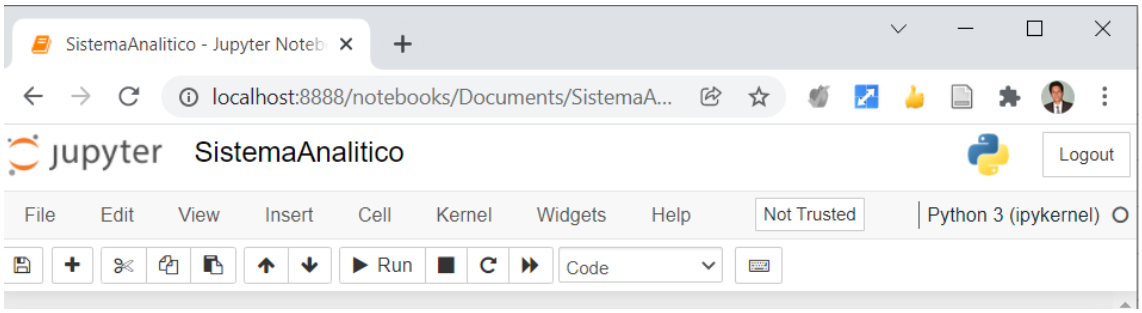
Etapa 3: Almacenamiento

Para el almacenamiento se ha trabajado con Archivos formateados en formato Excel y la base de datos relacional SQL Server 2000, Se ha integrado la información y se ha construido un archivo con formato csv para poder procesarlo en Jupyter notebook.

Etapa 4: Infraestructura tecnológica

La institución cuenta con una infraestructura incipiente, con servidores alojados en la nube, se ha trabajado remotamente con los accesos respectivos como se puede observar en la figura 36.

Figura 36 Trabajo local sobre la web



Etapa 5: Ingeniería de característica

Esta etapa contiene los siguientes pasos:

Paso 1: Generar extracción de datos

Como se muestra en la Figura 37, se logra visualizar los 9869 registros integrados y con 29 características presentes.

Figura 37 Carga de datos

```
import pandas as pd
import numpy as np

nombrearchivo = 'rendimiento.csv'

datos = pd.read_csv(nombrearchivo,index_col="id")

datos.shape
```

```
(9869, 29)
```

A continuación, visualizamos en la Figura 38 los tipos de datos correspondientes a cada una de las características establecidas. Se puede apreciar que 6 son de tipo entero (int64), 5 son de tipo cadenas (object) y 18 son de tipo float64 (reales).

Figura 38 Características y su tipo de dato

```
datos.dtypes
```

```
codigoEscuela                 int64
DescripcionEsc               object
IdModalidad                   int64
DescripcionMod               object
Genero                        int64
Fechanac                     object
Edad_Ingreso                float64
TipoColegio                   int64
Ubigeo                        int64
CodigoDpto                    int64
Departamento                 object
Sem_Ingreso                  object
SemestresEst                float64
PromPonGeneral              float64
Totalcursosllevados         float64
Totalcursosaprobados        float64
Totalcursosdesaprobados     float64
C1_PromSem                  float64
C1_totalCursos              float64
C1_CursosApro               float64
C1_CursosDesap              float64
C2_PromSem                  float64
C2_totalCursos              float64
C2_CursosApro               float64
C2_CursosDesap              float64
RendProm                    float64
RendCursos                  float64
RendPromC1                  float64
RendCursosC1                float64
dtype: object
```

Se muestra a continuación un conjunto de estadísticas descriptivas sobre las características seleccionadas, las cuales se ven en la figura 39, 40 y 41.

Figura 39 Estadísticas descriptivas sobre las características (1)

```
datosest = datos[["Edad_Ingreso","PromPonGeneral","Totalcursosllevados","Totalcursosaprobados","Totalcursosdesaprobados"]]
datosest.describe()
```

	Edad_Ingreso	PromPonGeneral	Totalcursosllevados	Totalcursosaprobados	Totalcursosdesaprobados
count	9863.000000	9538.000000	9538.00000	9538.000000	9538.000000
mean	18.906823	13.158003	27.84242	24.989306	2.853114
std	2.195231	2.541330	16.22510	15.449407	4.242928
min	16.000000	0.000000	2.00000	0.000000	0.000000
25%	18.000000	11.988777	14.00000	12.000000	0.000000
50%	18.000000	13.575000	27.00000	23.000000	1.000000
75%	20.000000	14.896000	40.00000	37.000000	4.000000
max	50.000000	18.250000	76.00000	71.000000	30.000000

Figura 40 Estadísticas descriptivas sobre las características (2)

```
datosest1 = datos[["C1_PromSem","C1_totalCursos","C1_CursosApro","C1_CursosDesap"]]
datosest1.describe()
```

	C1_PromSem	C1_totalCursos	C1_CursosApro	C1_CursosDesap
count	8333.000000	8333.000000	8333.000000	8333.000000
mean	12.847945	6.258010	5.498620	0.759390
std	2.648314	1.044456	1.771082	1.248262
min	0.000000	1.000000	0.000000	0.000000
25%	11.652000	6.000000	5.000000	0.000000
50%	13.261000	6.000000	6.000000	0.000000
75%	14.667000	7.000000	7.000000	1.000000
max	18.250000	9.000000	9.000000	9.000000

Figura 41 Estadísticas descriptivas sobre las características (3)

```
datosest2 = datos[["C2_PromSem","C2_totalCursos","C2_CursosApro","C2_CursosDesap"]]
datosest2.describe()
```

	C2_PromSem	C2_totalCursos	C2_CursosApro	C2_CursosDesap
count	6349.00000	6349.000000	6349.000000	6349.000000
mean	12.96245	6.009608	5.359112	0.650496
std	2.46240	1.116583	1.635269	1.089522
min	0.00000	1.000000	0.000000	0.000000
25%	11.89500	5.000000	5.000000	0.000000
50%	13.30200	6.000000	6.000000	0.000000
75%	14.54200	7.000000	6.000000	1.000000
max	18.10000	12.000000	10.000000	9.000000

También se especifica la cantidad de datos recopilados del archivo, utilizando para esto la función **count**().

Figura 42 Cantidad de registros por característica

```
datos.count()
```

```
codigoEscuela              9869
DescripcionEsc             9869
IdModalidad                9869
DescripcionMod             9869
Genero                     9869
Fechanac                   9863
Edad_Ingreso               9863
TipoColegio                9869
Ubigeo                     9869
CodigoDpto                 9869
Departamento               9869
Sem_Ingreso                9869
SemestresEst               9538
PromPonGeneral             9538
Totalcursosllevados        9538
Totalcursosaprobados       9538
Totalcursosdesaprobados    9538
C1_PromSem                 8333
C1_totalCursos             8333
C1_CursosApro              8333
C1_CursosDesap             8333
C2_PromSem                 6349
C2_totalCursos             6349
C2_CursosApro              6349
C2_CursosDesap             6349
RendProm                   9538
RendCursos                 9538
RendPromC1                 8333
RendCursosC1               8333
dtype: int64
```

Además, se muestra en porcentajes a los estudiantes etiquetados como 1 como Buen rendimiento y la etiquetada 2 como un mal rendimiento, como se aprecia en la figura 43.

Figura 43 Rendimiento de estudiantes en base a su promedio ponderado

```
datos["RendProm"].value_counts().plot(kind="pie",
    autopct='%.2f%%' ,title="Estudiantes según rendimiento en base a promedio ponderado")
plt.legend()
```

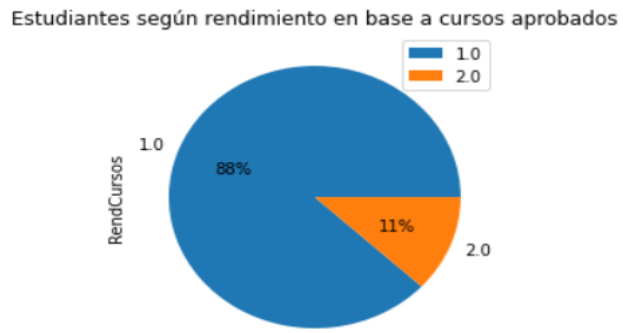
```
<matplotlib.legend.Legend at 0x29c3d3f3d30>
```

También, se muestra en figura 44 los porcentajes de estudiantes etiquetados como 1 con un **buen rendimiento** y los etiquetados con 2 como un **mal rendimiento**.

Figura 44 Rendimiento de estudiantes en base a sus cursos aprobados

```
datos["RendCursos"].value_counts().plot(kind="pie",
    autopct='%d%%' ,title="Estudiantes según rendimiento en base a cursos aprobados")
plt.legend()
```

```
<matplotlib.legend.Legend at 0x29c3d3db220>
```



Otra característica como se muestra en la figura 45 es el ***Genero*** en donde 1 representa a los hombres y 2 a las mujeres. Hay más hombres que mujeres en los datos que se están manipulando.

Figura 45 Cantidad de estudiantes según género

```
import matplotlib.pyplot as plt
datos["Genero"].value_counts().plot(kind="pie",
    autopct='%.2f%%' ,title="Estudiantes por género")
plt.legend()
```

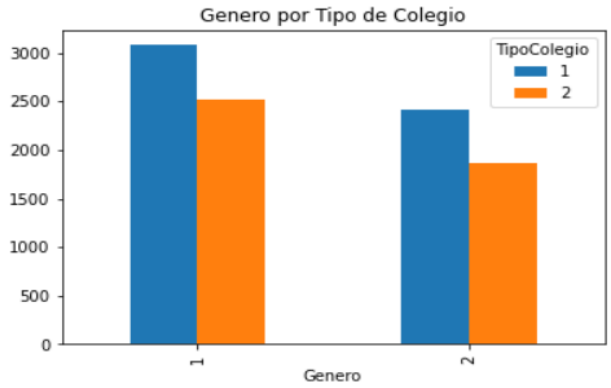
```
<matplotlib.legend.Legend at 0x29c3b9cca60>
```

Además, se muestra en la figura 46 una comparativa entre el género y el tipo de colegio de procedencia de los estudiantes al momento de ingresar a la universidad. Siendo en 1 el tipo de colegio nacional y 2 de colegio particular.

Figura 46 Cantidad estudiantes por Genero y Tipo de Colegio de procedencia

```
genero_tipocol = datos.groupby(['TipoColegio','Genero']).size()
genero_tipocol = genero_tipocol.reset_index()
genero_tipocol = pd.pivot_table(genero_tipocol,
                                columns='TipoColegio', index='Genero', values=0)
genero_tipocol.plot(kind='bar', title='Genero por Tipo de Colegio')
```

```
<AxesSubplot:title={'center':'Genero por Tipo de Colegio'}, xlabel='Genero'>
```

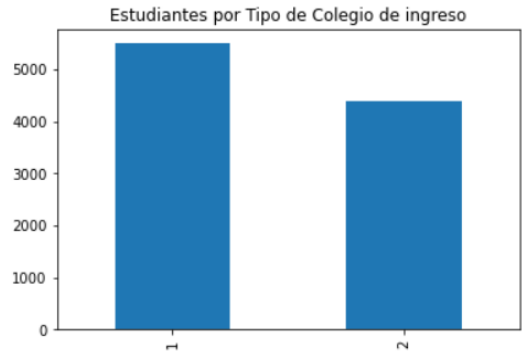


En la figura 47 se aprecia la cantidad de estudiantes de acuerdo al tipo de colegio de procedencia a su ingreso a la universidad, 1 representa a colegio nacional y 2 a colegio particular.

Figura 47 Cantidad de estudiantes por tipo de colegio de procedencia

```
datos["TipoColegio"].value_counts().plot(kind="bar", title="Estudiantes por Tipo de Colegio de ingreso")
```

```
<AxesSubplot:title={'center':'Estudiantes por Tipo de Colegio de ingreso'}>
```

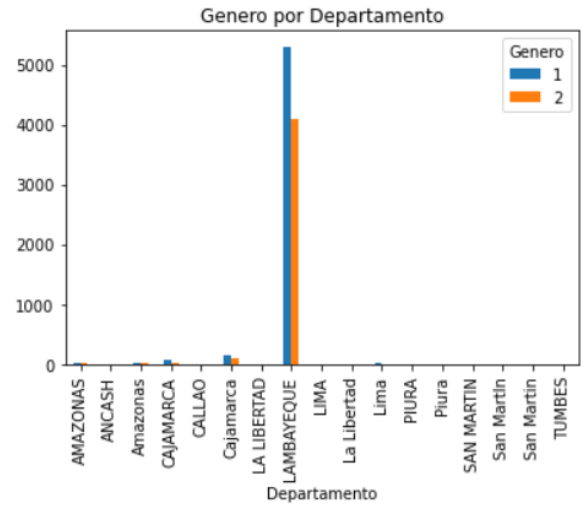


En la figura 48 se observa la cantidad de estudiantes de acuerdo al género por departamento de procedencia, la mayor cantidad de ingresantes son del departamento de Lambayeque.

Figura 48 Cantidad de estudiantes por Género y departamento

```
dpto_genero = datos.groupby(['Genero','Departamento']).size()
dpto_genero = dpto_genero.reset_index()
dpto_genero = pd.pivot_table(dpto_genero, columns='Genero', index='Departamento', values=0)
dpto_genero.plot(kind='bar', title='Genero por Departamento')
#dpto_genero
```

```
<AxesSubplot:title={'center':'Genero por Departamento'}, xlabel='Departamento'>
```



Paso 2: Clasificar

Utilizamos Random Forest para evaluar la importancia de las características identificadas como se puede apreciar en la figura 49, entre las características con importancia alta están: el promedio ponderado general, el promedio semestral del

primer ciclo, el promedio ponderado del segundo ciclo, la cantidad de cursos desaprobados del primer ciclo, el total de cursos desaprobados, la cantidad de cursos aprobados del primer ciclo, el total de cursos desaprobados y los totales de cursos tanto del primer ciclo como del segundo ciclo, por otro lado, se observa que ciertas características presenta muy baja importancia como el género, departamento, tipo de colegio de procedencia, la modalidad de ingreso, la edad de ingreso y el código de escuela.

Figura 49 Importancia de los predictores usando Random Forest

```
importancia_predictores = pd.DataFrame(
                            {'predictor': datoslimpios.drop(columns = "RendProm").columns,
                             'importancia': modelo.feature_importances_}
                            )
print("Importancia de los predictores en el modelo")
print("-------------------------------------------")
importancia_predictores.sort_values('importancia', ascending=False)
```

```
Importancia de los predictores en el modelo
-------------------------------------------
```

	predictor	importancia
6	PromPonGeneral	0.754611
10	C1_PromSem	0.120715
14	C2_PromSem	0.045391
13	C1_CursosDesap	0.024992
8	totalcursosaprobados	0.019938
12	C1_CursosApro	0.016091
9	Totalcursosdesaprobados	0.013040
7	Totalcursosllevados	0.001908
17	C2_CursosDesap	0.001508
16	C2_CursosApro	0.001256
0	codigoEscuela	0.000164
3	Edad_Ingreso	0.000154
1	IdModalidad	0.000080
11	C1_TotalCursos	0.000077
15	C2_totalCursos	0.000028
4	TipoColegio	0.000024
5	CodigoDpto	0.000014
2	Genero	0.000010

Paso 3: Preparar

De acuerdo a la evaluación de la importancia de los predictores, se toma en consideración la lista de características con mayor significancia para una mejor predicción. Por lo que se seleccionan estas características como se aprecia en la figura 50.

Figura 50 Preparación de datos con características de mayor importancia

```
# preparando los datos

datospreparados = datos[["PromPonGeneral","C1_PromSem","C2_PromSem","C1_CursosDesap",
                         "totalcursosaprobados","C1_CursosApro","Totalcursosdesaprobados",
                         "Totalcursosllevados","C2_CursosDesap","C2_CursosApro","RendProm"]]

datospreparadosylimpios = datospreparados.dropna()
datospreparadosylimpios
```

id	PromPonGeneral	C1_PromSem	C2_PromSem	C1_CursosDesap	totalcursosaprobados	C1_CursosApro	Totalcursosdesaprobados	Totalcursosllevados	C
A0001	13.433222	12.810	13.238	0.0	55.0	6.0	2.0	57.0	
A0002	13.471888	12.667	12.143	0.0	55.0	6.0	0.0	55.0	
A0003	14.678000	14.667	14.000	0.0	55.0	6.0	1.0	56.0	
A0004	14.253555	13.143	13.952	0.0	55.0	6.0	0.0	55.0	
A0005	13.123000	13.762	11.524	0.0	55.0	6.0	3.0	58.0	
...	...	...	...	...	...	...	...	...	
A6973	14.238000	14.130	14.440	0.0	18.0	7.0	0.0	18.0	
A6974	16.997666	16.522	17.960	0.0	18.0	7.0	0.0	18.0	
A6975	15.486000	15.522	15.520	0.0	18.0	7.0	0.0	18.0	
A6976	15.388000	15.043	15.960	0.0	18.0	7.0	0.0	18.0	
A6977	15.646000	15.217	16.360	0.0	18.0	7.0	0.0	18.0	

5481 rows × 11 columns

Paso 4: Analizar

Para el análisis se utilizó la correlación de pearson, como se muestra en la figura 51. Si el valor es positivo indica que las características se correlacionan directamente, si es negativo indica que las características se relacionan inversamente y si el caso es 0 indica que no es posible determinar una covariación entre las características.

Figura 51 Correlación de Pearson para las características

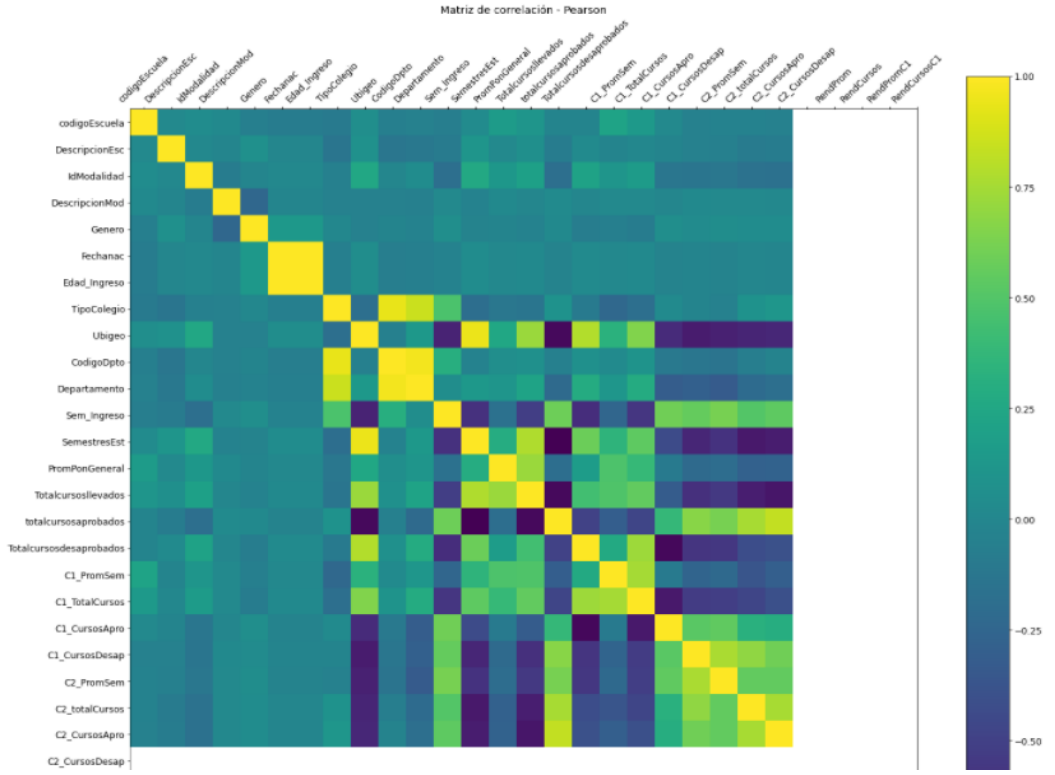
```
datoslimpios.corr("pearson")
```

	codigoEscuela	IdModalidad	Genero	Edad_Ingreso	TipoColegio	CodigoDpto	PromPonGeneral	To
codigoEscuela	1.000000	0.039335	-0.009100	-0.008624	-0.041055	-0.129172	-0.031488	
IdModalidad	0.039335	1.000000	0.013638	-0.014307	0.069599	-0.023958	0.070722	
Genero	-0.009100	0.013638	1.000000	-0.072574	-0.007432	0.024563	0.282068	
Edad_Ingreso	-0.008624	-0.014307	-0.072574	1.000000	-0.267453	-0.038929	-0.031404	
TipoColegio	-0.041055	0.069599	-0.007432	-0.267453	1.000000	0.158687	-0.052156	
CodigoDpto	-0.129172	-0.023958	0.024563	-0.038929	0.158687	1.000000	0.016951	
PromPonGeneral	-0.031488	0.070722	0.282068	-0.031404	-0.052156	0.016951	1.000000	
Totalcursosllevados	0.067272	0.024033	0.030086	0.009553	-0.050796	-0.056544	0.214500	
totalcursosaprobados	0.061034	0.039692	0.099795	-0.005717	-0.070370	-0.050284	0.461522	
Totalcursosdesaprobados	0.010878	-0.046968	-0.201251	0.042438	0.060826	-0.012034	-0.724152	
C1_PromSem	0.012861	0.072523	0.295499	-0.022859	-0.049063	0.004953	0.913120	
C1_TotalCursos	0.144673	0.019676	0.114808	0.025237	-0.042125	-0.007230	0.276944	
C1_CursosApro	0.099915	0.057422	0.208574	-0.005043	-0.054379	-0.011577	0.692898	
C1_CursosDesap	-0.022345	-0.066127	-0.203805	0.028445	0.042901	0.010593	-0.764746	
C2_PromSem	-0.082863	0.053998	0.199774	-0.003520	-0.056715	0.023290	0.760718	
C2_totalCursos	0.166870	0.021830	0.087755	0.004579	-0.030255	-0.003282	0.289869	
C2_CursosApro	0.068190	0.037429	0.146286	0.011325	-0.061090	0.008859	0.616379	
C2_CursosDesap	0.060937	-0.032815	-0.125943	-0.011853	0.058666	-0.015793	-0.606546	
RendProm	-0.003846	-0.042014	-0.138106	0.011302	0.047797	-0.000210	-0.685305	

Además, se presenta en la figura 52 el mapa de calor según los resultados obtenidos de la correlación de pearson.

Figura 52 Mapa de calor para la Correlación de Pearson

```
# mapa de calor para la correlacion de pearson

import matplotlib.pyplot as plt

f = plt.figure(figsize=(23,20))
plt.matshow(datos.corr(), fignum=f.number)
plt.xticks(range(datos.shape[1]), datos.columns, fontsize=12, rotation=45)
plt.yticks(range(datos.shape[1]), datos.columns, fontsize=12)
cb = plt.colorbar()
cb.ax.tick_params(labelsize=12)
plt.title('Matriz de correlación - Pearson', fontsize=14)
plt.show()
```



Paso 5: Generar datos limpios

Para la generación de datos limpios se procedió a utilizar la función dropna(), la cual permite eliminar todos los registros que tienen valores nulos, generando una data consistente y sin valores vacíos que mejore las predicciones del modelo. Como se muestra en la figura 53.

Figura 53 Generando datos limpios

```
datoslimpios = datoslimpios.dropna()
```

```
datoslimpios
```

id	codigoEscuela	IdModalidad	Genero	Edad_Ingreso	TipoColegio	CodigoDpto	PromPonGeneral	Totalcursosllevados	totalcursosaprobados	Totalcursosd
A0001	3	1	1	20.0	2	14	13.433222	57.0	55.0	
A0002	3	1	1	19.0	1	6	13.471888	55.0	55.0	
A0003	3	1	2	17.0	2	14	14.678000	56.0	55.0	
A0004	3	1	1	18.0	1	14	14.253555	55.0	55.0	
A0005	3	1	2	19.0	1	14	13.123000	58.0	55.0	
...	...	...	...	...	...	...	...	...	...	
A6973	11	1	2	20.0	2	14	14.238000	18.0	18.0	
A6974	11	1	2	18.0	2	14	16.997666	18.0	18.0	
A6975	11	1	2	18.0	2	14	15.486000	18.0	18.0	
A6976	11	1	2	20.0	2	14	15.388000	18.0	18.0	
A6977	11	1	2	17.0	2	14	15.646000	18.0	18.0	

5479 rows × 19 columns

Paso 6: Distribuir

Como se aprecia en la figura 54, proceso en el cual se distribuyen los datos para el tratamiento en el modelo, se generan dos bloques uno para entrenamiento y otro para pruebas.

Figura 54 Distribución de datos

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt

from sklearn.datasets import load_boston
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.model_selection import RepeatedKFold
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import ParameterGrid
from sklearn.inspection import permutation_importance
import multiprocessing

nombrearchivo = 'rendimiento.csv'

datos = pd.read_csv(nombrearchivo,index_col="id")

datoslimpios = datos[["codigoEscuela", "IdModalidad","Genero","Edad_Ingreso","TipoColegio","CodigoDpto",
                      "PromPonGeneral","Totalcursosllevados","totalcursosaprobados","Totalcursosdesaprobados",
                      "C1_PromSem","C1_TotalCursos","C1_CursosApro","C1_CursosDesap","C2_PromSem","C2_totalCursos",
                      "C2_CursosApro","C2_CursosDesap","RendProm"]]
datoslimpios = datoslimpios.dropna()
datoslimpios

X_train, X_test, y_train, y_test = train_test_split(
                                       datoslimpios.drop(columns = "RendProm"),
                                       datoslimpios['RendProm'],
                                       random_state = 123
                                   )
```

Etapa 6: Procesamiento computacional

En esta fase se selecciona diferentes algoritmos para evaluar el rendimiento académico de los estudiantes a través de los datos limpios obtenidos anteriormente.

Se establece un conjunto de algoritmos para su despliegue y evaluación de resultados como se muestra en la figura 55.

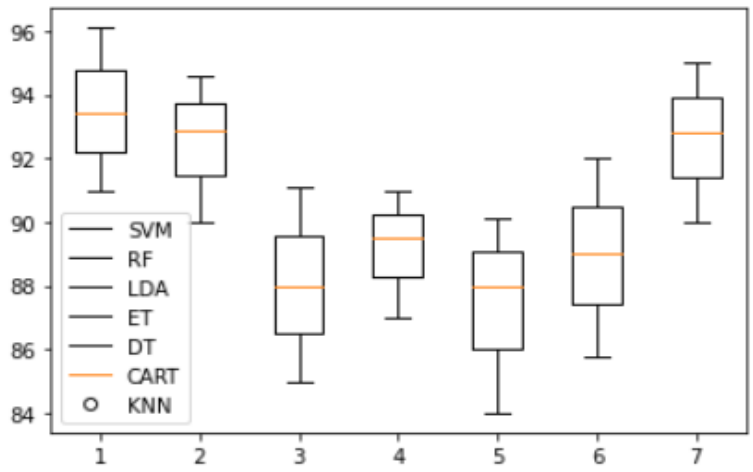
Figura 55 Listado de algoritmos para procesamiento computacional

```
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier

modelosdefinidos = []
modelosdefinidos.append(("SVM", SVC() ))
modelosdefinidos.append(("RF", RandomForestClassifier() ))
modelosdefinidos.append(("LDA", LinearDiscriminantAnalysis() ))
modelosdefinidos.append(("ET", ExtraTreesClassifier() ))
modelosdefinidos.append(("DT", BaggingClassifier(base_estimator=DecisionTreeClassifier()) ))
modelosdefinidos.append(("CART", DecisionTreeClassifier() ))
modelosdefinidos.append(("KNN", KNeighborsClassifier() ))

modelosdefinidos
```

```
[('SVM', SVC()),
 ('RF', RandomForestClassifier()),
 ('LDA', LinearDiscriminantAnalysis()),
 ('ET', ExtraTreesClassifier()),
 ('DT', BaggingClassifier(base_estimator=DecisionTreeClassifier())),
 ('CART', DecisionTreeClassifier()),
 ('KNN', KNeighborsClassifier())]
```

La evaluación de los resultados se realizó utilizando los parámetros por defecto para cada uno de los algoritmos establecidos. Se obtuvo los siguientes resultados para cada uno de los algoritmos como se muestra en la figura 56, en donde los que tuvieron mejor resultado fueron el SVM, RF y KNN.

Figura 56 Evaluación de cada algoritmo



El entrenamiento con los datos segmentados, se aplicó a estos 3 algoritmos identificados anteriormente como se aprecia en la figura 57.

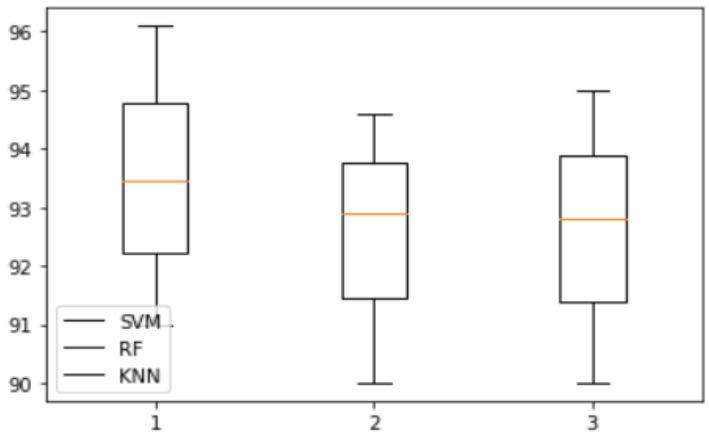
Figura 57 Algoritmos seleccionados para entrenamiento

```
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier

modelosseleccionados = []
modelosseleccionados.append(("SVM", SVC() ))
modelosseleccionados.append(("RF", RandomForestClassifier() ))
modelosseleccionados.append(("KNN", KNeighborsClassifier() ))

modelosseleccionados
```

```
[('SVM', SVC()),
 ('RF', RandomForestClassifier()),
 ('KNN', KNeighborsClassifier())]
```

Luego del entrenamiento se compararon los resultados de las precisiones de cada uno de los algoritmos, según figura 58, en donde la precisión de SVM es de 93.45%, les sigo con el RF con 92.9% y finalmente el KNN con un valor de 92.8%.

Figura 58 Comparación de resultado



Los resultados como se aprecia en la figura anterior establecen que el algoritmo SVM es el mejor de todos. Por lo que su resultado se muestra en la figura 59.

Figura 59 Resultados del algoritmo

	SVM
Accuracy	93.45678
Desv_Stand	1.30564

Se almacena el modelo para su posterior uso, como se aprecia en la figura 60.

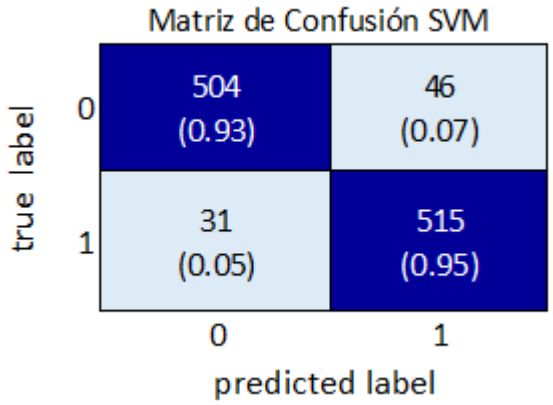
Figura 60 Almacenamiento del Modelo

```
import joblib  jbl

modelo.fit(X_train, y_train)
archivoModelo = 'Modelo/ModeloSVM.pkl'

jbl.dump(modelo, archivoModelo)
```

Etapa 7: Visualización

Se utiliza el conjunto de datos de validación con el modelo obtenido para evaluar su rendimiento, del total de datos procesados 5479 se trabajó con el 80% (4383) para datos de entrenamiento y 20% (1096) para los datos de validación, como se observa en la figura 61.

Figura 61 Matriz de confusión del modelo evaluado



En la tabla 10 se muestra la información obtenida en la matriz de confusión se puede indicar que 504 estudiantes fueron detectados correctamente con mal rendimiento, mientras que 515 estudiantes fueron detectados correctamente con buen rendimiento, además se observa que 46 fueron clasificados incorrectamente, 31 también fueron clasificados incorrectamente.

Tabla 10 Matriz de Confusión

	Clases	Resultados del Sistema Propuesto		
		Buen Rendimiento	**Mal Rendimiento**	**Total**
Rendimiento académico	Buen Rendimiento	504 (VP)	46 (FN)	550
	Mal Rendimiento	31 (FP)	515 (VN)	546

Resultado del Sistema Propuesto se aprecian en la tabla 11. Donde un 92.97% es la proporción de clasificación correcta global del sistema y un 94.21 corresponde a la precisión del sistema al clasificar al estudiante.

Tabla 11 Rendimiento del Sistema Propuesto

Métrica Evaluada	**Fórmula**	**Resultado**
Accuracy	Accuracy = (VP+VN) / (VP+VN+FP+FN)	92.97%
Classification Error	Classification Error = (FP+FN)/(VP+VN+FP+FN)	7.03%
Recall (TVP)	recall = VP / (VP+FN)	91.64%
Specificity (TVN)	Specificity = VN / (VN+FP)	94.32%
Precision	Precision = VP / (VP + FP)	94.21%

Etapa 8: Decisiones

Una vez finalizado la generación del modelo, se puede poner a prueba y evaluar de acuerdo a nuevos datos de los estudiantes con la finalidad de clasificarlos en estudiante con buen rendimiento o estudiante con mal rendimiento y de esta manera en el caso de ser mal rendimiento buscar acciones que beneficien al estudiante como, por ejemplo:

- Reforzar programas de tutoría.
- Programa de apoyo orientado a sus necesidades académicas.

3.7. Valoración y corroboración de los resultados

De los resultados planteados en el diagnóstico contextual, realizado en base a los estudios previos sobre el rendimiento académico se determinó que existe una diferencia en el accuracy de 8.03% lo cual indica una mayor proporción de clasificación correcta global del sistema propuesto con respecto al sistema base.

Se utilizaron un total de 1096 registros académicos de estudiantes clasificados en buen rendimiento y mal rendimiento. Considerando 4383 registros académicos para el entrenamiento del modelo.

El valor recall (TVP) presenta un incremento significativo de 11.31% lo cual indica una correcta clasificación como se aprecia en la tabla 12.

Tabla 12 Comparación de rendimiento

Métrica Evaluada	**Sistema Base**	**Sistema Propuesto**	**Diferencia**
Accuracy	84.95%	92.97%	8.03%
Classification Error	15.05%	7.03%	8.03%
Recall (TVP)	80.32%	91.64%	11.31%
Specificity (TVN)	89.76%	94.32%	4.56%
Precision	89.09%	94.21%	5.12%

Se obtuvo también un TVN del 4.56% lo que nos indica que se mejoró la clasificación correcta de los estudiantes con buen rendimiento. Por último, se obtiene un 5.12% en la precisión del rendimiento de los estudiantes como clasificados correctamente desde el sistema propuesto.

IV. CONCLUSIONES

- Se caracterizó el proceso de procesamiento de datos de acuerdo a los planteamientos conceptuales establecidos en la manipulación de los datos tanto en su captura, integridad, transformación, análisis y visualización.
- Se determinó las tendencias históricas del proceso de procesamiento de datos en base a criterios de recolección de datos, calidad, seguridad, tratamiento, rendimiento y soporte de los datos evidenciándose un alto volumen de datos a procesar además de un pobre rendimiento de los mismos.
- 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 para esto características del rendimiento como los promedio finales y semestrales de los estudiantes. El modelo se compone de 4 dimensiones: la de soporte tecnológico, analítica del negocio, analítica de datos y las decisiones basadas en datos.
- Se elaboró un Sistema Analítico basado en un modelo predictivo para el tratamiento de datos académicos, apoyando el logro de los objetivos trazados.
- Se corroboró los resultados de manera estadística evaluándose los indicadores teniéndose una mejora de 8.03% en el rendimiento global del sistema, un 11.3% de incremento en la clasificación correcta del mal rendimiento académico y un 5.12% de clasificación correcta del buen rendimiento académico.

V. RECOMENDACIONES

- Realizar una recolección de características de los estudiantes no solo en el ámbito académico, sino también 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 de manera clara para poder aplicar el modelo que se ajuste a las necesidades, y que también permita evaluar temas como la deserción estudiantil, influencia de herramientas tecnológicas en la virtualidad, entre otros, influencia del aula virtual en los cursos de los estudiantes, etc.

VI. REFERENCIAS

Maestro Cano, I. C. (2016). Reflexiones epistemológicas sobre Big Data. *Revista de filosofía - eikasia*, 451-474.

Alvarez Valle, J. (11 de 11 de 2013). *«El Big Data: Una Gran Oportunidad»*. Obtenido de http://www.lne.es/opinion/2013/11/10/big-data-una-granoportunidad/1497242.html

Ángeles, M. d. (2021). Sistema de archivos, gestores de base de datos y Sistema de archivos, gestores de base de datos y. *Revista Digital Universitaria (RDU), 22*(6). doi:http://doi.org/10.22201/cuaieed.16076079e.2021.22.6.6

Ayala Franco, E., López Martínez, R. E., & Menéndez Dominguez, V. H. (2021). Modelos predictivos de riesgo académico en carreras de computación con minería de datos educativos. *RED. Revista de Educación a Distancia, 21*(66). doi:http://dx.doi.org/10.6018/red

Britos, L., Di Gennaro, M., Gil Costa, V., Kasian, F., Lobos, J., Ludueña, V., . . . Trabes, G. (2016). Busquedas en Grandes Volúmenes de Datos. *XVIII Workshop de Investigadores en Ciencias de la Computación WICC 2016*, (págs. 283-287). Concordia - Entre Ríos.

Brusil Cruz, C. A. (Febrero de 2020). Análisis comparativo entre aprendizaje supervisado y aprendizaje semi-supervisado para la clasificación de señales sísmicas vulcanológicas del volcán Cotopaxi. Quito.

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

Contreras, L., Fuentes, H., & Rodríguez, J. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria, 13*(5), 233-245. doi:http://dx.doi.org/10.4067/S0718-50062020000500233

Cravero, A., Sepúlveda, S., & Muñoz, L. (2020). Big Data Architectures for the Climate Change Analysis: A Systematic Mapping Study. *IEEE Latin America Transactions*, 1793 - 1806.

De Battista, A., Cristaldo, P., Ramos, L., Nuñez, J. P., Retamar, S., & Bouzenard, D. (2016). Minería de Datos Aplicada a Datos Masivos. *XVIII Workshop de Investigadores en Ciencias de la Computación*, (págs. 288-292). Concordia - Entre Ríos.

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*, 95-109.

Elmasri, R., & Navathe, S. B. (2007). *Fundamentos de Sistemas de Bases de Datos* (Quinta Edición ed.). España: Pearson - Addison Wesley.

Escobar Borja, M., & Mercado Pérez, M. (2019). big data: un analisis documental de su uso y aplicación en el contexto de la era digital. *Revista La Propiedad Inmaterial Nº 28*, 273 - 293.

Espinoza Montalvo, S. (2019). Predicción de postulantes que cometerán fraude interno en una compañía con algoritmos de aprendizaje supervisado. *Interfases*, 49-60.

Flores Vivar, J. M. (2018). Algoritmos, aplicaciones y Big data, nuevos paradigmas en el proceso de comunicación y deenseñanza-aprendizaje del periodismo de datos. *Revista de Comunicación, 17*(2), 268-291. doi:https://doi.org/10.26441/RC17.2-2018-A12

García, S., Ramírez Gallego, S., Luengo, J., & Herrera, F. (julio de 2016). Big Data: Preprocesamiento y calidad de datos. *Big Data: Preprocesamiento y calidad de datos*. Granada, España.

Gómez Degraves, Á. (2021). *Big Data, un sistema de gestión de datos.*

Guzman Ponce, A., Valdovinos Rosas, R. M., Marcial Romero, J. R., & Alejo Eleuterio, R. (2018). Entornos de trabajo para procesamiento de datos masivos y aprendizaje automatico. *Research in Computing Science*, 225–237. Obtenido de https://rcs.cic.ipn.mx/2018_147_5/Entornos%20de%20trabajo%20para%20procesamiento%20de%20datos%20masivos%20y%20aprendizaje%20automatico.pdf

Hernández Leal, E. J., Duque Méndez, N. D., & Moreno Cadavid, J. (2017). Big Data una exploración de investigaciones, tecnologías y casos de aplicación. *TecnoLógicas*, 15-38.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación.* México: Mc Graw Hill.

Luna Perejón, F. (21 de Julio de 2020). Estudio e integración en sistemas empotrados de algoritmos de Aprendizaje Supervisado basados en Redes Neuronales Artificiales para el análisis de perturbaciones y eventos asociados a la marcha. Sevilla.

Malberti, A., Klenzi, R., & Beguerí, G. (2016). 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)*, (págs. 233-237). Entre Ríos.

Menacho Chiok, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos, 78*(1), 26-33. doi:https://doi.org/10.21704/ac.v78i1.811

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. *Revista Cubana de Ciencias Informáticas*, 105-118. Obtenido de

http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992016000300008&lng=es&tlng=es

Orihuela Maita, G. Y. (2019). *Aplicación de Data Science para la Predicción del Rendimiento Académico de los Estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú.* Huancayo: Universidad Nacional del Centro del Perú.

Peñaloza Báez, M. J. (2017). Big data y analítica del aprendizaje en aplicaciones de salud y educación médica. *Revista Investigación en educación médica*, 61-66. doi:https://doi.org/10.1016/j.riem.2017.11.003

Pojon, M. (2017). Using Machine Learning to Predict Student Performance.

Puyol Moreno, J. (2014). Una aproximación a Big Data. *Revista De Derecho De La UNED (RDUNED)*, 471–506.

Quinteros, O. E., Funes, A., & Ahumada, H. C. (2016). 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 - WICC 2016*, (págs. 223-226). Entre Ríos.

Quiroz Martinez, M., Aguilar Duarte, R. A., & Intriago Cedeño, D. (2020). Proceso de diseño de una arquitectura Big Data para el análisis de grandes volúmenes de datos e información. *Opuntia brava*.

Russo, C. C., Charme, J., Piergallini, M. R., Guasch, M. M., Torriggino, A., & Smail, A. (2016). Aplicación de minería de datos espacial en el área de salud en la zona de influencia de la UNNOBA. *XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina)*, (págs. 135-138). Entre Ríos.

Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., & Tessore, J. (2016). Tratamiento masivo de datos utilizando técnicas de machine learning. *XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina)*, 131-134.

Sandoval, L. J. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista Tecnológica*(11), 36-40.

Santos Grueiro, I. (2015). EL TAMAÑO SÍ ES IMPORTANTE. *Revista web de la Facultad de Ingeniería de la Universidad de Deusto*.

Schab, E., Rivera, R., Bracco, L., Coto, F., Cristaldo, P., Ramos, L., . . . De Battista, A. (2018). Minería de Datos y Visualización de Información. *XX Workshop de Investigadores en Ciencias de la Computación - RedUNCI - UNNE*, 325-329.

Silberschatz, A., Korth, H., & Sudarshan, S. (2014). *Fundamentos de Bases de Datos* (Quinta Edición ed.). España: Mc Graw Hill.

Tascón, M., & Coullaut, A. (2016). *Big Data y el internet de las cosas. Qué hay detrás y cómo nos va a cambiar.* Madrid: Catarata.

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

Tepepa Cantero, A., Pérez Meana, H. M., & Nakano Miyatake, M. (2018). Algoritmos de aprendizaje supervisado para la clasificación de géneros musicales caracterizados mediante modelos estadísticos. *Research in Computing Science*, 119-128.

Tolosa, G. H., Banchero, S., Ríssola, E., Delvechio, T., & Feuerstein, E. (2016). Grandes Datos y Algoritmos Eficientes para Búsquedas de Escala Web. *XVIII Workshop de Investigadores en Ciencias de la Computación - WICC 2016 - WICC 2016*, (págs. 227-232). Entre Ríos.

Treviño-Reyes, R., Rivera-Rodríguez, F. S., & Garza-Alonso, J. A. (2020). La analítica de datos como ventaja competitiva en las organizaciones. *Vinculatégica*, 1063-1074.

Vargas Guzmán, W. C., Moreno Cadena, A. G., Oñate Escalante, A. M., & Sanabria Hivon, M. (2020). Importancia del big data en un gestor documental para las entidades públicas de Colombia. *SIGNOS, Investigación en Sistemas de Gestión, 13*(1).

Vargas Neira, A. R. (2021). *Big data analytics aplicada en la integración de datos de internet de las cosas, caso de uso: Agricultura de precisión.*

Vite Cevallos, H., Townsend Valencia, J., & Carvajal Romero, H. (2020). Big Data e internet de las cosas en la producción de banano orgánico. *Universidad y Sociedad vol.12 no.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. *Universidad Nacional de Río Negro*. Obtenido de http://hdl.handle.net/20.500.12049/150

Anexos

Anexo 01: Matriz de consistencia

Anexo 02: Operacionalización de las variables.

Anexo 03: Instrumentos

Anexo 04: Validación de instrumentos por juicio de expertos

Anexo 05: Consentimiento Informado

Anexo 06: Aprobación del Informe de Tesis

ANEXO N° 1 MATRIZ DE CONSISTENCIA

Manifestaciones del problema	– 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.
Problema	El inadecuado procesamiento de los datos usando técnicas predictivas aplicados a grandes volúmenes de datos y el análisis de la big data limita el tratamiento de los datos académicos.
Causas que originan el Problema	– Insuficiente referencia teórica sobre técnicas predictivas aplicadas a la Big data, en el proceso de generación del conocimiento. – Limitaciones de técnicas predictivas que den un soporte para la extracción de conocimiento útil en grandes volúmenes 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.
Objeto de la Investigación	Proceso de Procesamiento de datos en la big data
Objetivo General de la Investigación	Aplicar un **Sistema Analítico** basado en un **modelo predictivo** que tenga en cuenta la relación entre las técnicas predictivas integradas y los grandes volúmenes de datos para mejorar el procesamiento de los datos en la big data.
Objetivos específicos	– Caracterizar epistemológicamente el procesamiento de datos y su dinámica. – Determinar las tendencias históricas del procesamiento de datos y su dinámica. – Diagnosticar el estado actual del procesamiento de datos en la Universidad Nacional Pedro Ruiz Gallo. – Elaborar un modelo predictivo para el tratamiento de datos académicos.

	– Elaborar un sistema analítico basado en un modelo predictivo. – Validar los resultados de la investigación.
Campo de la investigación	Dinámica del proceso de procesamiento de datos en la big data.
Título de la Investigación	Sistema analítico basado en un modelo predictivo de procesamiento de datos en la big data para el sector académico en la educación superior.
Hipótesis	Si se aplica un **Sistema Analítico** basado en un **modelo predictivo** que tenga en cuenta la relación entre las técnicas predictivas integradas y los grandes volúmenes de datos, **entonces** se contribuye a la mejora en el procesamiento de los datos en la big data.
Variables	Variable Independiente: Sistema analítico basado en un modelo predictivo. Variable Dependiente: Procesamiento de datos en la big data.

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 2 OPERACIONALIZACIÓN DE VARIABLES

VARIABLE INDEPENDIENTE	DIMENSIONES	DESCRIIPCIÓN
Sistema analítico basado en un modelo predictivo	Introducción-Fundamentación.	Se establece el contexto y ubicación de la problemática a resolver. Ideas y puntos de partida que fundamentan la estrategia. Se indica la teoría en que se fundamenta el aporte propuesto.
	Diagnóstico	Indica el estado real del objeto y evidencia el problema en torno al cual gira y se desarrolla la estrategia, protocolo, o programa, según el aporte practico a desarrollar.
	Planteamiento del objetivo general.	Se desarrolla el objetivo general del aporte práctico. Se debe tener en cuenta que no es el de la investigación.
	Planeación estratégica	Se definen metas u objetivos a corto y mediano plazo que permiten la transformación del objeto desde su estado real hasta el estado deseado. Planificación por etapas de las acciones, recursos, medios y métodos que corresponden a estos objetivos. Considerando las siguientes dimensiones: Recolección, calidad, seguridad, manipulación y rendimiento.
	Instrumentación	Explicar cómo se aplicará, bajo qué condiciones, durante qué tiempo, responsables, participantes.
	Evaluación	Definición de los logros obstáculos que se han ido venciendo, valoración de la aproximación lograda al estado deseado.

VARIABLE DEPENDIENTE	**DIMENSIONES**	**INDICADORES**	**TÉCNICAS E INSTRUMENTOS DE LA INVESTIGACIÓN**	**FUENTES DE VERIFICACIÓN (FUENTES DE INFORMACIÓN)**
Tratamiento de datos en la big data	Recolección	Volumen de los datos	- Encuesta - Observación - Análisis documental	Cuestionario Usuarios Base de Datos Ficha de datos
		Frecuencia de recopilación		
	Seguridad	Número de incidentes de datos registrados		
		Tiempo medio para resolver un incidente		
	Manipulación	Exactitud		
		Sensibilidad		
		Especificidad		
		Precisión		

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 3 INSTRUMENTO

Guía de Encuesta

Esta encuesta, es dirigida 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 en una institución universitaria.

La información que nos facilite es anónima y la mejor manera de colaborar con nosotros es siendo analítico y veraz en sus respuestas, para que estas reflejen los problemas reales que se afrontan al respecto.

Finalmente queremos agradecerle su disposición a colaborar en este estudio el cual puede ayudar a solucionar los problemas del tratamiento de datos académico.

INSTRUCCIONES:

Al responder este cuestionario debe tener en cuenta lo siguiente:

- Lea detenidamente cada pregunta, antes de contestarla, así como sus posibles respuestas.

- Encontrará una forma fundamental de responder las preguntas.

DIMENSIÓN: RECOLECCIÓN

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

   ( ) Totalmente de acuerdo

   ( ) De acuerdo

   ( ) Indeciso

   ( ) En desacuerdo

   ( ) Totalmente en descuerdo

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

   ( ) Totalmente de acuerdo

   ( ) De acuerdo

   ( ) Indeciso

   ( ) En desacuerdo

   ( ) Totalmente en descuerdo

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

   ( ) Muy frecuentemente

   ( ) Frecuentemente

   ( ) Ocasionalmente

   ( ) Raramente

   ( ) Nunca

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

   ( ) Totalmente de acuerdo

   ( ) De acuerdo

   ( ) Indeciso

   ( ) En desacuerdo

   ( ) Totalmente en descuerdo

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

   ( ) Muy frecuentemente

   ( ) Frecuentemente

   ( ) Ocasionalmente

   ( ) Raramente

   ( ) Nunca

DIMENSIÓN: MANIPULACIÓN

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

   ( ) Muy frecuente

   ( ) Frecuente

   ( ) Ocasional

   ( ) Raramente

   ( ) Nunca

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

   ( ) Totalmente de acuerdo

   ( ) De acuerdo

   ( ) Indeciso

   ( ) En desacuerdo

   ( ) Totalmente en descuerdo

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.

   ( ) Totalmente de acuerdo

   ( ) De acuerdo

   ( ) Indeciso

   ( ) En desacuerdo

   ( ) Totalmente en descuerdo

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

   ( ) totalmente de acuerdo

   ( ) de acuerdo

( ) indeciso

( ) en desacuerdo

( ) totalmente en desacuerdo.

DIMENSIÓN: CALIDAD

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

( ) Muy frecuentemente

( ) Frecuentemente

( ) Ocasionalmente

( ) Raramente

( ) Nunca

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

( ) Totalmente de acuerdo

( ) De acuerdo

( ) Indeciso

( ) En desacuerdo

( ) Totalmente en desacuerdo.

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

( ) Muy frecuentemente

( ) Frecuentemente

( ) Ocasionalmente

( ) Raramente

( ) Nunca

DIMENSIÓN: RENDIMIENTO

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

( ) Muy buena

( ) Buena

( ) Regular

( ) Baja

( ) Mala

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

( ) Muy frecuentemente

( ) Frecuentemente

( ) Ocasionalmente

( ) Raramente

( ) Nunca

DIMENSIÓN: SEGURIDAD

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

( ) Totalmente de acuerdo

( ) De acuerdo

( ) Indeciso

( ) En desacuerdo

( ) Totalmente en desacuerdo.

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

( ) Totalmente de acuerdo

( ) De acuerdo

( ) Indeciso

( ) En desacuerdo

( ) Totalmente en desacuerdo.

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

( ) Siempre

( ) Casi siempre

( ) Ocasionalmente

( ) Raramente

( ) Nunca

18. Los datos académicos son solo modificados mediante autorización

( ) Siempre

( ) Casi siempre

( ) Ocasionalmente

( ) Raramente

( ) Nunca

DIMENSION: SOPORTE

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

( ) Siempre

( ) Casi siempre

( ) Ocasionalmente

( ) Raramente

( ) Nunca

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

    ( ) Siempre

    ( ) Casi siempre

    ( ) Ocasionalmente

    ( ) Raramente

    ( ) Nunca

UNIVERSIDAD SEÑOR DE SIPÁN

Guía de Entrevista

Esta entrevista, es dirigida a los jefes de oficina académicas centrales, con el objetivo de diagnosticar el estado actual de la dinámica del procesamiento de datos del área académica en una institución universitaria

Finalmente queremos agradecerle su disposición a colaborar en este estudio el cual pretende colaborar en la solución de los problemas de tratamiento de datos académicos.

INSTRUCCIONES

- El tiempo de la entrevista es aproximadamente 30 minutos.
- Consta de 16 preguntas que serán formuladas por el entrevistador.

DIMENSIÓN: RECOLECCIÓN Y MANIPULACIÓN

1. ¿Qué tipo de información manipula el área académica?
    a. Estructurada: _____
    b. No estructurada: _____

    (datos / video / audio / multimedia)

2. ¿Cuál es el volumen de información que se manipula en el área académica por cada tipo? (unidades de medida)

    _____
    _____

3. ¿Cuál es la cantidad de servidores (físicos/virtuales) con los que cuenta la institución y que data se almacena ahí?

    _____
    _____

DIMENSIÓN: CALIDAD Y RENDIMIENTO

4. El nivel de disponibilidad de los sistemas académicos se estima en:
    a. Más del 90%
    b. Entre 71 a 90%
    c. Entre 61 a 70%
    d. Menor a 61%
5. ¿Cómo califica el tiempo de respuesta de los sistemas académicos en responder una consulta o demanda?
    a. Muy bueno
    b. Bueno
    c. Regular
    d. Malo
    e. Muy malo
6. ¿Cómo califica la integridad y confiabilidad de los datos académicos?
    a. Muy bueno
    b. Bueno
    c. Regular
    d. Malo
    e. Muy malo
7. ¿La Información académica que requieren las oficinas es suficiente para el desarrollo de sus actividades?
    a. Siempre
    b. Casi siempre
    c. Normalmente
    d. A veces
    e. Nunca
8. ¿Se realiza un seguimiento de las consultas que realizan los usuarios con respecto a los datos académicos?
    a. Siempre
    b. Casi siempre
    c. Normalmente
    d. A veces
    e. Nunca

DIMENSIÓN: SEGURIDAD

9. ¿Se generan copias de seguridad de los datos académicos?
    a. Si
    b. No
10. ¿Dónde se tienen almacenados los backups?
    a. Discos externos
    b. Storage Server
    c. Almacenamiento en la nube
    d. Otros. Especifique
11. ¿Con cuánta frecuencia se realizan las copias de seguridad?
    a. Diario
    b. Semanal
    c. Mensual
    d. Anual
12. ¿Cómo calificaría los niveles de seguridad lógica que aplica la Institución para proteger los datos académicos que administra?
    a. Muy bueno
    b. Bueno
    c. Regular
    d. Malo
    e. Muy malo
13. ¿Cómo calificaría los niveles de seguridad física que aplica la Institución para proteger los datos académicos que administra?
    a. Muy bueno
    b. Bueno
    c. Regular
    d. Malo
    e. Muy malo

DIMENSION: SOPORTE

14. ¿A través de qué medios se atienden los requerimientos nuevos que soliciten a su área?
    a. Por correo electrónico
    b. Por teléfono
    c. A través de un sistema informático
    d. A través de documento escrito
    e. Otro medio. Especifique
15. ¿Cuál es el tiempo promedio en el que se atiende los requerimientos solicitados?
    a. En menos de un día
    b. En promedio 3 días
    c. En promedio 7 días
    d. Más de 7 días
16. ¿Cuál sería el principal problema para no atender en el menor tiempo posible los requerimientos de los usuarios?

    _____

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 4 INSTRUMENTO DE VALIDACION NO EXPERIMENTAL POR JUICIO DE EXPERTOS

1. NOMBRE DEL JUEZ		Jessie Leila Bravo Jaico
2.	PROFESIÓN	Ing. computación y sistemas
	ESPECIALIDAD	Transformación digital y seguridad inform.
	GRADO ACADÉMICO	Dra. en ciencias Computación y Sistemas
	EXPERIENCIA PROFESIONAL (AÑOS)	27 años
	CARGO	Docente universitaria.
Título de la Investigación: Sistema analítico basado en un modelo predictivo de procesamiento de datos en la Big data en la educación superior.		
3. DATOS DEL TESISTA		
3.1	NOMBRES Y APELLIDOS	ROGER ERNESTO ALARCÓN GARCÍA
3.2	PROGRAMA DE POSTGRADO	Doctorado en ciencias de la Computación y Sistemas
4. INSTRUMENTO EVALUADO		1. Entrevista ( ) 2. Cuestionario (X) 3. Lista de Cotejo ( ) 4. Diario de campo ( )
5. OBJETIVOS DEL INSTRUMENTO		GENERAL Diagnosticar el estado actual del procesamiento de datos académicos en la Universidad Nacional Pedro Ruiz Gallo.
		ESPECÍFICOS – Diagnosticar el estado actual de la recolección de datos académicos. – Evaluar el estado actual de la calidad de los datos. – Diagnosticar la seguridad de los datos.

	– Diagnosticar la manipulación y rendimiento de los datos.

A continuación se le presentan los indicadores en forma de preguntas o propuestas para que Ud. los evalúe marcando con un aspa (x) en "A" si está de ACUERDO o en "D" si está en DESACUERDO, SI ESTÁ EN DESACUERDO POR FAVOR ESPECIFIQUE SUS SUGERENCIAS

N	6. DETALLE DE LOS ITEMS DEL INSTRUMENTO	
01	¿Considera que los datos académicos que se registran son suficientes para la toma de decisiones? Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X) D( ) SUGERENCIAS:
02	¿Considera que los datos académicos obtenidos en los procesos de su área son claros? Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X) D( ) SUGERENCIAS:
03	¿Con que frecuencia se recopilan los datos académicos? Escala de medición: Muy frecuentemente, Frecuentemente, Ocasionalmente, Raramente y Nunca.	A(X) D( ) SUGERENCIAS:
04	¿Considera que los sistemas académicos son intuitivos y fáciles de manipular? Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X) D( ) SUGERENCIAS:
05	¿Considera que los datos que se registran en los sistemas académicos se validan? Escala de medición: Muy frecuentemente, Frecuentemente, Ocasionalmente, Raramente y Nunca.	A(X) D( ) SUGERENCIAS:
06	¿Considera que los datos que se registra en los sistemas académicos se valida? Escala de medición: Muy frecuentemente, Frecuentemente, Ocasionalmente, Raramente y Nunca.	A(X) D( ) SUGERENCIAS:
07	¿Qué tan frecuente se encuentran los datos disponibles para cuando usted lo necesita? Escala de medición: Muy frecuentemente, Frecuentemente, Ocasionalmente, Raramente y Nunca.	A(X) D( ) SUGERENCIAS:

08	¿Los reportes académicos obtenidos le permite un análisis completo para los requerimientos de su oficina? Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X)	D ( )
		SUGERENCIAS:	
09	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. Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X)	D ( )
		SUGERENCIAS:	
10	Los datos académicos que tiene almacenada la institución son de fácil acceso a los usuarios que lo requieran. Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X)	D ( )
		SUGERENCIAS:	
11	¿Qué tan frecuente se detectan datos erróneos en el procesamiento de la información? Escala de medición: Muy frecuentemente, Frecuentemente, Ocasionalmente, Raramente y Nunca.	A(X)	D ( )
		SUGERENCIAS:	
12	¿Está de acuerdo en que la institución aplique algún estándar de calidad para el procesamiento de datos? Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X)	D ( )
		SUGERENCIAS:	
13	La institución captura los datos académicos centrado en las necesidades organizacionales Escala de medición: Muy frecuentemente, Frecuentemente, Ocasionalmente, Raramente y Nunca.	A(X)	D ( )
		SUGERENCIAS:	
14	¿El tiempo de respuesta de las aplicaciones al solicitar datos es? Escala de medición: Muy buena, Buena, Regular, Baja y Mala.	A(X)	D ( )
		SUGERENCIAS:	
15	¿Los datos académicos se obtienen en tiempo real? Escala de medición: Muy frecuentemente, Frecuentemente, Ocasionalmente, Raramente y Nunca.	A(X)	D ( )
		SUGERENCIAS:	

16	¿Considera usted que los datos académicos están totalmente seguros? Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X)	D ( )
		SUGERENCIAS:	
17	¿Cree usted que los datos académicos son accesibles sólo por personal autorizado? Escala de medición: Totalmente de acuerdo, De acuerdo, Indeciso, En desacuerdo y Totalmente en desacuerdo.	A(X)	D ( )
		SUGERENCIAS:	
18	¿La institución periódicamente le solicita que actualice claves para acceder a los sistemas académicos? Escala de medición: Siempre, Casi siempre, Ocasionalmente, Raramente y Nunca.	A(X)	D ( )
		SUGERENCIAS:	
19	Los datos académicos son solo modificados mediante autorización Escala de medición: Siempre, Casi siempre, Ocasionalmente, Raramente y Nunca.	A(X)	D ( )
		SUGERENCIAS:	
20	Ante un requerimiento nuevo, la atención de la OTI es rápida Escala de medición: Siempre, Casi siempre, Ocasionalmente, Raramente y Nunca.	A(X)	D ( )
		SUGERENCIAS:	
21	¿Ante un incidente relacionado con los sistemas académicos el tiempo de atención es rápido? Escala de medición: Siempre, Casi siempre, Ocasionalmente, Raramente y Nunca.	A(X)	D ( )
		SUGERENCIAS:	
	PROMEDIO OBTENIDO:	A(X)	D ( ):
6	COMENTARIOS GENERALES		
7	OBSERVACIONES		

Juez Experto

Colegiatura Nº 71194

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXO N° 4 INSTRUMENTO DE VALIDACION NO EXPERIMENTAL POR JUICIO DE EXPERTOS

1. NOMBRE DEL JUEZ		Jessie deila Bravo Taico
2.	PROFESIÓN	Ing. computación y sistemas
	ESPECIALIDAD	Transform. digital y seguridad inform.
	GRADO ACADÉMICO	Dra. en ciencias computación y sistemas
	EXPERIENCIA PROFESIONAL (AÑOS)	27 años
	CARGO	Docente universitaria
Título de la Investigación: Sistema analítico basado en un modelo predictivo de procesamiento de datos en la Big data en la educación superior		
3. DATOS DEL TESISTA		
3.1	NOMBRES Y APELLIDOS	ROGER ERNESTO ALARCÓN GARCÍA
3.2	PROGRAMA DE POSTGRADO	Doctorado en ciencias de la Computación y Sistemas
4. INSTRUMENTO EVALUADO		1. Entrevista (**X**) 2. Cuestionario ( ) 3. Lista de Cotejo ( ) 4. Diario de campo ( )
5. OBJETIVOS DEL INSTRUMENTO		<u>GENERAL</u> Diagnosticar el estado actual de la dinámica del procesamiento de datos del área académicos en la Universidad Nacional Pedro Ruiz .
		<u>ESPECÍFICOS</u> – Diagnosticar el estado actual de la recolección de datos académicos.

	– Evaluar el estado actual de la calidad de los datos. – Diagnosticar la seguridad de los datos. – Diagnosticar la manipulación y rendimiento de los datos.

A continuación se le presentan los indicadores en forma de preguntas o propuestas para que Ud. los evalúe marcando con un aspa (x) en "A" si está de ACUERDO o en "D" si está en DESACUERDO, SI ESTÁ EN DESACUERDO POR FAVOR ESPECIFIQUE SUS SUGERENCIAS

N	6. DETALLE DE LOS ITEMS DEL INSTRUMENTO	
01	¿Qué tipo de información manipula el área académica? Escala de medición: Estructurada, No estructurada.	A(X) D ( ) SUGERENCIAS:
02	¿Cuál es el volumen de información que se manipula en el área académica por cada tipo? Escala de medición: Numérica.	A(X) D ( ) SUGERENCIAS:
03	¿Cuál es la cantidad de servidores (físicos/virtuales) con los que cuenta la institución y que data se almacena ahí? Escala de medición: Numérica.	A(X) D ( ) SUGERENCIAS:
04	El nivel de disponibilidad de los sistemas académicos se estima en Escala de medición: Más del 90%, entre 71 y 90%, entre 61 y 70% y menor a 61%.	A(X) D ( ) SUGERENCIAS:
05	¿Cómo califica el tiempo de respuesta de los sistemas académicos en responder una consulta o demanda? Escala de medición: Muy bueno, Bueno, Regular, Malo y Muy malo.	A(X) D ( ) SUGERENCIAS:
06	¿Cómo califica la integridad y confiabilidad de los datos académicos? Escala de medición: Muy bueno, Bueno, Regular, Malo y Muy malo.	A(X) D ( ) SUGERENCIAS:
07	¿La Información académica que requieren las oficinas es suficiente para el desarrollo de sus actividades? Escala de medición: Siempre, Casi siempre, Normalmente, A veces y Nunca.	A(X) D ( ) SUGERENCIAS:

08	¿Se realiza un seguimiento de las consultas que realizan los usuarios con respecto a los datos académicos? Escala de medición: Siempre, Casi siempre, Normalmente, A veces y Nunca.	A(X) D( ) SUGERENCIAS:
09	¿Se generan copias de seguridad de los datos académicos? Escala de medición: Si y No.	A(X) D( ) SUGERENCIAS:
10	¿Dónde se tienen almacenados los backups? Escala de medición: Discos externos, Storage server, Almacenamiento en la nube, otros.	A(X) D( ) SUGERENCIAS:
11	¿Con cuánta frecuencia se realizan las copias de seguridad? Escala de medición: Diario, Semanal, Mensual y Anual.	A(X) D( ) SUGERENCIAS:
12	¿Cómo calificaría los niveles de seguridad lógica que aplica la Institución para proteger los datos académicos que administra? Escala de medición: Muy bueno, Bueno, Regular, Malo y Muy malo.	A(X) D( ) SUGERENCIAS:
13	¿Cómo calificaría los niveles de seguridad física que aplica la Institución para proteger los datos académicos que administra? Escala de medición: Muy bueno, Bueno, Regular, Malo y Muy malo.	A(X) D( ) SUGERENCIAS:
14	¿A través de qué medios se atienden los requerimientos nuevos que soliciten a su área? Escala de medición: Por correo electrónico, Por teléfono, A través de un sistema informático, A través de documento escrito y otro medio.	A(X) D( ) SUGERENCIAS:
15	¿Cuál es el tiempo promedio en el que se atiende los requerimientos solicitados? Escala de medición: En menos de un día, En promedio 3 días, En promedio 7 días y Más de 7 días.	A(X) D( ) SUGERENCIAS:

16	¿Cuál sería el principal problema para no atender en el menor tiempo posible los requerimientos de los usuarios? Escala de medición: Textual	A(X)	D( )
		SUGERENCIAS:	
	PROMEDIO OBTENIDO:	A(X)	D( ):
8	COMENTARIOS GENERALES		
9	OBSERVACIONES		

Juez Experto

Colegiatura Nº.....71194.....

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXOS N° 5 CONSENTIMIENTO INFORMADO

Institución: Universidad Señor de Sipán

Investigador: **Roger Ernesto Alarcón García**

Título: **Sistema analítico basado en un modelo predictivo de procesamiento de datos en la Big Data en la Educación Superior**

Yo, CARLOS HERIBERTO RUIZ OLIVA, identificado con DNI 16449928, DECLARO:

Haber sido informado de forma clara, precisa y suficiente sobre los fines y objetivos que busca la presente investigación **"Sistema analítico basado en un modelo predictivo de procesamiento de datos en la Big Data en la Educación Superior"**, así como en qué consiste mi participación.

Estos datos que yo otorgue serán tratados y custodiados con respeto a mi intimidad, manteniendo el anonimato de la información y la protección de datos desde los principios éticos de la investigación científica. Sobre estos datos me asisten los derechos de acceso, rectificación o cancelación que podré ejercitar mediante solicitud ante el investigador responsable. Al término de la investigación, seré informado de los resultados que se obtengan.

Por lo expuesto otorgo MI CONSENTIMIENTO para que se realice la **Entrevista** que permita contribuir con el objetivo de la investigación de **"Elaborar un sistema analítico basado en un modelo predictivo que tenga en cuenta la relación entre las técnicas predictivas integradas y los grandes volúmenes de datos para el procesamiento de los datos en la big data"**.

Las entrevistas serán grabadas y degrabadas fielmente.

Firmado digitalmente por: RUIZ OLIVA CARLOS HERIBERTO
Motivo: Director - Dirección de Servicios Académicos
Fecha: 06/10/2021 12:32:33-0500

DNI: 16449928

UNIVERSIDAD SEÑOR DE SIPÁN

ANEXOS N° 6

APROBACIÓN DEL INFORME DE TESIS

El Docente:

Dr. JUAN CARLOS CALLEJAS TORRES

De la Asignatura:

SEMINARIO DE INVESTIGACIÓN VI: INFORME DE TESIS

APRUEBA:

El Informe de Tesis: ***"SISTEMA ANALÍTICO BASADO EN UN MODELO PREDICTIVO DE PROCESAMIENTO DE DATOS EN LA BIG DATA EN LA EDUCACIÓN SUPERIOR"***

Presentado por:

Mg. ROGER ERNESTO ALARCÓN GARCÍA

Chiclayo, de 14 de diciembre del 2021

_____

Dr. JUAN CARLOS CALLEJAS TORRES