PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Propuesta de Diseño de una Arquitectura Big Data para el Monitoreo de los Cultivos de las Empresas Azucareras en la región La Libertad en el Perú

Tesis para obtener el grado académico de Maestro en Gestión de la Ingeniería

que presenta:

Miguel Angel Rodríguez Saldaña

Asesor:

Ing. Jonatán Edward Rojas Polo

Lima, 2025

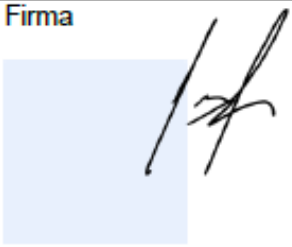
Informe de Similitud

Yo, Jonatan Edward Rojas Polo, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis titulada(o) Propuesta de Diseño de una Arquitectura Big Data para el Monitoreo de los Cultivos de las Empresas Azucareras en la región La Libertad en el Perú, de el autor Miguel Ángel Rodríguez Saldaña, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 17%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 19/06/2025.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de investigación, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.
-

Lugar y fecha:

Lima, 19 de junio de 2025.

<table>
<tr><td colspan="2">Apellidos y nombres del asesor / de la asesora:
Rojas Polo, Jonatan Edward</td></tr>
<tr><td>DNI:
42529429</td><td rowspan="2">Firma
  </td></tr>
<tr><td>ORCID:
https://orcid.org/0000-0003-2522-3422</td></tr>
</table>

Agradecimientos

A mis padres Segundo y Nicida, y a mis hermanos César, Mike y Jackeline.

A mi asesor, Mgs. Jonatan, quién me enseño los temas relacionados al Big Data y acompañarme durante todo el desarrollo de esta tesis.

Dedicatoria

A mis padres Segundo y Nicida.

A mis abuelos César e Eunice.

RESUMEN

Las empresas agrícolas en el Perú se enfrentan con mayor frecuencia a interrupciones laborales de diversas índoles, generando un gran interés por la adopción de tecnologías de automatización. Las tecnologías en la actualidad aplicables en la agroindustria como automatización, robótica, Business Analitics, vehículos no tripulados, visión artificial, Internet de las Cosas y Big Data ofrecen un buen resultado en la optimización de procesos, mejora continua, mejor relación entre inversión necesaria e impacto en la operatividad

La agroindustria en el país es uno de los sectores con mayor crecimiento económico, producción, exportación para nuestra economía. Además, ayuda a dar oportunidad de trabajo a las personas locales y actividad agrícola.

Las empresas azucareras con suficientes fuentes de agua, suelos fértiles y clima vegetal adecuado tienen un gran potencial, pero existen problemas como dificultad en supervisar y obtener información sobre el estado de los cultivos, falta de nitrógeno y otros nutrientes, falta de agua o zonas afectadas por malas hierbas, malezas y plagas.

Este documento presenta una arquitectura de Big Data que permite el uso, gestión y análisis de datos de diversas fuentes, como el clima, la información del suelo, imágenes satelitales/de drones y sensores en áreas agrícolas para ayudar a identificar y cuantificar el retraso en el crecimiento y las variables relacionadas.

Existen grandes oportunidades de aplicar el análisis en Big Data en el sector agroindustrial con el objetivo de ir hacia una agricultura inteligente que ayude a tomar decisiones en los procesos y dar soluciones a los problemas complejos de la agricultura actual.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

INTRODUCCIÓN

El Perú tiene una ventaja competitiva en el sector agroindustrial, por su ubicación geográfica, diversidad de recursos naturales y la ampliación de tierra disponible para la actividad agrícola. Pero la fuerza de este potencial se está disminuyendo por la gestión del gobierno, baja inversión en infraestructura y tecnología, y en la constante asociatividad de la cadena de valor de la agricultura (Fung, 2014).

El azúcar es un producto nacional que contribuye enormemente al PIB agrícola en los países en desarrollo. Por tanto, el mayor desafío para la industria azucarera es reducir los costos de producción y operación. Los países con mayor producción y exportación son Brasil, India, Colombia, Tailandia, China y Australia (Aguilar, 2017).

Para (Asociados&Apoyo, 2017) el cultivo de caña en el Perú se concentra en la costa norte (78,8%, INEI). Según datos del INEI, la superficie cosechada alcanzó las 90.400 hectáreas al cierre de 2014, la mayor superficie cosechada en la historia del Perú. En el 2018 se cultivaron 24 mil nuevas hectáreas. En resumen, debido a la continua recuperación y consolidación de la industria.

Se proyecta que para el año 2050 se necesitará que la producción global de alimentos se duplique (FAO, 2009). El desafío social de alimentar a un número cada vez mayor de personas crea desafíos ambientales, ya que el trabajo debe realizarse sin aumentar la cantidad de agua, tierras de cultivo y fertilizantes utilizados para cultivar alimentos. Otro problema es el cambio climático, que se espera que afecte significativamente la producción agrícola mundial, debido a vulnerabilidad a los desastres naturales y que cuenten con sistemas de pronóstico y estrategias de mitigación débiles (Zhao, D., & Li, Y. R., 2015). En este contexto, a las empresas agroindustriales peruanas se les presenta los retos de adaptarse a estas condiciones cambiantes y la oportunidad de aprovechar la gran demanda y posicionarse como una agricultura líder. Para ello, es importante delinear estrategias, darle nuevos valores agregados y principalmente el uso de la tecnología.

La agricultura de precisión (AP) es una tecnología que combina sensores, sistemas de información y equipos avanzados para incrementar la

productividad de las cosechas y reducir el impacto en el medio ambiente, sin perjudicar su calidad (Pierce, F. J., & Nowak, P., 1999). Muchos datos información que proporciona la AP requiere nuevos métodos para aumentar el conocimiento. Por lo tanto, se está investigando el Big Data para encontrar soluciones innovadoras para analizar grandes conjuntos de datos (Bendre, M. R., Thool, R. C., & Thool, V. R., 2015).

Las tecnologías que impulsa esta incluye el uso de sensores inteligentes e inalámbricos habilitadas para IoT que recopilan datos del suelo en tiempo real, datos de rendimiento, diversas características ambientales, comportamiento animal y estado de la máquina. Al calcular y analizar los datos de este sensor mediante IoT, los agricultores pueden obtener información valiosa sobre el clima y el pronóstico, el monitoreo de cultivos y el pronóstico del rendimiento, la detección de enfermedades animales y vegetales (Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A., & Bhansali, S., 2019).

La AP ha adoptado las tecnologías de teledetección en el monitoreo estratégico utilizando imágenes satelitales para la toma de decisiones regionales (Delgado, J. A., Short, N. M., Roberts, D. P., & Vandenberg, B., 2019), y más recientemente para el monitoreo y control estratégico a partir de la información de datos de detección remota a baja altitud para el tratamiento específico del sitio a escala de campo (Huang, Y., Chen, Z. X., Tao, Y. U., Huang, X., Z., & Gu, X. F.). La teledetección agrícola es una tecnología clave que produce datos e información georreferenciada para la planificación de operaciones agrícolas de precisión. Los datos generados deben analizarse rápidamente evaluar y decidir, lo que se logra mediante la integración gradual en la ciencia y tecnologías de datos. Uno de los temas de investigación es el procesamiento de Big Data y su transformación en "small data" para problemas específicos o áreas de actividad agrícola.

Para obtener valor de Big Data, debe procesarse y analizarse de manera oportuna. Finalmente, una combinación con Big Data y otras tecnologías puede analizar la información y ayudar a sacar conclusiones de forma rápida y automática, lo que a su vez permite que las máquinas inteligentes inicien o detengan el riego y la fertilización, y obtengan información sobre las condiciones climáticas, la calidad del suelo y sus efectos nutricionales.

Además, puede guiar una buena planificación de cultivos, ya que ayuda a predecir las tendencias de precios y demanda del mercado (Agtech, 2017).

El estudio de este trabajo se centra en la investigación, análisis y diseño de una solución de Big Data para los cultivos de caña de azúcar en la región Libertad.

Capítulo 1: Marco teórico

Este capítulo se describe los conceptos del Big Data y características de las 5Vs del Big Data, también conceptos de Machine Learning, Base de Datos e inteligencia artificial que serán las herramientas pilares para el desarrollo de esta tesis.

1.1. Big Data

1.1.1. Definición Big Data

Con el rápido desarrollo de la tecnología de la información, los productos electrónicos han proliferado en varios campos como la salud, educación, servicios ambientales, agricultura, el clima, las redes sociales, el ejército, la vida web y la industria (Agtech, 2017). Los datos se recopilan de todo en el proceso de comportamiento continuo como se muestra en la Ilustración 1., La detección generalizada y de bajo costo y una caída dramática en los costos de almacenamiento de datos han expandió la cantidad de información extraíble sin esfuerzo con una velocidad y volumen. La cantidad de datos ha experimentado un crecimiento significativo crecido enormemente durante los últimos diez años y ahora se clasifica como "Big Data" (Bendre, M. R., Thool, R. C., & Thool, V. R., 2015).

Ilustración 1. Qué causa el Big Data

Según un informe de International Business Machines (IBM), cada día se generan dos quintillones de bytes de datos en el mundo, y el 90% se han generado en los últimos dos años. Big data no es una tecnología independiente, sino sucesora de las revoluciones tecnológicas de los últimos cincuenta años o más (Huang, Y., Chen, Z. X., Tao, Y. U.,

Huang, X., Z., & Gu, X. F.). Big Data se define como grandes volúmenes de datos de alta velocidad, variables y complejos que necesita técnicas y tecnologías con el fin de lograr la extracción, almacenamiento, distribución, gestión y análisis. Es el proceso de recopilar, administrar y analizar datos para generar información valiosa y descubrir patrones ocultos. Conjuntos de datos que no pueden ser capturados, administrados y procesados por una computadora típica, según la definición de Hadoop. Todos los datos recopilados a nuestro alrededor se consideran una gran fuente de datos (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021).

El objetivo de Big Data es extraer información valiosa de los datos. Procesa grandes cantidades de datos, que están lejos del análisis más clásico. Con esta herramienta, las empresas pueden aprender sobre los clientes, la competencia, el medio ambiente y mucho más (Holmes, 2017) .

1.1.2. Características de Big Data

Sin comprender las características de los grandes datos, que se pueden resumir en 5 modelos, es imposible gestionar de forma eficaz estos datos masivos que son: variedad, velocidad, volumen, veracidad y valor como se ilustra en el siguiente esquema (Corea, 2019).

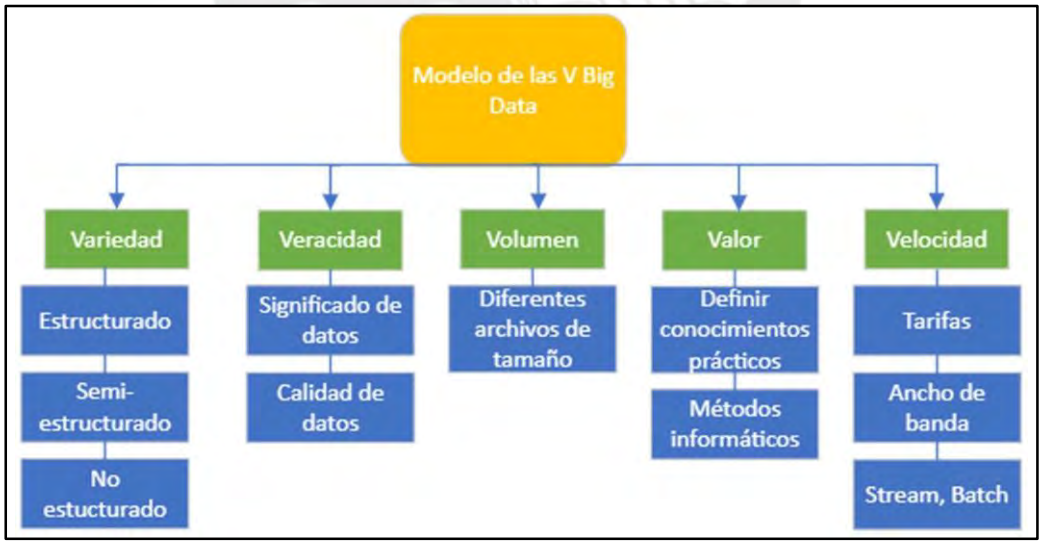


Ilustración 2. Modelo V del Big Data

Volumen

Uno de los problemas que surgen con los grandes datos es la gran cantidad de datos que deben administrarse. Significa que la cantidad de datos electrónicos que se recopilan y

almacenan hoy, está creciendo a un ritmo cada vez mayor. Esta avalancha de datos, principalmente de redes sociales y nuevas plataformas, tiene un enorme potencial y desafíos. Su flujo continuo da como resultado un ciclo de vida de datos muy corto. Sin embargo, los avances tecnológicos han permitido nuevas formas avanzadas de almacenamiento, gestión, desarrollo y análisis (Corea, 2019).

Variedad

Esta propiedad corresponde a cómo se estructuran los datos. Hay tantos datos que es necesario agruparlos. Estos datos pueden ser estructurados, semiestructurados y no estructurados. La forma en que interpretamos los datos afecta los resultados que queremos lograr, y malinterpretar los datos puede distorsionar los resultados.

Las nuevas plataformas y redes sociales han creado más fuentes de datos. Flat incluye texto, imágenes, datos web, tweets, audio, video y más. Esto brinda oportunidades ilimitadas para Big Data y dificulta su almacenamiento, procesamiento y análisis (Corea, 2019).

Velocidad

Otra característica del Big Data está relacionada con la velocidad. Un buen análisis puede llevar horas o incluso días, lo más probable es que esto último. Es importante tomarse el tiempo necesario para procesar la información, ya que esto hará o deshará el análisis.

El período de vida de los datos es corto y requiere respuesta rápida, un procesamiento rápido y evitar la obsolescencia. La capacidad de respuesta es fundamental cuando se trata de optimizar y monetizar el uso correcto de los datos. Esto mejora la exactitud y calidad de los resultados (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021).

Veracidad

Uno de los retos del Big Data es saber si hay datos falsos entre todos los datos que se procesan. Es esta cantidad de datos lo que hace que la gente dude de su autenticidad. Tantos, muchos pueden ser incorrectos o incompletos y pueden ser engañosos cuando provienen de otros países. Las empresas necesitan realizar un análisis integral para verificar la precisión de los datos para que puedan usarlos para lograr sus objetivos (Corea, 2019).

La precisión afecta la calidad en los datos. Y también se debe verificar la integridad y autenticidad de este. Las decisiones basadas en datos solo son correctas si los datos son buenos, confiables, claros y verdaderos. (Corea, 2019).

Valor

Como antes, los datos analizados no solo deben ser reales, sino también aportar valor. El análisis debe comprobar y descartar las que fallan y conservar las que fallan, siempre que ambas sean verdaderas.

El valor es el factor individual más importante en Big Data y, hasta cierto punto, incluye todos los demás factores. Después de todo, si todos los demás parámetros son correctos, se puede obtener más información y conocimiento. Así que estos serán datos muy valiosos. Valor significa rentabilizar los datos, extraer toda la información que contienen y crear una ventaja competitiva. (Corea, 2019).

Las otras "V"

Variabilidad

Es la cantidad de inconsistencia en los datos. Deben detectarse utilizando anomalías y técnicas de detección de anomalías antes de que pueda ocurrir algo significativo. (Corea, 2019).

Otro es la gran cantidad de dimensiones y múltiples fuentes de datos típicas y diversas. Esta característica se relaciona a la velocidad inconsistente a la que se cargan en la base de datos.

Validez

Esto se refiere a la limpieza, exactitud y corrección de los datos cuando se utilizan. Los beneficios son tan buenos como los datos subyacentes, por lo que se deben aplicar buenas prácticas y para garantizar una calidad de datos consistente, definiciones comunes y metadatos. (Corea, 2019).

Volatilidad

O en momentos en que es necesario guardar datos. Antes de Big Data, la gente tendía a almacenar datos indefinidamente, porque pequeñas cantidades de datos significaban tarifas bajas. En una base de datos se puede guardar en tiempo real sin problemas de rendimiento. (Corea, 2019).

Sin embargo, debido a la velocidad y volumen, su volatilidad debe ser considerada importante. Dicho esto es necesario establecer reglas para la disponibilidad y validez de estos datos y, si es necesario, para asegurar la rápida recuperación de la información.

Visualización

Otra característica del Big Data es la complejidad de su visualización. Por ejemplo, no puede confiar en un gráfico tradicional que trae trillones de filas de datos, por lo que no hay necesidad de diferentes representaciones como agrupaciones o mapas, coordenadas, gráficos, etc. (Corea, 2019).

1.1.3. Fuentes de Big Data

Varias fuentes están dando lugar a un gran aumento de volumen de datos. Gran parte de este crecimiento se puede atribuir a la digitalización de casi todo en el mundo. Pago electrónico de facturas, compras en línea, comunicación a través de redes sociales, transacciones por correo electrónico en diversas organizaciones, visualización digital de datos organizacionales, etc. son algunos ejemplos de esta digitalización. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

Sensores: Los sensores que contribuyen al gran volumen de datos se enumeran a continuación.

- Sensores de acelerómetro instalados en dispositivos móviles para detectar las vibraciones y otros movimientos.
- Sensores en vehículos y dispositivos médicos.

Cuidado de la salud: Las principales fuentes en el campo de la salud son:

- Los registros de salud electrónicos recopilan y muestran información del paciente tales como antecedentes médicos, prescripciones de los médicos y resultados de pruebas de laboratorio.
- Un portal para pacientes brinda a los pacientes acceso a sus registros médicos personales almacenados en el EHR.
- El repositorio de datos clínicos agrega registros de pacientes individuales de varias fuentes clínicas y los consolida para brindar una vista unificada del historial del paciente.

Caja negra: Los datos son generados por la caja negra en aviones y helicópteros. Las cajas negras registran las actividades de vuelo, los informes de la tripulación y la información de rendimiento de la aeronave.

Datos Web: Los minoristas en línea capturan los datos generados al hacer clic en un enlace en un sitio web. Se trata de realizar un análisis del flujo de clics para analizar los intereses de los usuarios y las tendencias de compra para generar recomendaciones basadas en los intereses de los clientes y publicar anuncios relevantes para los consumidores.

Datos organizativos: Las transacciones de correos electrónicos y los documentos que se generan dentro de las organizaciones contribuyen juntos a los datos de estas.

La Ilustración 3 ilustra los datos generados por varias fuentes que fueron discutido anteriormente.

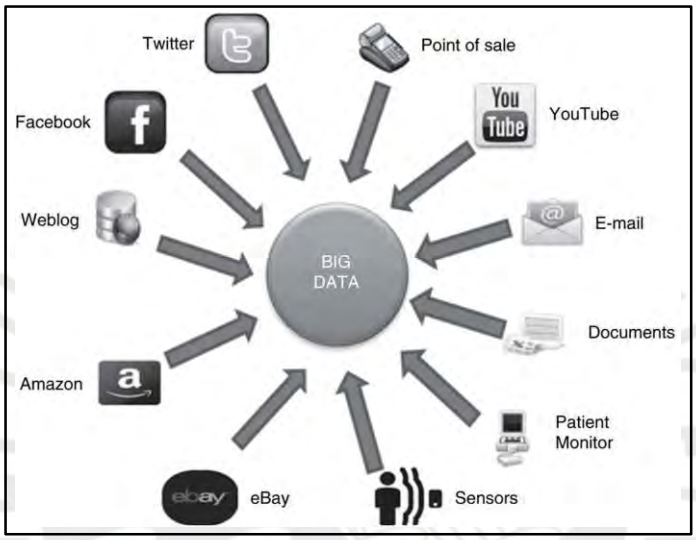


Ilustración 3. Fuentes de Big Data.

1.1.4. Diferentes tipos de datos

Los datos pueden ser generados por máquinas o por humanos. Los datos generados por humanos se refieren a los datos generados como resultado de las interacciones de los humanos con las máquinas. Correos electrónicos, documentos, publicaciones de Facebook son algunos de los datos generados por humanos. Los datos generados por máquinas se refieren a los datos generados por aplicaciones informáticas o dispositivos de hardware sin intervención humana activa. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021). Los datos de sensores, sistemas de alerta de desastres, sistemas de pronóstico del tiempo y datos satelitales son algunos de los datos generados por máquinas. Los datos generados por máquinas y humanos pueden representarse mediante los siguientes tipos:

- Datos estructurados
- Datos no estructurados
- Datos semiestructurados

Datos estructurados

Según (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021), los datos que se pueden guardar en un formato tabular con filas y columnas o en una base de datos relacional se denominan datos estructurados. Datos estructurados a menudo generados por empresas, las empresas exhiben un alto grado de organización y pueden procesarse fácilmente utilizando herramientas de minería de datos y pueden consultarse y recuperarse utilizando el campo de clave principal.

Datos no estructurados

Los ejemplos de datos no estructurados incluyen video, audio, imágenes, correos electrónicos, archivos de texto, publicaciones en redes sociales y archivos PDF. Los datos no estructurados normalmente residen en archivos de texto o archivos binarios. Estos no tienen ninguna estructura interna identificable.

Datos semiestructurados

Los datos semiestructurados son aquellos que tienen una estructura, pero no encajan en una BD relacional. Los datos semiestructurados se organizan para un análisis más sencillo que los datos no estructurados. JSON y XML son ejemplos de datos semiestructurados.

1.1.5. Infraestructura de Big Data

Los componentes principales de la tecnología de Big Data son las herramientas y tecnologías que permiten que los datos se almacenen, procesen y analicen. El método de almacenamiento de datos en tablas ya no es compatible con la evolución de datos con las 3V (es decir, volumen, velocidad y variación). (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

El sólido RBDMS ya no era rentable. La escalabilidad de RDBMS para guardar y procesar una gran cantidad de datos se volvió costosa. Esto condujo al surgimiento de nueva tecnología, que era altamente escalable a muy bajo costo.

Las tecnologías clave incluyen:

- Hadoop
- HDFS
- MapReduce

Hadoop

Apache Hadoop es un software escrito en Java y puede almacenar grandes cantidades de datos estructurados, semiestructurados y no estructurados en un sistema de archivos distribuido y procesarlos en paralelo. Es una plataforma de almacenamiento altamente escalable y rentable.

La escalabilidad de Hadoop se refiere a la capacidad de mantener el rendimiento incluso cuando la carga aumenta significativamente a medida que se agregan más nodos. Los archivos de Hadoop se escriben una vez y se leen varias veces. El contenido del archivo no se puede cambiar. Una gran cantidad de computadoras interconectadas que trabajan juntas como un sistema se denomina clúster. Los clústeres de Hadoop están diseñados para almacenar y analizar de manera rentable grandes cantidades de datos diversos en un entorno informático distribuido.

HDFS

Es un sistema de archivos distribuido para almacenar los datos en clústeres de computadoras. Fue desarrollado originalmente como parte del proyecto Apache Hadoop.

HDFS está diseñado para ser sumamente escalable y tolerante a fallos, lo que significa que puede manejar grandes cantidades de datos y seguir funcionando incluso si hay fallas en el hardware o en los nodos individuales del clúster. Es especialmente adecuado para el procesamiento de datos en lotes (batch processing) y aplicaciones que requieren un acceso de lectura eficiente.

Los datos se dividen en bloques más pequeños y se distribuyen en diferentes nodos del clúster. Estos se replican en varios nodos para ser función a la tolerancia a fallos y disponibilidad. Esto significa que, incluso si un nodo falla, los datos aún están disponibles en otros nodos. No requiere hardware altamente confiable y costoso.

MapReduce

MapReduce es el modelo de programación de procesamiento por lotes para el marco Hadoop, que adopta un principio de divide y vencerás. Es altamente escalable, confiable y tolerante a fallas, capaz de procesar datos de entrada con cualquier formato en entornos informáticos paralelos y distribuidos que solo admiten cargas de trabajo por lotes.

Su rendimiento reduce significativamente el tiempo de procesamiento en comparación con el paradigma de procesamiento por lotes tradicional, ya que el enfoque tradicional consistía en mover los datos de la plataforma de almacenamiento a la plataforma de procesamiento, mientras que el paradigma de procesamiento de MapReduce reside en el marco donde los datos en realidad residen.

1.2. Ciclo de Vida de Big Data

Big data produce grandes beneficios, desde ideas comerciales innovadoras hasta formas no convencionales de tratar enfermedades, superando los desafíos. Los desafíos surgen porque gran parte de los datos son recopilados por la tecnología actual. Las tecnologías de Big Data son capaces de capturarlos y analizarlos de manera efectiva. La infraestructura de Big Data implica nuevos modelos informáticos con la capacidad de procesar cálculos distribuidos y paralelos con almacenamiento y rendimiento altamente escalables. Algunos de los componentes de Big Data incluyen Hadoop (marco), HDFS (almacenamiento) y MapReduce (procesamiento). En la Ilustración 4 se ilustra el ciclo de vida. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

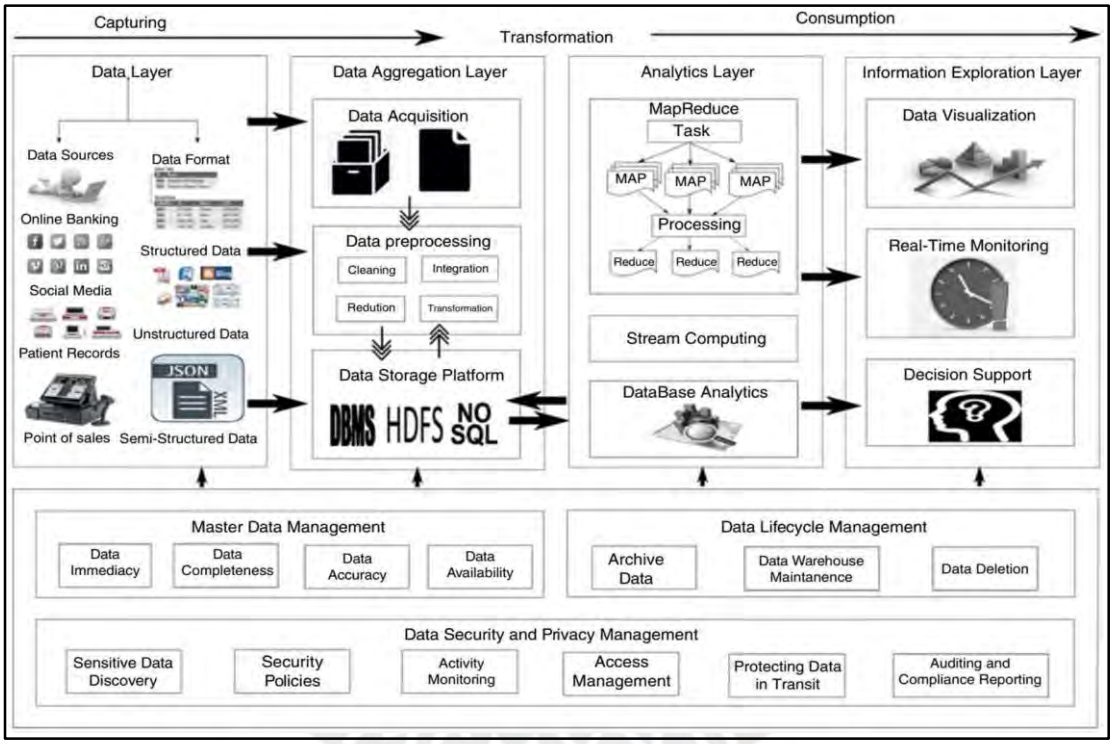


Ilustración 4. Ciclo de Vida de Big Data.

Big data produce grandes beneficios, desde ideas comerciales innovadoras hasta formas no convencionales como de tratar enfermedades y superar los desafíos. Los desafíos surgen porque gran parte de los datos son recopilados por la tecnología actual. Las tecnologías de Big Data son capaces de capturarlos y analizarlos de manera efectiva. La infraestructura de Big Data implica nuevos modelos informáticos con la capacidad de procesar cálculos distribuidos y paralelos con almacenamiento y rendimiento altamente escalables. Algunos de los componentes incluyen Hadoop (marco), HDFS almacenamiento) y MapReduce (procesamiento).

Capture datos que ingresan a alta velocidad desde múltiples fuentes con diferentes formatos. Los datos recopilados se almacenan en HDFS, NoSQL y otras plataformas de almacenamiento y luego se procesan previamente para que los datos sean adecuados para el análisis. Los datos preprocesados almacenados en la plataforma de almacenamiento luego pasan a la capa de análisis, donde los datos se procesan utilizando herramientas de Big Data como MapReduce e YARN para descubrir el conocimiento oculto de ellos. La analítica y el aprendizaje automático son conceptos importantes durante todo el proceso. El análisis de texto es un tipo de análisis realizado en datos textuales no estructurados. Con el crecimiento de las redes sociales y las transacciones por correo electrónico, ha aumentado la importancia del análisis de texto. El análisis predictivo sobre el comportamiento del consumidor y el análisis de interés del consumidor se realizan en los datos de texto extraídos de varias fuentes en línea, como redes sociales, sitios web de venta minorista en línea y mucho más. El aprendizaje automático ha hecho posible el análisis de texto. Los datos analizados se representan visualmente mediante herramientas de visualización como Tableau para que el usuario final pueda comprenderlos fácilmente para tomar decisión.

1.2.1. Generación de Big Data

La primera etapa es la generación de datos. La gama de datos de diferentes fuentes se está ampliando gradualmente.

1.2.2. Agregación de datos

La etapa de recopilación de datos incluye la recopilación de datos sin procesar, la transferencia de datos a plataformas de almacenamiento y el preprocesamiento. Esto quiere decir que se están registrando más datos a un ritmo cada vez mayor.

El preprocesamiento incluye limpieza, integración, transformación y reducción de datos para que sean confiables, sin errores, consistentes y precisos. Los datos capturados suelen ser ser redundantes, ocupar espacio de almacenamiento y aumentar los costos de almacenamiento, lo que se puede evitar procesando previamente los datos. Además, la mayoría de los datos recopilados pueden no ser significativos para fines de análisis y, por lo tanto, deben comprimirse durante el preprocesamiento. Por lo tanto, el preprocesamiento de datos efectivo es esencial para un almacén de datos rentable. Los datos preprocesados luego se transmiten para varios propósitos, como el modelado y análisis de datos. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

1.2.3. Procesamiento de datos

Es un proceso que convierte lo datos sin procesar en un formato entendible y proporciona acceso a datos consistentes y exactos. Los datos obtenidos de múltiples fuente son incorrectos, incompletos e inconsistentes debido a su volumen masivo y fuentes heterogéneas, y no tiene sentido almacenar estos. Además, algunas aplicaciones analíticas tienen un requisito crucial para la calidad de los datos. Por lo tanto, para un análisis de datos efectivo, eficiente y preciso, el preprocesamiento sistemático de datos es esencial. La calidad de los datos de origen se ve afectada por varios factores. Por ejemplo, los datos pueden tener errores, como un campo de salario con un valor negativo (por ejemplo, salario = −2000), que surge debido a errores de transmisión o errores tipográficos o entrada de datos incorrecta intencional por parte de los usuarios que no desean divulgar su información personal. Incompleto implica que el campo carece de los atributos de interés (por ejemplo, Educación = ""), lo que puede provenir de un campo no aplicable o errores de software. La inconsistencia en los datos se refiere a las discrepancias de estas, por ejemplo, la fecha de nacimiento y la edad pueden ser inconsistentes. Las inconsistencias en los datos surgen cuando los datos recopilados provienen de diferentes fuentes, debido a inconsistencias en las convenciones de nomenclatura entre diferentes países e inconsistencias en el formato de entrada (por ejemplo, el campo de fecha DD/MM cuando se interpreta como MM/DD). Las fuentes de datos a menudo tienen datos redundantes en diferentes formas y, por lo tanto, duplicados en los datos también deben eliminarse en el preprocesamiento de datos para que sean significativos y libre de errores. Hay varios pasos involucrados en el preprocesamiento de datos:

1.2.3.1. Integración de Datos

Esta implica combinar datos de diferentes fuentes para brindar a los usuarios finales una vista de datos unificada. Se enfrentan varios desafíos al integrar datos; por ejemplo, al extraer datos del perfil de una persona, el nombre y el apellido pueden intercambiarse en una determinada cultura, por lo que en tales casos la integración puede ocurrir de manera incorrecta. Las redundancias de datos a menudo ocurren al integrar datos de múltiples fuentes. La Ilustración 5 ilustra que fuentes diversificadas, como organizaciones, teléfonos inteligentes, computadoras personales, satélites y sensores, generan datos dispares, como correos electrónicos, detalles de empleados, mensajes de chat de WhatsApp, publicaciones en redes sociales, transacciones en línea, imágenes satelitales y datos sensoriales. De acuerdo a los diferentes tipos de datos deben integrarse y presentarse como datos unificados para la limpieza, el modelado y almacenamiento de datos y para

extraer, transformar y cargar (ETL) los datos. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

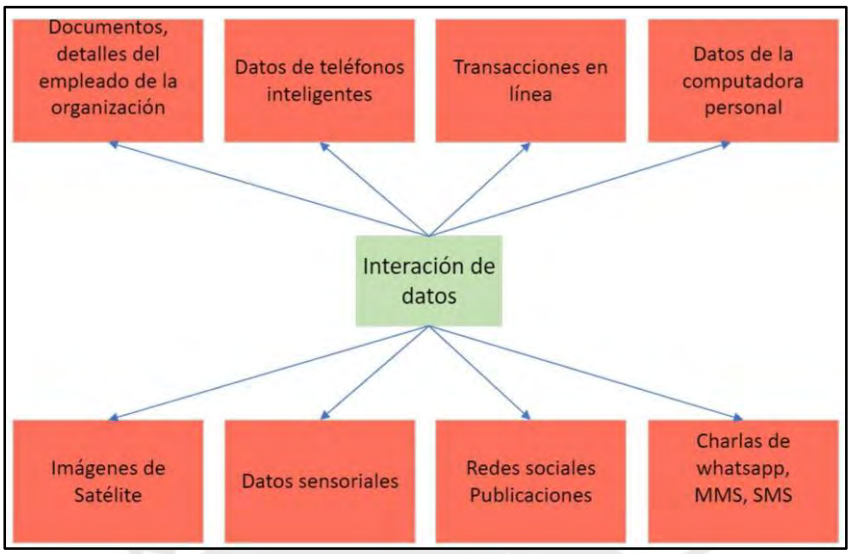


Ilustración 5. Integración de datos.

1.2.3.2. Limpieza de datos

Este proceso completa los valores faltantes, corrige errores e inconsistencias y elimina la duplicación de datos para mejorar la calidad. Cuanto mayor es la heterogeneidad de la fuente de datos. Por lo tanto, pueden estar involucradas varias etapas de limpieza. Se requiere un análisis detallado de los datos para determinar los tipos de errores e inconsistencias encontrados. La redundancia en los datos es la duplicación de estas, que aumenta los costos de almacenamiento y transmisión; y reduce la precisión y confiabilidad de los datos. Los valores faltantes se pueden completar manualmente, pero esto es engorroso, requiere mucho tiempo y no es adecuado para grandes volúmenes de datos. Los valores faltantes se pueden completar usando constantes globales, pero este enfoque causa problemas al integrar los datos y, por lo tanto, no es un enfoque infalible. Los datos ruidosos se pueden manejar utilizando cuatro métodos, a saber, regresión, agrupamiento e inspección manual. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

1.2.3.3. Reducción de datos

Es el concepto de reducir el volumen de datos o reducir las dimensiones de los datos, es decir, la cantidad de atributos. Las técnicas de reducción de datos se utilizan para analizar datos en un formato reducido sin perder la integridad de los datos reales, pero aun así producen resultados de alta calidad. Además, esta técnica incluye compresión de datos,

reducción de tamaño y reducción de volumen. Utilice las técnicas de compresión de datos para lograr una representación comprimida o simplificada de los datos reales. Cuando los datos originales se recuperan de los datos comprimidos sin perder ninguna información, se denomina recuperación de datos sin pérdidas.

Por otro lado, si la recuperación de datos es solo parcial, entonces se denomina reducción de datos con pérdida. La reducción de dimensionalidad es la reducción de un número de atributos, y las técnicas incluyen transformaciones wavelet donde los datos originales se proyecta en un espacio más pequeño y la selección de subconjuntos de atributos, un método que implica la eliminación de atributos irrelevantes o redundantes. La eliminación de ruido es una técnica para reducir el volumen mediante la selección de datos alternativos más pequeños. La reducción de cantidad se logra utilizando métodos paramétricos y no paramétricos. En el método paramétrico, los datos reales no se almacenan, solo los parámetros. Los métodos no paramétricos conservan una representación simplificada de los datos originales.

1.2.3.4. Transformación de datos

Se refiere a convertir o combinar datos en un formato adecuado, transformarlos en información lógica y significativa para la gestión. Los verdaderos desafíos de la transformación de datos surgen cuando los campos de un sistema no coinciden con los campos de otro sistema. Antes de la transformación de datos, se realiza la limpieza y manipulación de datos. Las organizaciones recopilan cantidades masivas de datos y la cantidad de datos crece rápidamente. Transforme los datos recopilados utilizando herramientas ETL. La transformación de datos incluye las siguientes estrategias:

Smoothing, que elimina el ruido de los datos mediante la incorporación de técnicas de agrupación, clustering y regresión.
Agregación, que aplica resumen o agregación sobre los datos para dar un dato consolidado. (Por ejemplo, la ganancia diaria de una organización puede agregarse para obtener una facturación mensual o anual consolidada).
Generalización, lo que normalmente se considera como escalar en la jerarquía donde los atributos se generalizan a un nivel superior pasando por alto los atributos de un nivel inferior. (Por ejemplo, el nombre de la calle puede generalizarse como el nombre de una ciudad o una jerarquía de nivel superior, a saber, el nombre del país).

Discretización, que es una técnica en la que los valores sin procesar de los datos (p. ej., edad) se reemplazan por etiquetas conceptuales (ej., adolescente, adulto, mayor) o etiquetas de intervalo (ej., 0–9, 10–19, etc.)

1.3. Análisis del Big Data

Según (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021), se define como la implementación de técnicas de análisis sofisticadas. incluida la minería de datos, el análisis estadístico y el análisis predictivo, entre otras técnicas. Se refiere a la fase de examinar y analizar grandes cantidades de datos utilizando variables típicas para llegar a conclusiones, descubrir patrones y relaciones ocultos, tendencias y otra información comercial importante para mejorar la operatividad y buscar nuevos mercados y oportunidades.

Para (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021), el análisis de datos se introdujo para favorecer el análisis de datos usando marcos a escala, para ejecutarse en paralelo para extraer información de Big Data. Por lo tanto, brinda la oportunidad de transferir el comportamiento colectivo a métodos alternativos como los métodos de estimación de parámetros. Se define además como tuberías para la recolección, extracción, purificación, integración, agregación, así como visualización, análisis, modelado e interpretación.

Este proceso tiene cinco etapas principales: integración de datos, procesamiento, preprocesamiento, extracción de datos y presentación de información.

El análisis de Big Data se centra en extraer información significativa utilizando algoritmos eficientes en los datos capturados para procesar, analizar y visualizar los datos. Esto comprende enmarcar el algoritmo efectivo y el sistema eficiente para integrar datos, analizando el conocimiento así producido para hacer soluciones de negocio. Por ejemplo, en el comercio minorista en línea, el análisis de los enormes datos generados por las transacciones en línea es la clave para mejorar la percepción de los comerciantes.

Las tres fases principales del análisis son el análisis descriptivo, el análisis predictivo y el análisis prescriptivo, como se muestra en la Ilustración 6. El análisis predictivo predice lo que sucederá, mientras que el análisis descriptivo describe lo que ya sucedió. En todas las disciplinas, como la simulación, el aprendizaje, las estadísticas, las máquinas y la optimización, el análisis prescriptivo combina el análisis predictivo y descriptivo para

proporcionar información prospectiva y técnicas de aprovechamiento. La analítica avanzada es un conjunto integral de métodos analíticos como Big Data, Inteligencia Artificial (IA), Machine Learning, etc. (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021).

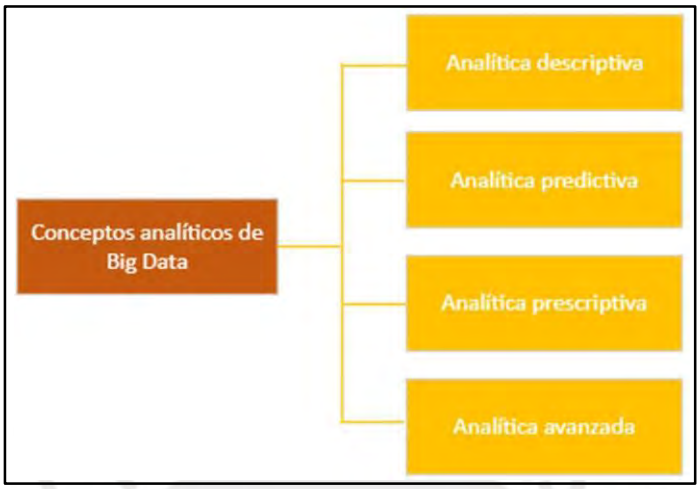


Ilustración 6. Analítica del Big Data

1.3.1. Terminología de Análisis de Big Data

1.3.1.1. Data Warehouse

También denominado Enterprise Data Warehouse (EDW), es un repositorio de los datos que recopilan varias organizaciones y empresas comerciales. Recopila los datos de diversas fuentes para que los datos estén disponibles para un acceso y análisis unificados por parte de los analistas de datos (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021).

Su función principal es proporcionar una vista unificada y coherente de los datos de una organización, independientemente de su ubicación o formato original. Para lograr esto, los datos se extraen, transforman y cargan (ETL) desde sistemas operativos, bases de datos transaccionales y otras fuentes de datos. Durante el proceso de transformación, los datos se limpian, se combinan y se reestructuran según un esquema de datos predefinido, lo que garantiza que estén estandarizados y listos para el análisis.

1.3.1.2. Business Intelligence

BI o Inteligencia Empresarial, es el proceso de analizar los datos que producen un resultado deseable para que las organizaciones y los usuarios finales tomen decisiones. Los beneficios del análisis de Big Data son aumentar los ingresos, mejorar la eficiencia y el rendimiento, y competir con los competidores comerciales al identificar las tendencias del mercado. Los datos de BI comprenden tanto los datos del almacenamiento (datos que se

capturan y almacenan previamente) como los datos que se transmiten, lo que ayuda a las organizaciones a tomar decisiones estratégicas. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

1.3.1.3. Analítica

Los científicos de datos analizan datos sin procesar para tomar decisiones comerciales. La inteligencia de negocios está más enfocada. El enfoque metódico destaca la diferencia entre la analítica de datos y la inteligencia empresarial. Ambos se utilizan con el fin de enfrentar los desafíos en el negocio y allanar el camino para nuevas oportunidades comerciales.

1.3.1.4. Métodos de análisis.

Se pueden dividir en métodos clásicos y modernos. Los métodos clásicos son el análisis de texto, el análisis de audio y el análisis de video. El análisis de texto se refiere a técnicas para extraer información de datos textuales. (Corea, 2019).

Para (Holmes, 2017), las redes sociales, correo electrónico, blogs, foros en línea, respuestas a encuestas, documentos de la empresa, noticias y otras fuentes de datos. El análisis de audio analiza datos de audio no estructurados y extrae información de ellos. Cuando se aplica al lenguaje hablado humano, el análisis de audio también se denomina análisis del habla.

El análisis del habla utiliza dos enfoques técnicos, la transcripción y el habla. Los sistemas de reconocimiento continuo de voz de vocabulario extenso (LVCSR) siguen un proceso de dos pasos, indexación y búsqueda. Los sistemas basados en el habla utilizan sonidos o fonemas. El análisis de video, conocido como análisis de contenido de video (VCA), cuenta con diferentes de técnicas para monitorear, analizar y extraer información significativa de los clips de video. Las técnicas modernas de análisis de datos, como los árboles de decisión, las redes neuronales y las máquinas de vectores de soporte, son adecuadas para identificar patrones lineales y no lineales en los datos (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021).

Las capacidades de análisis de Big Data se definen como “la capacidad de aprovechar los recursos y realizar tareas de análisis de negocios basadas en interacciones entre los activos de TI y otros recursos de la empresa. Arquitectura de análisis de Big Data de mejores prácticas que es flexible compuesto por cinco capas arquitectónicas principales

como se muestra en la Ilustración 8: (1) datos, (2) datos agregación, (3) análisis, (4) exploración de información y (5) gobernanza de datos.

La capa de datos contiene todas las fuentes necesarias para proporcionar la información que necesita. Además, apoya las operaciones diarias y resuelva los problemas comerciales. La capa de agregación es responsable de procesar datos de varias fuentes. En esta capa, los datos se procesan de manera inteligente a través de tres pasos: adquisición para leer datos entregados desde diferentes canales de comunicación, frecuencias, tamaños y formatos, limpieza de transformación, descomposición, traducción, fusión, clasificación y validación. (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021).

La capa de análisis es responsable de procesar todo tipo de datos y realizar los análisis apropiados. La capa de exploración de información genera resultados como una variedad de informes visuales, monitoreo de información en tiempo real y conocimientos comerciales significativos derivados de la capa de análisis para los usuarios de la organización. La capa de gestión consiste en la gestión de datos maestros (MDM), ciclo de vida, seguridad y protección de datos. La operación de análisis se desarrolla como se muestra en la Ilustración 7 (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021).

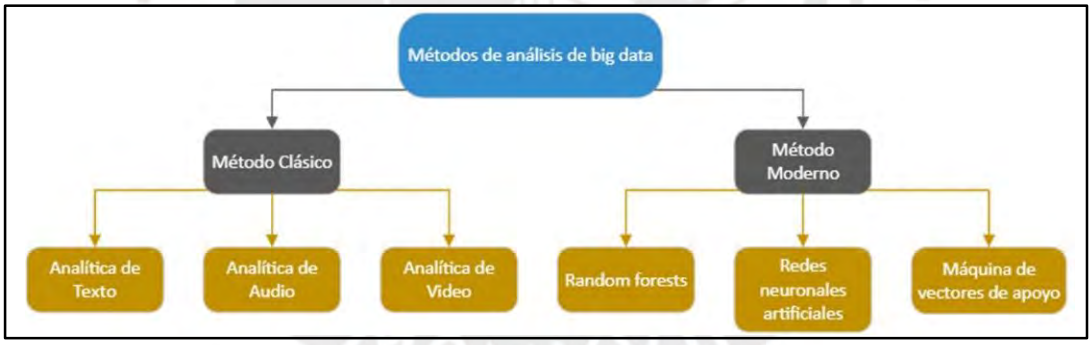


Ilustración 7. Métodos de Analítica de Big Data

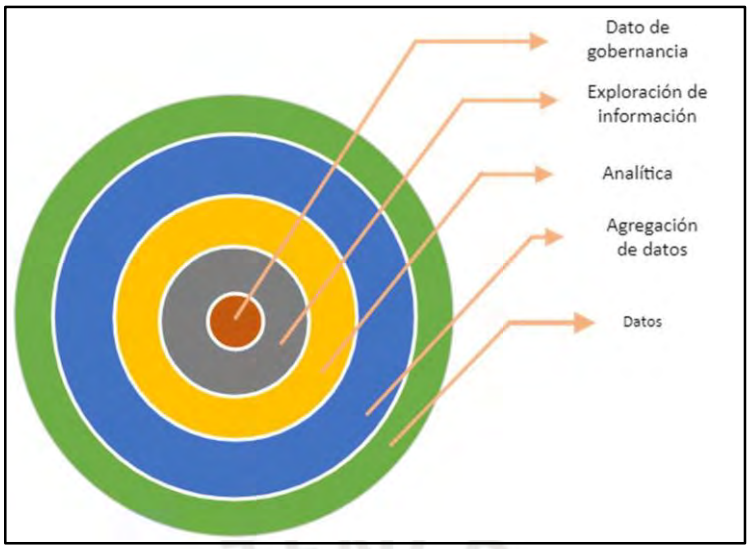


Ilustración 8. Arquitectura de Análisis Big Data

1.3.2. Técnicas de Análisis de Big Data

Varias técnicas de análisis involucradas en Big Data son:

- Análisis Cuantitativo
- Análisis Cualitativo
- Análisis Estadístico

1.3.2.1. Análisis Cuantitativo

Los datos cuantitativos son datos basados en números. La finalidad de este análisis estadístico es la cuantificación. En efecto la muestra de población puede generalizarse a toda la población de estudio. Hay diferentes tipos de datos cuantitativos que se someten a un análisis cuantitativo:

- Datos nominales: es un tipo de datos categóricos donde los datos se describen en función de categorías. Este tipo de dato no tiene valor numérico. No se pueden realizar operaciones aritméticas en este tipo de datos. Por ejemplo: sexo (masculino, femenino) y altura (alto, bajo).
- Datos ordinales: el orden o la clasificación de los datos es lo que importa en los datos ordinales, en lugar de la diferencia entre los datos. Se utilizan operadores aritméticos > y <. Por ejemplo, cuando se le pide a una persona que exprese su felicidad en una escala del 1 al 10, una puntuación de 8 significa que la persona es más feliz que una puntuación de 5, que es más que una puntuación de 3. Estos valores simplemente expresan el orden de felicidad. Otros ejemplos son las

clasificaciones que van desde una estrella hasta cinco estrellas, que se utilizan en varias aplicaciones, como la clasificación de películas, el consumo de corriente de un dispositivo electrónico y el rendimiento de la aplicación de Android.

- Datos de intervalo: en este punto no solo importa el orden de los datos, pero la diferencia entre ellos también importa. Uno de los ejemplos comunes de datos ordinales es la diferencia de temperatura en Celsius. La diferencia entre 50°C y 60°C es lo mismo que la diferencia entre 70°C y 80°C. A tiempo escala los incrementos son consistentes y medibles.
- Datos de razón: una variable de razón es esencialmente un dato de intervalo con la propiedad adicional de que los valores pueden tener cero absolutos. El valor cero en la razón indica que la variable no existe. La altura, el peso y la edad son ejemplos de datos de proporciones. Por ejemplo 40 de 10 años. Mientras que esos datos como la temperatura son variables de relación ya que 0°C no significa que la temperatura no exista.

1.3.2.2. Análisis cualitativo

Según (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021), es el análisis de datos en su entorno natural. Los datos cualitativos no se pueden reducir fácilmente a números. Las historias, los artículos, las reseñas de investigaciones, las transcripciones, las conversaciones, la música, los gráficos, el arte y las imágenes son datos cualitativos. El análisis cualitativo responde básicamente a las preguntas "cómo", "por qué" y "qué". Básicamente, existen dos enfoques para el análisis de datos cualitativos, a saber, deductivo e inductivo. El análisis deductivo es el uso de preguntas de investigación para agrupar datos investigados y luego buscar similitudes y diferencias entre ellos. Los métodos inductivos utilizan nuevos sistemas de consulta para agrupar datos y luego buscar relaciones entre ellos. Un análisis cualitativo tiene los siguientes tipos básicos:

A. Análisis de contenido: se utiliza con fines de clasificación, tabulación y resumen. El análisis de contenido puede ser descriptivo (¿qué son realmente los datos?) o interpretativo (¿qué significan los datos?).
B. Análisis narrativo: los análisis narrativos se utilizan para transcribir los datos de observación o entrevista. Los datos deben ser mejorados y presentados al lector en una forma revisada. Por lo tanto, la actividad central de un análisis narrativo es reformular los datos presentados por personas en diferentes contextos en función de sus experiencias.
C. Análisis del discurso: el análisis del discurso se utiliza para analizar datos, como textos escritos o una conversación natural. El análisis se centra principalmente en

cómo las personas utilizan los idiomas para expresarse verbalmente. Algunas personas hablan de manera simple y directa, mientras que otras personas hablan de manera vaga e indirecta.

D. Análisis del marco: el análisis del marco se utiliza para identificar el marco inicial, que se desarrolla a partir del problema en cuestión.
E. Teoría fundamentada: la teoría fundamentada básicamente comienza con el examen de un caso particular de la población y la formulación de una teoría general sobre toda la población.

1.3.2.3. Análisis Estadístico

El análisis estadístico utiliza métodos estadísticos para analizar datos. Las técnicas de análisis estadístico descritas son:

- Pruebas A/B
- Correlación
- Regresión

1.3.3. Inteligencia de Negocios de Big Data

Según (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021), es el proceso de analizar los datos y producir un resultado deseable para las organizaciones y los usuarios finales para ayudarlos a encontrar la mejor decisión. El beneficio del análisis de Big Data es aumentar los ingresos, aumentar la eficiencia y el rendimiento, y superar a los rivales comerciales al identificar las tendencias del mercado. Los datos de BI comprenden tanto los datos del almacenamiento (datos previamente capturados y almacenados) como los datos que se transmiten, lo que ayuda a las organizaciones a tomar decisiones estratégicas.

1.3.3.1. Procesamiento de transacciones en línea (OLTP)

Se se utiliza para procesar y administrar aplicaciones orientadas a transacciones. Las solicitudes se procesan en tiempo real y no por lotes; de ahí el nombre OLTP. Se utilizan en transacciones en las que se requiere que el sistema responda de inmediato a las solicitudes del usuario final. Como ejemplo, la tecnología OLTP se utiliza en aplicaciones de procesamiento de transacciones comerciales, como cajeros automáticos (ATM). Las aplicaciones OLTP se utilizan para recuperar un conjunto de registros y ponerlos a disposición de los usuarios finales; por ejemplo, una lista de productos de hardware que se venden en una tienda en un día determinado. Las aerolíneas, los bancos y los supermercados utilizan OLTP para muchas aplicaciones, incluidas la banca electrónica, el

comercio electrónico, las transacciones electrónicas, la nómina, los sistemas POS, los sistemas de reserva de aerolíneas y la contabilidad. Un solo sistema OLTP puede admitir miles de usuarios y las transacciones pueden variar de simples a complejas. Las transacciones OLTP típicas tardan segundos, no minutos, en completarse. Las características principales de los sistemas OLTP son la integridad de los datos mantenida en un entorno de acceso múltiple y el procesamiento rápido de consultas y la disponibilidad de transacciones en segundo lugar. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

El término “procesamiento de transacciones” se asocia con un proceso en el que una tienda minorista en línea o un sitio web, que procesa el pago de un cliente en tiempo real por los bienes y servicios adquiridos. Durante la OLTP, el sistema de pago del comerciante se conectará automáticamente con el banco del cliente, después de lo cual se realizarán verificaciones de fraude y otras verificaciones de validez, y se autorizará la transacción si se determina que es legítima.

1.3.3.2. Procesamiento analítico en línea (OLAP)

Según (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021), se utilizan para procesar consultas de análisis de datos y realizar análisis efectivos en cantidades masivas de datos. En comparación con OLTP, los sistemas OLAP manejan un número relativamente menor de transacciones. En otras palabras, las tecnologías OLAP se utilizan para recopilar, procesar y presentar a los usuarios comerciales datos multidimensionales para su análisis. Los diferentes tipos de sistemas OLAP son el procesamiento analítico en línea multidimensional (MOLAP), en línea relacional (ROLAP) y el híbrido (HOLAP), que es una comunicación de estas. Se los denomina mediante una definición de cinco palabras clave: Análisis rápido de información multidimensional compartida (FASMI).

- Se rápido, indica la velocidad a la que el sistema OLAP entrega respuestas a los usuarios finales, quizás en segundos.
- El análisis, es la capacidad del sistema para proporcionar una rica funcionalidad analítica. Se espera que el sistema responda la mayoría de las consultas sin programación.
- El uso compartido, se define como capacidad del sistema para admitir el uso compartido y, en ese momento, debe poder implementar los requisitos de seguridad para mantener la confidencialidad y la administración de acceso simultáneo cuando se requieren reescrituras múltiples.

- Multidimensional, es el requisito básico del sistema OLAP, que proporciona una vista multidimensional de los datos. Esta matriz multidimensional de datos se denomina comúnmente cubo.
- La información, se especifica como la capacidad del sistema, para manejar grandes volúmenes de datos obtenidos del almacén de datos.

En un sistema OLAP, se presenta la información en lugar de los datos al usuario final. La tecnología OLAP se utiliza en pronósticos y minería de datos. Se utilizan para predecir las tendencias actuales en las ventas y predecir los precios futuros de los productos básicos.

1.3.3.3. Plataforma de análisis en tiempo real (RTAP)

La aplicación de técnicas analíticas a los datos en movimiento transforma los datos en conocimientos empresariales e información procesable. La computación de transmisión es crucial para realizar análisis en movimiento de datos de múltiples fuentes a velocidades y volúmenes sin precedentes. La computación de transmisión es esencial para procesar los datos a diferentes velocidades y volúmenes, aplicar técnicas analíticas apropiadas en esos datos y producir información procesable instantáneamente para que las acciones apropiadas se puedan tomar de forma manual o automática.

Las aplicaciones de la plataforma de análisis en tiempo real (RTAP) se pueden usar para alertar a los usuarios finales cuando ocurre una situación y también les brindan las opciones y recomendaciones para tomar las medidas apropiadas. Las alertas son adecuadas en aplicaciones en las que el sistema RTAP no debe tomar acciones automáticamente. Por ejemplo, un sistema de monitoreo de pacientes alertaría a un médico o enfermera para que tome una acción específica para una situación. Las aplicaciones RTAP también se pueden usar en la detección de fallas cuando una fuente de datos no genera datos dentro del tiempo estipulado. Las fallas en ubicaciones remotas o problemas en las redes se pueden detectar usando RTAP. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

1.3.4. Procesamiento de análisis en tiempo real

La accesibilidad de nuevas fuentes como videos, imágenes y datos de redes sociales brinda una gran oportunidad para obtener información más profunda sobre los intereses de los clientes, productos, etc. La velocidad volumen de los datos tradicionales y nuevos generados son significativamente más altos que antes. Las fuentes de datos tradicionales incluyen los datos del sistema transaccional que se almacenan en RDBMS y formatos de archivo planos. Estos son en su mayoría datos estructurados, como transacciones de

ventas y transacciones de tarjetas de crédito. Para aprovechar al máximo el poder de la analítica, es necesario capturar cualquier tipo de datos, ya sean no estructurados o semiestructurados. Los datos de redes sociales, weblogs, datos de máquinas, imágenes y videos capturados desde cámaras de vigilancia y teléfonos inteligentes, datos de aplicaciones y datos de dispositivos sensores, en su mayoría no están estructurados. Las organizaciones que capturan estos grandes datos de múltiples fuentes pueden descubrir nuevos conocimientos, predecir eventos futuros y obtener acciones recomendadas para escenarios específicos, e identificar y manejar los riesgos financieros y operativos. Como se muestra en la Ilustración 9 muestra la arquitectura de procesamiento de datos tradicionales y nuevas, su procesamiento, análisis, conocimientos procesables y sus aplicaciones. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

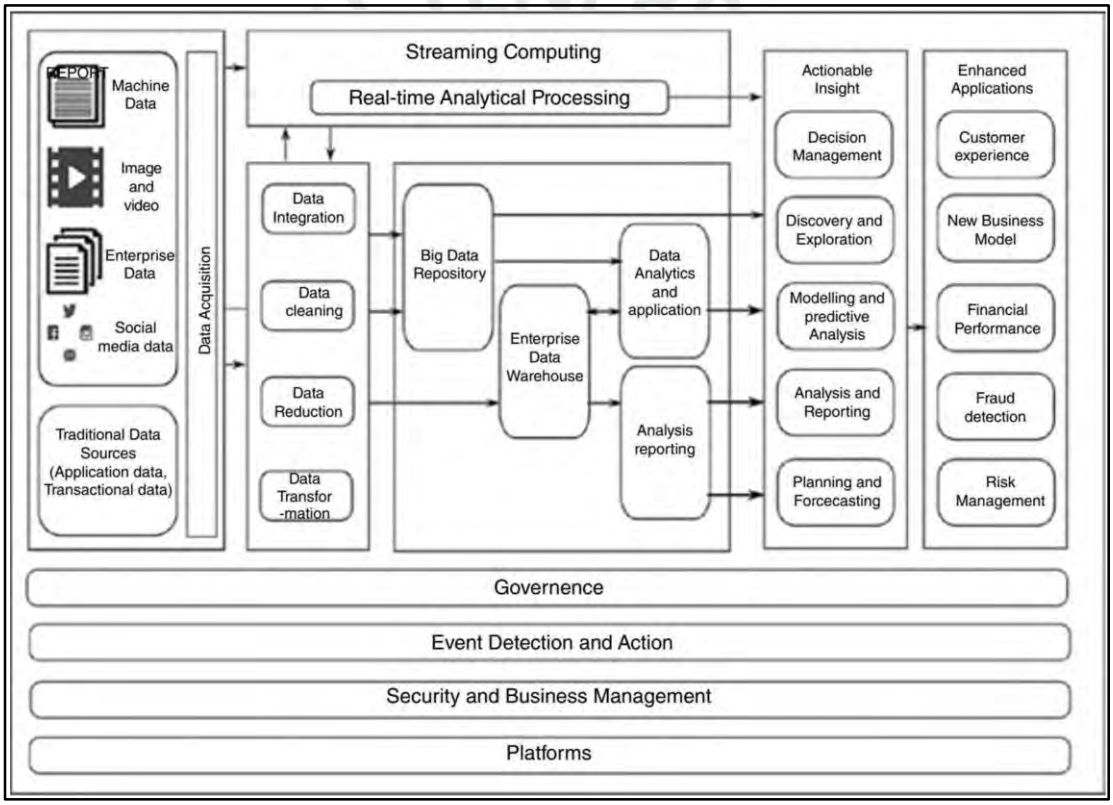


Ilustración 9. Arquitectura de Procesamiento de Análisis.

La información operativa compartida incluye datos maestros y de referencia, centro de actividad, centro de contenido y catálogo de metadatos. Los datos de transacciones son datos que describen eventos comerciales, como la venta de productos a clientes, la compra de productos a proveedores y la contratación y gestión de empleados. Los datos maestros son información comercial importante que respalda las transacciones. Los datos básicos son datos que describen a los clientes, productos, empleados, etc. involucrados en la transacción.

Los datos de referencia son datos asociados con una transacción con un conjunto de valores, como el estado del pedido del producto, el puesto del empleado o el código del producto. Content Hub es un destino único para los usuarios de la web que buscan contenido de redes sociales o cualquier tipo de contenido en forma de texto generado por el usuario o archivos multimedia. El Centro de actividades gestiona toda la información sobre las actividades recientes.

1.3.5. Almacén de datos empresarial

ETL (Extraer, Transformar y Cargar) se usa para cargar datos en el almacén de datos, primero se transforman antes de cargarlos, lo que requiere un hardware costoso separado. Otro enfoque rentable es cargar primero los datos en el almacén y luego transformarlos en la propio BD. El marco Hadoop ofrece una plataforma de procesamiento y almacenamiento de bajo costo que vuelca datos sin procesar directamente en HDFS y luego aplica técnicas de transformación a los datos.

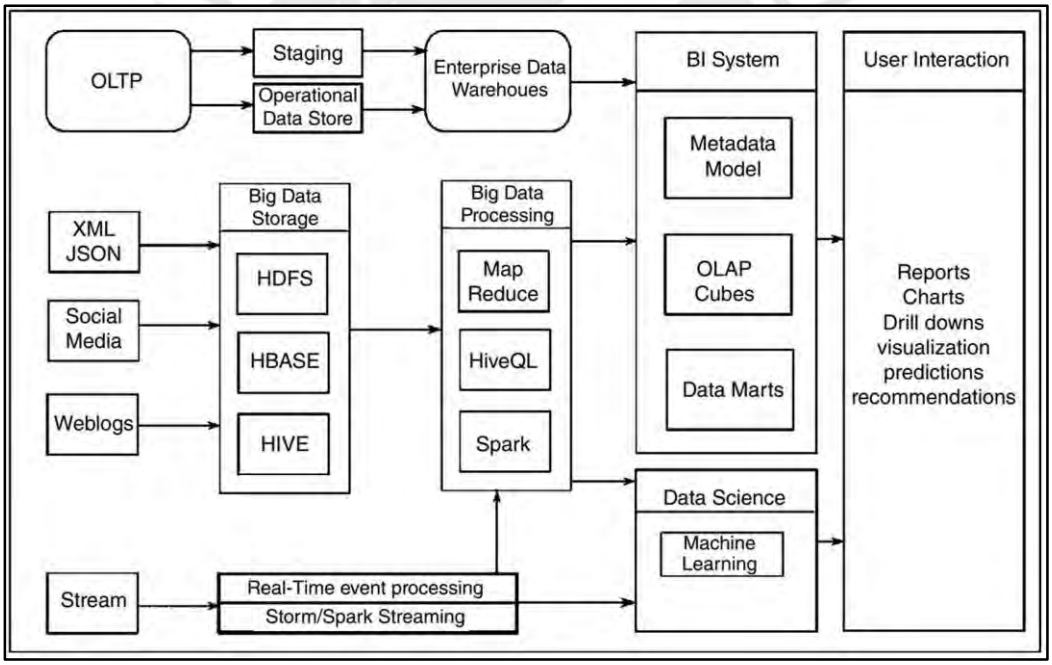


Ilustración 10. Arquitectura EDW integrada con tecnologías de Big Data.

En la Ilustración 10 muestra la arquitectura de un EDW integrado con tecnologías de Big Data. La capa superior del diagrama muestra un sistema de inteligencia empresarial tradicional con almacenamiento de datos operativos (ODS), base de datos provisional, EDW y varios otros componentes. La capa intermedia del diagrama muestra varias tecnologías para almacenar y procesar grandes volúmenes de datos no estructurados que

parten de múltiples fuentes. Se almacena en paradigmas de almacenamiento como HDFS, HBase y Hive y se procesa mediante paradigmas de procesamiento como MapReduce y Spark. Los datos procesados se pueden acceder a ellos directamente a través de sistemas de baja latencia. La capa inferior del diagrama muestra el procesamiento de datos en tiempo real. Las organizaciones utilizan técnicas de aprendizaje automático para comprender mejor a sus clientes, ofrecer un mejor servicio y proponer nuevas recomendaciones de productos. Más entrada de datos con mejores técnicas de análisis produce mejores recomendaciones y predicciones. El resultado es un reporte de visualización de datos presentados al usuario final. Además, se presentan predicciones y recomendaciones a las organizaciones. (Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi, 2021)

1.3.6. Lenguajes de Programación

Cuando se trata de análisis, debemos discutir las herramientas que necesita saber para realizar análisis efectivos para tomar buenas decisiones en el momento adecuado.

Actualmente hay muchos lenguajes en la ciencia de datos que utiliza un analista o científico de datos. Entre ellos están Python, R, SQL, Java, C y C++.

Python

Según (Yan, 2023), Python es un lenguaje creado en 1991 por Guido Wang. Este lenguaje de alto nivel es utilizado para desarrollar diversas aplicaciones. es un lenguaje de programación de alto nivel, interpretado y de uso general, reconocido por su facilidad de lectura y sencillez; además, facilita a los programadores la redacción de código de forma más rápida y sencilla debido a su sintaxis precisa y breve.

Ilustración 11. Lenguaje de programación Python

En ciencia de datos, Python se usa a menudo para manipular datos, implementar algoritmos de análisis de datos y entrenarlos para utilizarlos en el aprendizaje automático y profundo. Python es compatible con muchas estructuras de datos y utiliza una sintaxis sencilla en inglés, lo que lo convierte en un lenguaje ideal para los programadores principiantes.

R

Para (Gandrud), R fue creado en 1995 por Ross Ihaka y Robert Gentleman con el objetivo de proporcionar análisis de datos, estadísticas y modelos gráficos fáciles de usar. Python es de propósito general, mientras que R está más especializado para el análisis estadístico y la visualización intuitiva.

R está diseñado para manejar grandes conjuntos de datos y procesamiento complejo usando RStudio. Su sintaxis específica de estadísticas es intuitiva para los investigadores capacitados en estadística, y sus poderosas capacidades de visualización permiten una comunicación más intuitiva de los resultados.

Ilustración 12. Lenguaje de programación R

SQL

SQL (lenguaje de consulta estructurado) se se utiliza para crear bases de datos. Aunque este lenguaje todavía se encuentra entre los 20 principales de los índices de lenguaje de muchos científicos de datos, en realidad está siendo reemplazado por motores de base de datos NoSQL como Cassandra, Redis, Riak, HBase o Infinitegraph.

Ilustración 13. Lenguaje de programación SQL

1.3.7. Árbol de Decisión

Definición

Son algoritmos de aprendizaje automático es utilizado en la ciencia de datos para procesar grandes cantidades de datos y resolver problemas. También son llamados algoritmos estadísticos que permiten clasificar en función de determinadas características o atributos o hacer una regresión de las relaciones entre diferentes variables para realizar predicciones en modelos de análisis de datos predictivos de Big data. (Ghosh, S. & Koley, S., 2014)

El árbol de decisión es una estructura que está compuesta por ramas y nodos de distintos tipos:

Este es una representación gráfica de cómo se define un árbol de decisión. Que cuenta con:

- Nodos internos, son las características que se evalúan para tomar una decisión
- Las ramas, son las decisiones acertadas.
- Los nodos finales son el resultado.

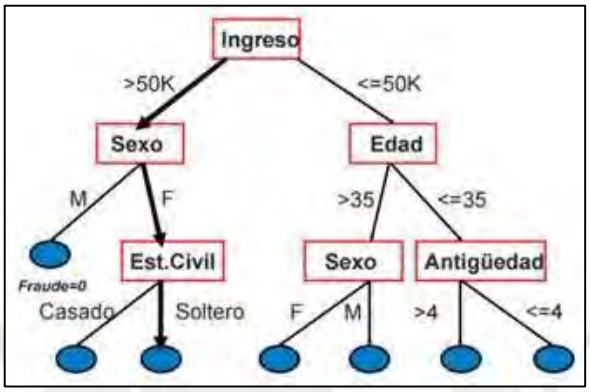


Ilustración 14. Ejemplo de Árbol de Decisión

Modelos

Existen varios modelos de árboles de decisión: regresión, clasificación, árboles embolsados, bosques aleatorios, etc. Los dos anteriores son muy similares excepto por el tipo de datos de la variable objetivo, los regresores se enfocan en variables numéricas y categóricas.

En un modelo de clasificación, el valor de una variable se predice clasificando sobre la base de otras variables. Podríamos predecir quién comprará un producto al clasificar entre consumidores y no consumidores o qué marca de computadora portátil comprará cada uno. Los valores esperados, es decir, están predeterminados. El conjunto de valores posibles se utiliza para determinar el resultado.

1.3.8. Clustering

Definición

Este procedimiento posibilita que los algoritmos de aprendizaje automatizado adquieran los datos que emplearán para realizar sus labores. Este procedimiento permite que las máquinas desarrollen su habilidad para analizar en cantidades grandes con menos fallos.

La función primordial del clustering consiste en reunir los datos para generar lo que se denomina clústeres. Cada uno de estos grupos reúne una serie de datos parecidas entre ellos.

Este algoritmo es utilizado en modelos de Machine Learning no supervisado. Y permite realizar las siguientes tareas:

- Analizar datos.
- Encontrar posibles errores.
- Dividir los datos en grupos similares para facilitar el proceso.

Modelos

Los métodos de clustering están enmarcados en las técnicas de machine learning y de aprendizaje no supervisado. Aunque existen multitud de métodos, los dos más conocidos son el K means y el Clustering jerárquico.

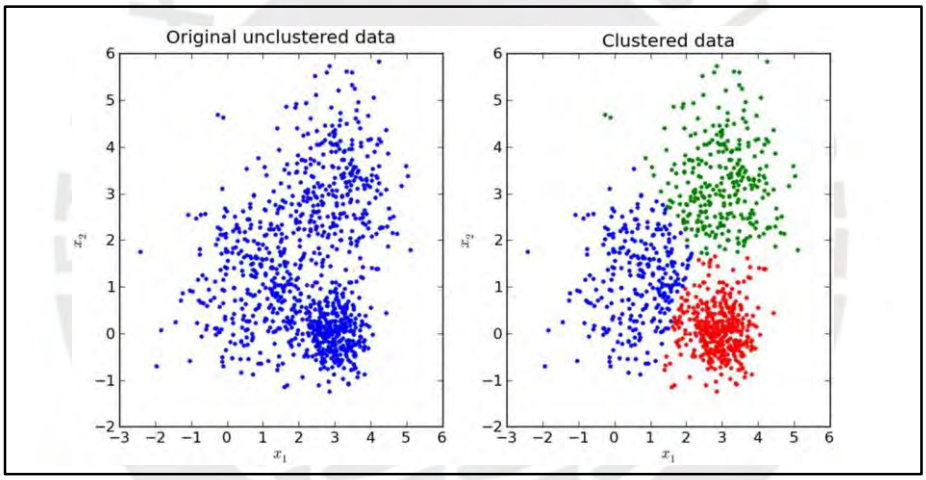


Ilustración 15. Ejemplo de Clustering

El K means clustering es un procedimiento no jerárquico empleado para reunir objetos y dividir el conjunto de datos en diversos clústeres. Además, no sean intrusivos. Un aspecto relevante de este procedimiento es que la cantidad de subgrupos o clústeres debe estar previamente definida antes de iniciar.

Clustering Jerárquico nos permite obtener representaciones basadas en árboles que se conocen como dendrogramas. Los dendrogramas son una representación que sirve para ilustrar la organización jerárquica entre diferentes elementos y que puede hacerlo de forma horizontal o vertical.

1.4. Machine Learning y Deep Learning en la agricultura

A causa del crecimiento poblacional, la necesidad de alimentos sigue en aumento, lo que convierte a la agricultura en una de las actividades más relevantes de la economía de una nación. Para satisfacer las crecientes necesidades de los cultivos, el sector agrícola necesita avances como hacer cálculos precisos con respecto a la producción de rendimiento y utilizar los mejores y más recientes equipos agrícolas. La agricultura digital, la agricultura precisa, la agricultura inteligente, la agricultura intensiva, la agricultura continua, la agricultura orgánica y la agroindustria son nombres para la agricultura moderna debido a estos avances (Rai, 2021).

La agricultura de precisión comprende en descubrir, calcular y actuar sobre las inconsistencias en el mismo campo y otros rendimientos. El objetivo principal del estudio de la agricultura de precisión es optimizar el rendimiento de los insumos y al mismo tiempo ahorrar recursos y establecer un sistema de apoyo a la evaluación para la gestión general en los cultivos. La predicción del clima y el efecto de fertilizantes son otros factores importantes en este campo. Y para obtener información de ellos se utiliza sensores remotos y de gran precisión (Pierce, F. J., & Nowak, P., 1999).

Los rendimientos agrícolas están relacionados con la agroindustria. Este término tiene las siguientes características como mejoramiento continuo, producción de rendimiento, agroquímicos, equipos, provisión de semillas y estrategias de comercialización y distribución. Wang y col. (2006) introdujo sensores inalámbricos en la agricultura y la industria alimentaria.

Otro aspecto significativo de la agricultura moderna es el manejo de los problemas en términos de rendimiento, impacto atmosférico, seguridad alimentaria y sostenibilidad en las circunstancias imperantes. Como la demanda mundial global está constante aumento, es necesario incrementar la producción de cultivos añadiendo la disponibilidad a tiempo y alta calidad nutricional. Esto puede alcanzarse salvaguardando el ecosistema natural a través de prácticas de agricultura sostenible. La gestión agrícola se fundamenta en identificar, evaluar y responder a la variación creciente entre campos. (FAO, 2009).

Actualmente la industria agrícola actual se enfrenta a desafíos como el inadecuado tratamiento de las granjas, diferentes dolencias que prevalecen en los animales, infección por plagas, riego irregular, etc. el rendimiento y también resultan peligrosos para el ecosistema debido al uso excesivo de productos químicos en él. No se puede proporcionar una respuesta unificada a todos los porblemas. Para tratar estos asuntos, la ecología

irregular debe ser abordada a través de la observación e investigación de todos los elementos y sucesos. Una solución a esta situación consiste en emplear la inteligencia artificial y el aprendizaje automático. El aprendizaje automático aporta a los agricultores información para incrementar la producción agrícola, reducir los costos de adquisición y compensar las pérdidas sufridas durante los desastres naturales. (Pierce, F. J., & Nowak, P., 1999) revisaron una nueva técnica de aplicación informática en el sector agrícola y alimentario para aumentar la calidad del producto. Además, (Rai, 2021) revisó la visión artificial y su aplicación en el rubro alimentos y la agricultura.

1.4.1. Machine Learning

El aprendizaje automático es un campo interdisciplinario que combina la informática y la estadística y se utiliza principalmente para el análisis y la clasificación, así como para tareas que normalmente realizan los humanos. Para hacer esto, necesitamos entrenar a las computadoras para que resuelvan problemas del mundo real con la mayor precisión. El aprendizaje automático se puede utilizar en diversas situaciones, como análisis facial, juegos de computadora, recuperación de información, pronósticos del mercado de valores, biología computacional, análisis de micromatrices de ADN para la clasificación del cáncer, detección de ataques epilépticos, optimización de datos en centros de servicio, clasificación automática de texto, aplicaciones y genética. Análisis de datos de expresión (Pierce, F. J., & Nowak, P., 1999). (Rai, 2021) dedujeron el algoritmo de reconocimiento facial con la ayuda del aprendizaje automático. La Ilustración 16 representa el modelo de trabajo de aprendizaje automático.

1.4.1.1. Aprendizaje Supervisado

Como su nombre lo específica, el aprendizaje supervisado significa que se requiere la presencia de un supervisor para realizar la tarea. Generalmente, la máquina se entrena utilizando los datos recopilados que están bien etiquetados y se conocen como datos etiquetados, de modo que este algoritmo puede analizar los datos de entrenamiento y dar un resultado correcto utilizando datos etiquetados (Rai, 2021). El algoritmo utilizado en el funcionamiento del aprendizaje supervisado se representa en la Ilustración 17.

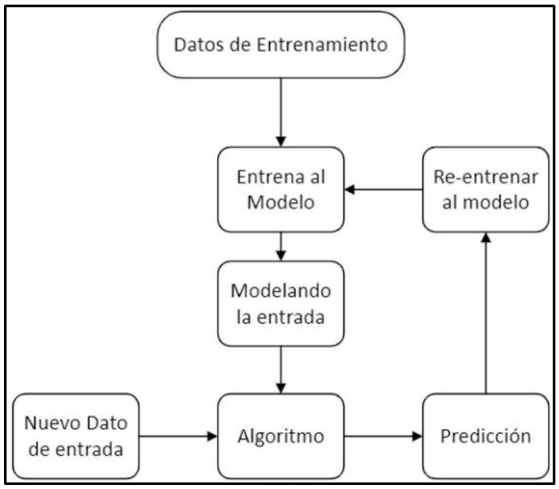


Ilustración 16. Entrenamiento de datos

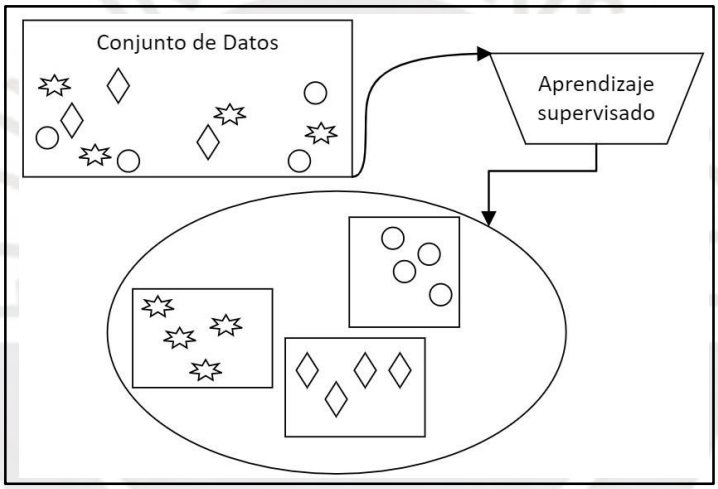


Ilustración 17. Aprendizaje Supervisado

El aprendizaje supervisado se divide en dos categorías: regresión y clasificación.

La regresión es una técnica utilizada para encontrar relaciones entre variables independientes y dependientes. La clasificación es el proceso de dividir los datos en clases separadas y definidas, asignamos una etiqueta a cada clase (Rai, 2021). El modelo de trabajo de aprendizaje supervisado de clasificación como se indica en la Ilustración 18.

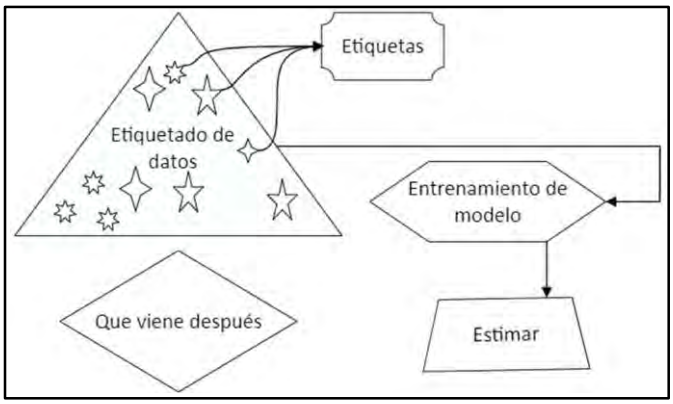


Ilustración 18. Modelo de entrenamiento de aprendizaje supervisado

1.4.1.2. Aprendizaje No Supervisado

En este aprendizaje no es necesario supervisar el modelo, sino que el trabajo se inicia con su propia información, donde se trata principalmente de datos no etiquetados. El algoritmo de aprendizaje no supervisado ofrece un mejor resultado al realizar tareas complejas en comparación con el aprendizaje supervisado (Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A., & Bhansali, S., 2019). Es más impredecible y también ayuda a encontrar el patrón desconocido en los datos. La Ilustración 19 representa el aprendizaje no supervisado y se puede clasificar en agrupación y asociación.

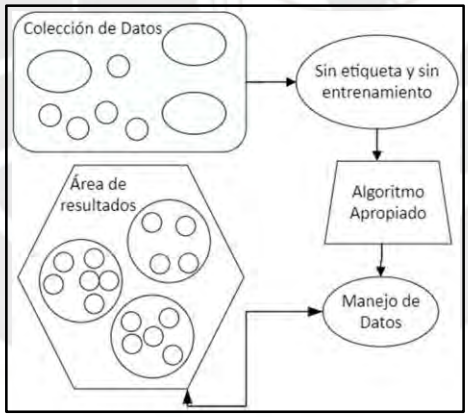


Ilustración 19. Aprendizaje No Supervisado

1.4.1.3. Aprendizaje Reforzado

En este aprendizaje el agente tiene la capacidad de interactuar con el entorno y encontrar un mejor resultado. Para ello, sigue fórmulas de hit y trail. Este aprendizaje se utiliza cuando no hay una forma adecuada de realizar una tarea, pero el modelo debe seguir algunas reglas estrictas para cumplir con su deber. En este tipo de aprendizaje no se requieren etiquetas. Existen dos tipos positivo y negativo (Roberge, 2021). El modelo de trabajo del aprendizaje por refuerzo se muestra en la Ilustración 20.

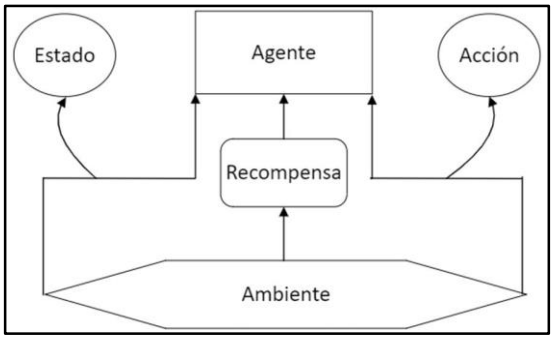


Ilustración 20. Aprendizaje Reforzado

1.4.2. Algoritmos de Machine Learning

A medida que el aprendizaje automático crece día a día, muchos autores utilizan varios algoritmos de aprendizaje automático para resolver problemas complejos. El aprendizaje automático es el campo de investigación actual en el que se están realizando muchas investigaciones (Rai, 2021). La Ilustración 21 muestra los diferentes algoritmos de aprendizaje automático. Algunos de ellos se explican brevemente.

1.4.2.1. Red Neuronal Artificial

En sus siglas (ANN), se forma utilizando una gran cantidad de elementos conocidos como neuronas, donde cada neurona toma decisiones simples y pasa esas decisiones a otras neuronas. Todas las neuronas están interconectadas entre sí, donde la interconexión entre estas neuronas se conoce como función de red. Una red neuronal poco profunda tiene tres capas: capa de entrada, capa oculta y capa de salida. La ANN no supervisada también se conoce como mapa de autoorganización. En esta red solo se proporcionan capas de entrada, donde la red necesita desarrollar sus propios incentivos de entrada, por lo que mediante la autoorganización se pueden recopilar todos los datos de entrada y encontrar las características inherentes al problema (Roberge, 2021).

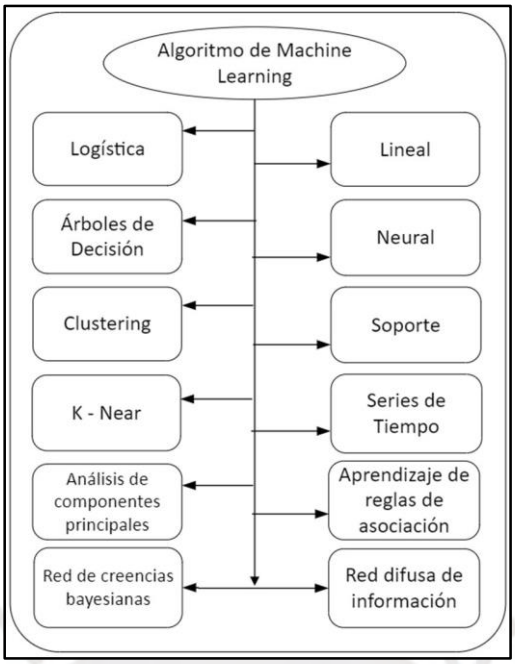


Ilustración 21. Algoritmo de Machine Learning

1.4.2.2. Máquinas de Vectores de Soporte (SVM)

Se emplea comúnmente este algoritmo para hallar respuestas a problemas de clasificación y regresión. El propósito principal del algoritmo SVM es establecer el límite de decisión capaz de segmentar el espacio n-dimensional en clases, para que en el futuro podamos categorizar nuevos datos en la categoría adecuada. El hiperplano es el límite de la decisión. El algoritmo SVM se emplea para identificar la cara, categorizar imágenes, categorizar texto, entre otros. Los algoritmos de SVM lineales y no lineales son distintos (Roberge, 2021).

Considere un felino con rasgos similares a los de un perro, por lo que para establecer si es un felino o un perro, se elabora un modelo empleando el algoritmo SVM. El primer paso del modelo consiste en reconocer imágenes de perros y gatos, de manera que pueda adquirir conocimientos sobre distintas particularidades de estos animales. Posteriormente, examina la criatura inusual, en la que el modelo establece un límite de decisión entre el gato y el perro y selecciona un caso extremo; es decir, tomará en cuenta el caso extremo de gato o perro por el vector de soporte.

1.4.2.3. Clustering

La agrupación implica agrupar datos para que cada grupo tenga datos similares. Es básicamente una recopilación de datos basada en sus similitudes y diferencias. La agrupación en clústeres es importante porque define agrupaciones básicas entre datos no

etiquetados. En ausencia de mediciones bien agrupadas, depende de cómo el usuario utilice los datos. Los métodos basados en la densidad, los métodos jerárquicos, los métodos de división y los métodos basados en cuadrículas son métodos de agrupamiento. Los métodos basados en la densidad tratan los clústeres como regiones comprimidas que comparten algunas similitudes distintas con las del resto del espacio. Estos métodos tienen una alta precisión y la capacidad de fusionarse en dos grupos. El enfoque jerárquico crea una estructura en forma de árbol basada en la jerarquía. Los nuevos grupos se crean utilizando grupos antiguos. Estos métodos se dividen en dos categorías: aglomerativos y divisivos (Rai, 2021).

1.4.2.4. Árbol de decisión

El análisis de árboles de decisión es una amplia herramienta de modelado analítico con aplicaciones en varios campos. Los árboles de decisión se construyen utilizando un enfoque algorítmico que identifica conjuntos de datos particionados en función de varios criterios. En un árbol de decisión, la partición de datos debe ser continua según ciertos factores. El algoritmo está descrito por dos factores llamados nodos de decisión y hojas. El nodo de decisión es el nodo donde se dividen los datos y las hojas son los resultados finales. Los árboles de clasificación y regresión son los principales tipos de árboles de decisión. El objetivo principal del algoritmo es crear un modelo que prediga la variable objetivo utilizando reglas simples. (Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A., & Bhansali, S., 2019).

1.4.2.5. Análisis de componentes principales

El análisis utiliza una transformación ortogonal en un proceso estadístico para transformar variables correlacionadas en variables no correlacionadas. El análisis de componentes se utiliza para estudiar la relación entre un conjunto de variables. Este algoritmo se utiliza para considerar grandes conjuntos de datos de variables interrelacionadas y seleccionar el mejor se ajuste al modelo. Este tipo de centralización de variables se denomina reducción de dimensionalidad. Este enfoque ayuda a reducir la complejidad del conjunto de variables. Este análisis también se llama análisis factorial general. (Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A., & Bhansali, S., 2019).

1.4.3. Aplicaciones de Machine Learning en la Agricultura

Se muestra las siguientes aplicaciones utilizadas:

- Predicción de rendimiento
- Detección de plagas y enfermedades

- Detección de malezas
- Manejo del suelo
- Reconocimiento una planta
- Gestión la calidad de cultivo
- Bienestar de los animales
- Previsión de ganado

1.4.3.1. Predicción de rendimiento

Existen numerosos elementos mediante los cuales un agricultor puede conseguir rendimientos óptimos en su labor agrícola. Uno de estos elementos es estimar el desempeño del cultivo. Este elemento abarca la fertilidad del terreno, el método de irrigación, las condiciones meteorológicas y la gestión de plagas. Si el agricultor no respeta adecuadamente estos cuatro aspectos durante la tarea agrícola, hay un alto peligro de perjudicar el cultivo. Veamos algunos de los modelos que se utilizan en el sector agrícola.

- La aplicación de aprendizaje automático ayuda a contar el número de semillas de café en una rama y también segrega los frutos del café en tres categorías de cosecha, no cosecha y semillas con etapa de maduración ignorada. Además, podemos estimar el peso de las semillas y el porcentaje de maduración de las semillas de café. (Ramos, P. J., Prieto, F. A., Montoya, E. C. & Oliveros, C.E., 2017) demostraron cómo medir automáticamente los frutos de café del cafeto a través de un sistema de visión artificial (MVS). Durante el desarrollo del cultivo, en la fase inicial y cuando no se llevó a cabo la cosecha, empleando la técnica MVS, evidenciaron que la estimación del conteo de semillas no será ni excesivamente alta ni excesivamente baja. Con esto, evidenciaron un valor de correlación superior de 0,90.
- (Amatya, S., Karkee, M., Gongal, A., Zhang, Q. & Whitting, M. D., 2016) diseñó un MVS que automáticamente mueve y sacude los árboles y recoge los frutos de la cereza durante la fase de recolección, además de identificar las ramas ocluidas y las cerezas que no son evidentemente perceptibles. En cambio, durante la recolección de cerezas se necesita más trabajadores, que constituyen aproximadamente el 50% de su costo de producción anual. Para disminuir este gasto, se han empleado tecnologías de recolección mecanizadas, tales como actuadores de ramas que provocan vibraciones en los frutos de la cereza, permitiendo su liberación de las ramas.

Esta herramienta ha generado una nueva era en el sector de la horticultura, ya que tiene mayor eficiencia y es económica para los agricultores. Los agricultores pueden utilizar esta técnica en su trabajo agrícola para aumentar la productividad.

1.4.3.2. Detección de plagas y enfermedades

La gestión de plagas y enfermedades es uno de los retos más significativos de la agricultura contemporánea. Uno de los enfoques para controlar enfermedades y plagas consiste en aplicar plaguicidas de manera homogénea sobre los cultivos. Esto demanda una alta eficacia, pero no resulta económico, y también conlleva el peligro de efectos secundarios como la polución del agua subterránea y el daño a la fauna y al ecosistema.

- (Ebrahimi, M.A., Khoshtaghaza, M.H., Minaei, S. & Jamshidi, B., 2017) diseñaron un aparato que facilita la detección de parásitos en el ambiente del invernadero mediante el procesamiento de imágenes. Este método SVM puede emplearse para la categorización y la orientación de parásitos. La técnica de procesamiento de imágenes y el método, con la elección adecuada de provincia e índice de color, demostraron ser eficaces en la identificación de metas con gran eficacia.
- (Moshou, D., Bravo, C., Jonathan, W., Wahlen, S., Cartney, M. A. & Ramona, H., 2015) diseñó un aparato óptico sencillo y económico para supervisión a distancia de dolencias, fundamentado en la reflectancia en múltiples bandas de ondas. Examinaron la distinción entre plantas saludables y enfermas durante las primeras fases de la enfermedad de la roya amarilla, mediante imágenes de campo obtenidas al situar un espectrógrafo en el punto de resonancia del rociado. Posteriormente, mediante la normalización de la intensidad, podemos reducir la elevada variabilidad espectral provocada por la estructura del dosel en distintos niveles de luz. El modelo de discriminación cuadrática se utiliza clasificar los espectros de salud y enfermedad tiene un elevado índice de éxito.

1.4.3.3. Detección de malezas

Para conseguir un rendimiento óptimo, prevenir las malas hierbas es una de las tareas primordiales. Es complicado distinguir las malezas de los cultivos, por lo que se recurre al aprendizaje automático a través de sensores inteligentes. Esta técnica lleva a la identificación y prevención exactas de las malas hierbas, con un costo reducido y que no perjudica al medio ambiente.

- Pant utilizó la teledetección para la clasificación de especies y la cartografía operacioanal de malezas. La exposición y el mapeo de parches de malezas de

Silybum marianum por medio de un mapa autoorganizado ordenado se informa utilizando una cámara multivisionaria que proporciona imágenes de alta resolución transportadas por el sistema de aeronaves no tripuladas (UAS).

1.4.3.4. Manejo del suelo

La administración de la tierra, tanto directa como indirectamente, tiene un rol crucial para asegurar la eficacia de las cosechas, la estabilidad del ecosistema y la salud de las personas. El suelo es un recurso natural diverso con procesos y mecanismos de difusión complejos, donde la temperatura del suelo también tiene una función importante en el estudio preciso del cambio climático en una región y su comportamiento ecológico. Los algoritmos de aprendizaje automático son utilizados para la medición de temperatura y humedad del suelo para explicar la dinámica del ecosistema y su efecto en la agricultura.

- (Ghosh, S. & Koley, S., 2014) introdujeron una nueva técnica llamada red de retro propagación que da mejores resultados para determinar las buenas propiedades del suelo en lugar de utilizar el método tradicional llamado modelo de regresión multivariante. El objetivo de esta técnica es entrenar el cultivo particular que tiene ciertas propiedades.

1.4.3.4. Reconocimiento de una planta

En comparación con el método tradicional de clasificación de plantas comparando las características de las hojas, el aprendizaje automático puede proporcionar más información sobre las características de las hojas mediante el análisis de la morfología de las venas de las hojas, lo que da como resultado resultados más rápidos y precisos. El objetivo principal es identificar y clasificar automáticamente diferentes especies de plantas, para evitar la experiencia humana y reducir el tiempo de clasificación.

- (Grinblat, G. L., Lucas, C. U., Mónica, G. L. & Granitto, P. M., 2016) emplearon una red de convolución profunda para el desafío de identificar plantas basándose en patrones de venas foliares. Se tomaron en cuenta tres tipos de leguminosas: frijol blanco, frijol rojo y patrones de venas foliares de soja, en los que se empleó la forma de las venas para recopilar la información de la hoja. Es uno de los instrumentos clave para identificar plantas basándose en su color y forma.
- (Weiss, U., Biber, P., Laible, S., Bohlmann, K. & Zell, A., 2010) odeló una metodología para distinguir las especies vegetales usando un sensor LiDAR tridimensional de baja resolución. Los autores han modelado un grupo de rasgos que poseen estadísticas comunes que no dependen del tamaño de la planta. Los

clasificadores han sido entrenados y comparados en este modelo con el conjunto de características que muestra una alta eficiencia en la identificación.

1.4.3.5. Gestión de la calidad del cultivo

Para incrementar el valor de la cosecha y minimizar el desperdicio, es necesario categorizar la calidad de la cosecha con el error más bajo posible. El cultivo se desarrolla en la penúltima subcategoría para identificar características vinculadas con la clase de cultivo.

- (Zhang, M., Changying, L. & Fuzeng, Y., 2017) desarrollaron un modelo para la detección y clasificación del material extraño botánico y no botánico enraizado dentro de la fibra de algodón en el momento del proceso de recolección.

1.4.3.6. Manejo de riego

El riego es una parte importante de la agricultura. Desempeña un papel importante en la productividad del rendimiento. No riegue demasiado ni demasiado poco, pero mantenga el equilibrio. Se deben tener en cuenta algunos factores como el tipo de suelo, la topografía del suelo, el clima, el tipo de cultivo, la calidad del agua, etc. para mantener estas condiciones.
El neuro goteo de (Hinnell, A. C., Lazarovitch, N., Furman, A., Poulton, M. & Warrick, A. W., 2010) es un algoritmo ANN basado en Excel diseñado para proporcionar una ilustración rápida de los patrones e índice de humedad del suelo a partir de emisores de riego por goteo de superficie.

1.4.3.7. Bienestar de los animales

Se encarga de la salud y la protección de los animales con el fin de preservar el balance en el ecosistema. El uso principal del aprendizaje automático es supervisar la conducta de los animales durante la exposición precoz a la infección.

- (Datta, R., Smith, D., Rawnsley, R., Bishop- Hurley, G., Hills, J., Timms, G. & Henry, D., 2015) siguió un marco de aprendizaje automático de dos etapas que es un método eficaz para clasificar el comportamiento del ganado. La tecnología de sensores de ganado y los clasificadores de ensamblaje se utilizan en el enfoque actual para categorizar y examinar los cambios de comportamiento en el ganado para mejorar su alimentación.
- (Pegorini, V., Karam, L.Z. & Pitta, L.S.R., 2015) propusieron una técnica basada en datos recopilados por sensores de rejilla de Bragg de fibra óptica que se proyectan

mediante la técnica de aprendizaje automático (clasificación de patrones). En este estudio, han considerado el proceso de masticación y la ingesta de alimentos del suplemento dietético. Además, se consideraron dos factores más del heno y el raigrás que son rumiadores y la inactividad para el comportamiento de ingestión. Demostraron que la clasificación de patrones diferencia los cinco patrones involucrados en el proceso de masticación.

1.4.3.8. Previsión de ganado

La producción de animales de granja se ocupa del problema del sistema de producción. El alcance principal de las aplicaciones de aprendizaje automático en esta área es el juicio preciso de los saldos monetarios, con la ayuda de los cuales los productores pueden obtener información basada en el monitoreo de la línea de producción y, por lo tanto, pueden obtener ganancias. Esto se debe a que los algoritmos de aprendizaje automático tienen el potencial para la detección temprana y la advertencia de problemas que brindan la información previa a los productores.

- (Morales, I.R., Cebrián, D. R., & Blanco, E. F., 2016) estudiaron cómo detectar la producción de huevos. SVM se utiliza para el reconocimiento de problemas en la producción de huevos con la ayuda de datos de producción de huevos de gallinas ponedoras. Al usar la configuración de parámetros ópticos de un modelo SVM, se puede indicar una alerta con un día de anticipación que puede ser útil para el diagnóstico preventivo de síntomas clínicos.
- (Craninx, M., Fieveza, V., Vlaeminck, B. & De Baets, B., 2008) introdujeron el algoritmo de aprendizaje automático para pronosticar el patrón de fermentación ruminal de los ácidos grasos de la leche. El mecanizado de la precisión de predicción del modelo dado se realiza con un modelo de regresión que depende de los ácidos grasos de la leche tanto impares como de secuencia múltiple.

1.4.4. Deep Learning

El aprendizaje profundo es un método moderno que se ha utilizado en varias aplicaciones en agricultura. Tiene varias aplicaciones, como el procesamiento de imágenes y la clasificación de texto. Dado que la tasa de éxito del aprendizaje profundo es muy alta en otros dominios, también se aplica a los métodos agrícolas. El aprendizaje profundo cubre varias capas de redes neuronales diseñadas para realizar tareas más cultas. Algunos de los modelos de aprendizaje profundo brindan resultados notables y, en términos de escala, no se corresponden con los humanos (Grinblat, G. L., Lucas, C. U., Mónica, G. L. & Granitto, P. M., 2016).

Cada capa toma como entrada el resultado de la anterior, y toda la red se entrena en una única secuencia. Una plataforma de aprendizaje profundo es una plataforma que ayuda a los usuarios a crear una variedad de arquitecturas de aprendizaje profundo o les permite aplicar fácilmente el aprendizaje profundo a una amplia gama de aplicaciones y servicios comerciales. Una de las principales diferencias entre el aprendizaje automático y el aprendizaje profundo es que el aprendizaje profundo requiere más datos para la clasificación, mientras que, en el aprendizaje automático, los datos pequeños son suficientes para la clasificación. Las herramientas de aprendizaje profundo más populares son theno, kera, tensorflow, py-torch y toolbox.

Algunos ejemplos son el procesamiento de imágenes y la clasificación de texto. Las Ilustraciones 22 y 23 muestran un modelo de trabajo de clasificación de imágenes y texto.

Algunos de los arquitectos del aprendizaje profundo se muestran en la Ilustración 24.

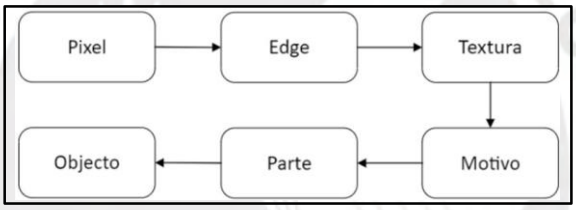


Ilustración 22. Modelo de clasificación de imágenes

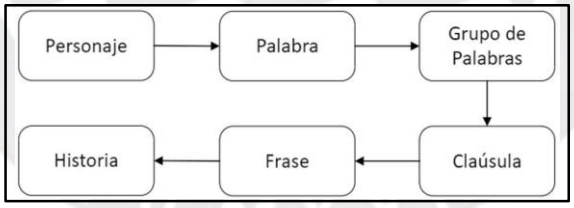


Ilustración 23. Modelo de clasificación de texto

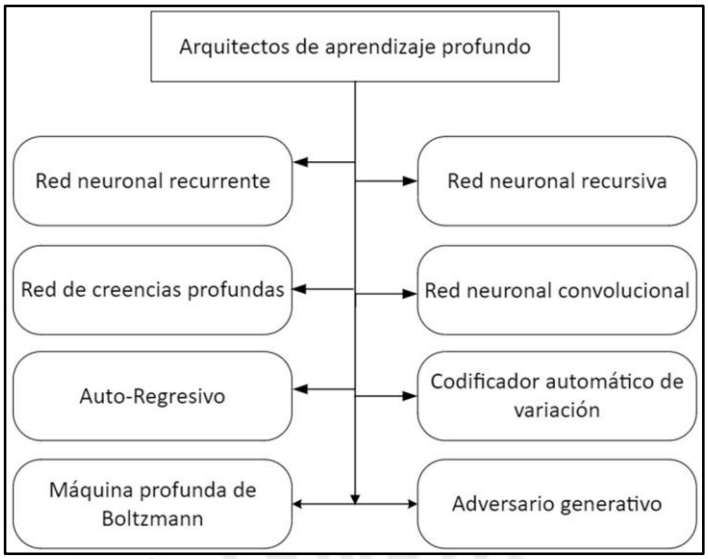


Ilustración 24. Arquitectura de aprendizaje profundo

1.4.4.1. Redes neuronales de convolución

La estructura de la red neuronal de convolución (CNN) se basa en una red neuronal de retroalimentación y está diseñada en una corteza animal y utiliza perceptrón de múltiples capas para este proceso. En CNN se utiliza a menudo la cantidad mínima de funciones de activación de unidad lineal rectificadas de preprocesamiento. Las aplicaciones generales incluyen reconocimiento de imagen/video, procesamiento de lenguaje natural, ajedrez, etc. La convolución se usa para encontrar características similares usando diferentes posiciones de imagen. Esto se hace usando filtros de aprendizaje que se pasan a través de los datos/imagen de entrada. La técnica utilizada para expandir los conjuntos de datos y mejorar la precisión de la CNN se denomina aumento de datos. Siempre que el gran conjunto de datos sea grande y aceptablemente grande, CNN mejorará la precisión de la clasificación correcta. Algunas aplicaciones de ANN son la decodificación de reconstrucción facial, el análisis de documentos, la recopilación de antecedentes y ambientales, la comprensión del clima, la publicidad, etc. (Aboul Ella Hassanien, Ashraf Darwish Editors, 2021).

1.4.4.2. Red neuronal recurrente

(Grinblat, G. L., Lucas, C. U., Mónica, G. L. & Granitto, P. M., 2016) mencionan que la red neuronal recurrente (RNN) es un tipo de red neuronal donde la salida del bucle anterior se considera como entrada para el bucle actual. Las aplicaciones generales de la red neuronal generativa son el reconocimiento de voz, el reconocimiento de escritura a mano, el análisis de secuencia de datos, etc. Además, la red neuronal generativa genera automáticamente

códigos de programación que dan un objetivo predefinido. El proceso de trabajo de RNN consiste en proporcionar información al modelo. La representación de los datos en la capa de entrada se calcula y se envía a la capa oculta, donde realiza el modelado de secuencia y el entrenamiento en direcciones hacia adelante o hacia atrás. También se pueden usar varias capas ocultas, sin embargo, la capa oculta final envía el resultado procesado a la capa de salida. La memoria RNN a corto plazo es actualmente un modelo RNN popular. Es eficaz en la secuencia de datos que requiere memoria o detalles de los últimos eventos. Algunas de las aplicaciones de RNN son el modelado y la predicción de idiomas, el reconocimiento de voz, la traducción automática, el reconocimiento de imágenes y la traducción. Estas redes se conocen como células. Estas celdas consideran la entrada del estado anterior como entrada actual y también deciden qué información debe tenerse en cuenta y cuál debe descuidarse. La condición anterior, la memoria y la entrada actuales se combinan para predecir la siguiente salida.

1.4.4.3. Redes generativas de confrontación

La novedad de las redes generativas adversarias (GAN) radica en el tecnicismo de su diseño. Es un tipo de aprendizaje automático no supervisado que incluye innovación computarizada para comprender las similitudes o prototipos de datos produciendo el resultado. Las GAN son modelos inteligentes para construir un sistema productivo modelando un problema que tiene dos submodelos como parte del aprendizaje supervisado. Son sistemas generativos que pueden educarse para crear ilustraciones. Las GAN son el escenario estimulante y en rápido ascenso. Funcionan por su potencial de modelo práctico generativo. La vecindad del dominio del sistema está relacionada con trabajos de conversión de imagen a imagen, por ejemplo, convertir imágenes de una estación en otra (por ejemplo, escenas de verano en escenas de invierno) o imágenes diurnas en nocturnas, al crear imágenes fotorrealistas de elementos, escenas e individuos que los individuos reconocen como falsificados (Grinblat, G. L., Lucas, C. U., Mónica, G. L. & Granitto, P. M., 2016).

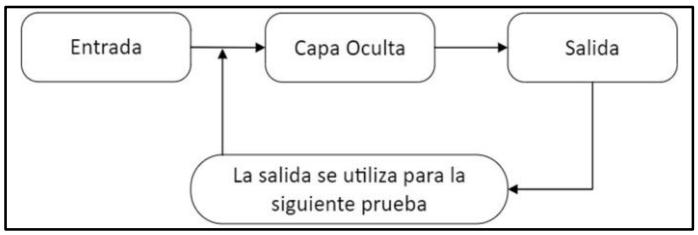


Ilustración 25. Red Generativa Adversaria (GAN)

1.4.5. Aplicación del aprendizaje profundo en agricultura

El aprendizaje profundo ha elevado el sector agrícola a su nuevo nivel. A su vez utiliza técnicas como la red neuronal convencional, RNN y GAN dando mejores resultados y ayuda al control de la agricultura. También emplea la tecnología de procesamiento de imágenes y la investigación de la información con eficacia. El intenso crecimiento en el campo del aprendizaje profundo ha mostrado muchos buenos resultados en general, pero está emergiendo como una bendición en el campo de la agricultura. Al comparar el aprendizaje profundo con un procedimiento común predominante, se puede decir que este método está superando la técnica de procesamiento de imágenes de uso común existente y tiene una mayor precisión. La inclinación profunda permite que los sistemas matemáticos que se componen de muchas etapas de procesamiento representen información con muchos niveles de abstracción. La red neuronal y la propagación inversa son la base del aprendizaje profundo (Rai, 2021).

1.4.5.1. CNN

Se emplea extensamente en el sector agrícola debido a su amplia capacidad para el procesamiento de imágenes. Se pueden categorizar las aplicaciones como la clasificación de plantas o cultivos, la predicción de plagas y rendimiento, la recolección de robots, la supervisión de catástrofes, entre otras. Berkley Vision and Learning Center ha desarrollado un nuevo marco de aprendizaje profundo para construir un modelo de detección de enfermedades. Este sistema es capaz de identificar alrededor de 10 a 15 casos de hojas enfermas de hojas sanas, también tiene la capacidad de separar las hojas de las plantas del entorno (Rai, 2021).

En 2007, para controlar e identificar malezas, se diseñó un enfoque que es una combinación de aprendizaje de características CNN y K-mean. El modelo manual para la detección de malezas conduce a un reconocimiento falso y una habilidad de extracción débil en la extracción de características. Para la división óptica de imágenes y la posterior restauración de la información faltante en una serie temporal de imágenes de satélite, se utilizan mapas Kohonen autoorganizados. En este método de configuración de posprocesamiento, se utilizan análisis geoespacial y varios algoritmos de filtrado. Aunque CNN tiene varios usos, enfrenta muchos desafíos que han ralentizado su aplicación en la clasificación de plantas. Por ejemplo, cada píxel de las imágenes de SAR transmitidas por el espacio se caracteriza por una fase de retrodispersión e intensidad en múltiples polarizaciones. Para la predicción del rendimiento y la recolección robotizada, el conteo de frutas es uno de los factores importantes. No podemos producir resultados satisfactorios

mediante el recuento tradicional o el recuento de imágenes de cámara o vídeo y, además, estos procesos requieren mucho tiempo. El preprocesamiento de este tipo de imágenes es un desafío debido a la oclusión y la iluminación. (Hansen, M. F., Smitha, M. L., Smitha. L. N., Michael, G., Salterb, M. G., Baxterc, E. M., Farishc, M. & Grieved, B., 2018) introdujeron una técnica para identificar al animal de ganado, como el cerdo, utilizando la función de reconocimiento facial de las CNN. Convencionalmente, las etiquetas de identificación por radiofrecuencia se usaban para detectar los animales, lo que anteriormente era un trabajo engorroso.

Para acompañar a una red de convolución completa, se propuso un método conocido como detección de manchas. El primer paso es recopilar las etiquetas formadas por humanos de un conjunto de imágenes de frutas y luego este modelo se entrena para un rendimiento de segmentación de imágenes. Luego, se usa CNN para contar las imágenes bifurcadas y dar una aproximación del número de frutas.

La última etapa del trabajo consiste en aplicar una ecuación de regresión para mapear la estimación del conteo de frutas intermedio con el conteo final de etiquetas generado por humanos. La precisión y la eficiencia aumentan al combinar el aprendizaje profundo con la detección de manchas.

El método de clasificación de la tierra se utiliza para identificar la tierra como uso y cobertura, para la evaluación del riesgo de desastres y para la alimentación y la agricultura. La idea general del método de aprendizaje profundo es integrar información desarrollada por múltiples fuentes heterogéneas utilizando técnicas de aprendizaje automático para proporcionar procesamiento de información y capacidad de representación. Este proceso incluye cuatro pasos: (i) filtrado de ruido y agrupamiento de datos, (ii) limpieza de la cobertura terrestre, (iii) posprocesamiento de mapas, (iv) análisis geoespacial. (Kussul, N., Lavreniuk, M., Skakun, S. & Shelestov, A., 2017) se implementa un enfoque de aprendizaje profundo multinivel para clasificar la cobertura terrestre y los tipos de cultivos utilizando imágenes satelitales multitemporales de múltiples fuentes.

Actualmente, se están introduciendo nuevas tecnologías en el ámbito agrícola; como vehículos aéreos no tripulados para el procesamiento de imágenes de alta resolución. Es bastante complicado para los agricultores manejar máquinas altamente autónomas. No resulta sencillo manejar estas máquinas sin supervisión. Por tanto, la detección del riesgo en tiempo real con estas máquinas de forma automática con alta fiabilidad se convierte en un requisito. Para la sostenibilidad de la tierra, es necesario tomar en cuenta algunas

condiciones. Se planea reducir las emisiones de CO2, minimizar la degradación de la tierra y potenciar los beneficios económicos a través de la información real de satélites. Para la toma de decisiones en la agricultura de precisión, la CNN y el algoritmo genético se han vuelto útiles en la traducción de imágenes de satélite.

La predicción del tiempo es uno de los elementos más relevantes que los agricultores pueden anticipar mediante el uso de la CNN. Además, la valoración del cultivo es un elemento crucial para los agricultores, los consumidores y los gobiernos y debe realizarse antes de la cosecha. Las CNN también es utilizado para la clasificación del comportamiento animal en la agricultura de precisión y la agroindustria, y las CNN y los algoritmos genéticos se han convertido en un método conveniente para la traducción utilizando imágenes satelitales. (Rai, 2021).

1.4.5.2. RNN

Uno de los desafíos en la agricultura es aprovechar la tecnología para identificar el tipo y calidad de tierra que se cuenta. Para tener un aspecto bueno de las tierras depende de algunos factores como el clima.

Para resolver los desafíos asociados a RRJ, se propone un modelo denominado NARX, que se refiere a un modelo autorregresivo no lineal con entrada exógena. En este modelo los valores previos se consideran entradas y los valores actuales exógenas. El sistema no solo analiza las entradas autónomas sino también la respuesta previa, lo que incrementa la potencia del sistema. (Kurumatani, 2018) propuso una técnica para pronosticar el precio de un producto agrícola utilizando RNN.

Esta técnica también se usa para pronosticar el clima. (Biswas, S. K., Sinha, N., Purkayastha, B. & Marbaniang, L., 2014) diseñaron tres modelos para la predicción del clima: autorregresivo no lineal con redes neuronales de entradas exógenas (NARX NN), modelo de razonamiento basado en casos y modelo de razonamiento basado en casos de segmentos. (Palangpour, P., Venayagamoorthy, G. K. & Duffy, K., 2006) produjo un modelo para identificar la ubicación de los animales en el bosque. En este modelo se utilizó el algoritmo de optimización de enjambre de partículas combinado con el modelo RNN, los resultados obtenidos por este modelo contienen menos errores.

1.4.5.3. GAN

GAN se considera una de las redes neuronales más útiles en muchos campos. Principalmente, GAN se utiliza para encontrar la pérdida de características en el procesamiento de imágenes causada por muestreo descendente. Cuando se comprime la imagen, es posible que se pierda parte de la información o se pierda la calidad de esa imagen, por lo que es posible que debamos recuperar todos los detalles originales. Para esta recuperación, se define una función de pérdida perpetua que comprende la pérdida adversa y la pérdida de contenido. A continuación, esta función se compara con la pérdida de error cuadrático medio (MSE) ampliamente utilizada. Mientras trabaja en una gran cantidad de imágenes, este modelo tiene como función mejorar la calidad de imágenes muy comprimidas. Esto se vuelve importante con todos los modelos que contienen trabajos de procesamiento de imágenes, principalmente en agricultura, porque ciertas aplicaciones dependen de imágenes de teledetección (Rai, 2021).

(Barth, R., Ijsselmuiden, J. M. M., Hemming, J. & Van Henten E. J., 2017) Sugirió un modelo para vencer las dificultades vinculadas a la amplia gama de datos recogidos en sistemas de aprendizaje profundo. En ausencia de información marcada manualmente, se utiliza una gran cantidad de datos (como en el modelo de aprendizaje profundo o en el modelo basado en GAN). A esto se le llama ciclo no supervisado, o sistema adversario generativo, para optimizar la practicidad de las imágenes agrícolas artificiales. Los autores han propuesto 10.500 fotografías empíricas artificiales, 50 anotadas empíricamente y 225 sin etiquetar para que su modelo funcione. La hipótesis planteada fue que existía una semejanza entre imágenes sintéticas e imágenes empíricas que pueden mejorarse cualitativamente para mejorar la transformación de características. Debido a este análisis, las imágenes artificiales se transformaron fácilmente en características locales como la difusión de la luz, el color y la consistencia en comparación con la traducción de características globales, lo cual no fue tan bueno.

1.4.6. Ventajas y desventajas en la agricultura

Generalmente, en el aprendizaje automático no es fácil analizar los datos no estructurados. La aplicación de métodos de aprendizaje profundo será más útil si podemos utilizar diferentes tipos de formatos de datos para el funcionamiento del algoritmo. Para encontrar conexiones entre diferentes áreas en diferentes disciplinas, podemos usar algoritmos de aprendizaje profundo. Los trabajadores a menudo están cansados, son irresponsables o ignoran cosas pequeñas, pero este no es el caso con los modelos de aprendizaje profundo. El algoritmo realizará miles de ciclos de trabajo en poco tiempo sin errores. (Rai, 2021).

El enfoque de aprendizaje tradicional, la identificación de características debe ser precisa, mientras que en los modelos de aprendizaje profundo tienen la capacidad de crear nuevas características por sí mismos. Generalmente, la resolución de problemas en el aprendizaje automático se realiza dividiendo las tareas grandes en tareas pequeñas y combinando los resultados de todas las tareas pequeñas para el resultado final, mientras que en el aprendizaje profundo las tareas se resuelven de un extremo a otro; y además necesita una gran porción de datos y es costoso.

Una de las principales desventajas es que no podemos encontrar cómo se realiza el análisis dentro del modelo. Generalmente, lo llamamos caja negra, pero a veces es importante conocer el algoritmo de análisis porque la interpretabilidad es necesaria en algunos dominios. Hoy en día, a medida que el aprendizaje automático está creciendo en todos los dominios de manera dramática, el principal temor es que el aprendizaje automático pueda llevar todo el trabajo de los humanos y llevar a los humanos al desempleo o la esclavitud (Rai, 2021).

1.5. Base de Datos

1.5.1. Definición

(Mannino, 2007) cita que es una colección de datos interrelacionados organizados de tal manera que se puede acceder a ellos automáticamente. En el siguiente ejemplo, una empresa que fabrica productos y registra de toda la información relacionada con ella, incluidas las materias primas utilizadas, proveedores, clientes, empleados, eventos en el proceso de producción, ventas, facturación, etc. Toda esta información conforma la base de datos de la empresa y tiene que alimentar información relevante y útil para los otros departamentos.

Para elaborar facturas es necesario obtener datos de clientes y productos, realizar pedidos, vincular proveedores y materias primas, realizar estudios de costes, datos de productos, materias primas y personal.

Para realizar todas las tareas requeridas, tales como guardar información, actualizar datos, recuperar información cuándo y de qué manera le interese, elaborar informes, entre otros, se requiere de un software denominado "base de datos".

Los hipervisores han experimentado una transformación radical desde la creación del concepto de base de datos en los años 60, generando diversos tipos de controladores para

satisfacer distintas demandas. Los administradores de "bases de datos bibliográficas" son los más habituales, cuya labor es ordenar y manejar la información en texto plano. Además, la utilización de esta herramienta facilita la búsqueda y elección de datos basándose en palabras del texto y palabras clave o descriptores vinculados a cada documento.

Para poder procesar la información se han desarrollado procedimientos para adaptar la teoría relacional y una matemática compleja que permite hacerse optimizar la organización y lograr resultados muy impresionantes a pesar de la complejidad del problema.

1.5.2. Características de la BD

Las bases de datos contienen datos simples y complejos. También guardan fotos, huellas dactilares, videos de productos, y resúmenes de libros. La recopilación de grandes cantidades de datos ha sido posible gracias a la popularización de internet y la automatización de la acumulación de registros. Solo basta con un clic del mouse para acceder a los datos almacenados. La gestión de bases de datos se ha vuelto una tarea de gran relevancia para muchas organizaciones (Mannino, 2007).

La BD tiene las siguientes características:

- Es Persistente: en otras palabras, los datos residen en un dispositivo de almacenamiento estable, como cualquier disco. Las organizaciones necesitan mantener estables los datos de los clientes, proveedores e inventario porque se siguen utilizando en tiempo actual. La variable de un programa de computadora no es estable porque reside en la memoria principal y se elimina una vez que se cierra el programa. De cualquier manera, su persistencia no significa que los datos coexistan en un estado o espacio físico para siempre; se eliminan cuando ya no se necesitan y generalmente se en cendran cuando un proveedor ya no está en operación.

  La durabilidad se basa en la relevancia de la aplicación buscada. Es crucial realizar un seguimiento del tamaño, espacio y tiempo, dado que el almacenamiento y conservación de datos es costosa, solo se deben conservar los datos pertinentes para la decisión.
- Es Compartida: se puede tener varios destinos y usuarios. Las bases de datos proporcionan memoria compartida para diversas funciones organizativas. Por ejemplo, una base de datos de empleados se puede usar para calcular salarios, realizar revisiones de desempeño, determinar los requisitos de informes públicos y más.

Múltiples usuarios pueden acceder a la base de datos simultáneamente; por ejemplo, muchos clientes pueden reservar una aerolínea al mismo tiempo. A menos que dos usuarios intenten modificar la misma parte de la base de datos al mismo tiempo, uno puede continuar sin esperar al otro.

- Tiene Correlación: significa que los datos guardados pueden ser vinculados como entidades distintas para conseguir una imagen más integral. Como ejemplo, para su procesamiento, una base de datos de clientes conecta los datos de un cliente (nombre, dirección, etc.) con los datos de un pedido (número de pedido, fecha del pedido, etc.).

Para evidenciar estas características, tomamos una base de datos sencilla universitaria, tal se muestra en la Ilustración 26. Esta base de datos simplificada incluye datos sobre alumnos, problemas, programas, propuestas de programas e inscripción. También abarca procesos como el registro de cursos, la asignación de docentes a los cursos proporcionados, el registro de notas y los horarios de los cursos proporcionados (Mannino, 2007).

Las relaciones en la base de datos de la universidad sirven para responder a preguntas como:

- ¿Qué ofertas están disponibles para un curso en el periodo académico actual?
- ¿Quién es el instructor de un curso ofrecido?
- ¿Qué estudiantes están inscritos en un curso?

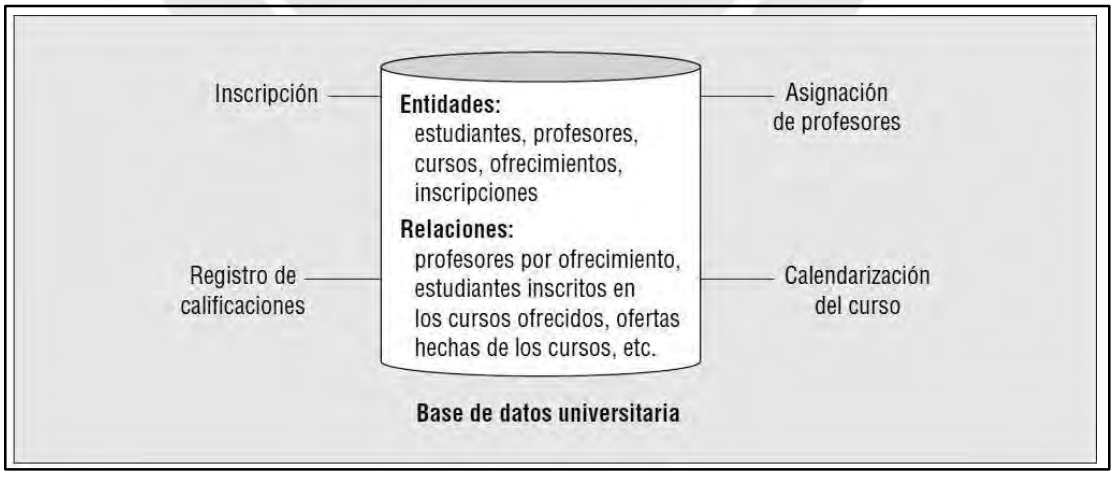


Ilustración 26. Figura simplificada de la base de datos universitaria

Ahora considere la base de datos del hospital en la Figura 27. Las bases de datos de los hospitales ayudan a los médicos a tratar a los pacientes. Los médicos hacen un diagnóstico y desarrollan un plan de tratamiento basado en los síntomas. Múltiples proveedores de atención médica leyeron y participaron en los registros médicos del paciente. Las enfermeras controlan los síntomas y administran medicamentos. El personal de la concesionaria de alimentos gestiona y prepara las comidas de acuerdo con el plan de comidas. Los médicos recetan nuevos tratamientos en función de los resultados de los tratamientos anteriores y los síntomas del paciente.

Las relaciones en la base de datos sirven para responder a preguntas como:

- ¿Qué síntomas son los más recientes de un paciente?
- ¿Quién registró el tratamiento a un paciente?
- ¿Qué diagnóstico hizo el médico a un paciente?

Las bases de datos deben tener muchas más entidades, relaciones y usos. Sin embargo, estas bases de datos simples tienen las características de las bases de datos comerciales: datos persistentes, múltiples usuarios y objetos conectados, múltiples relaciones entre entidades.

Ilustración 27. Base de datos de un hospital

Capítulo 2: Estudios de Casos

Según la FAO (2009), para 2050, la población aumentará significativamente y se prevé un aumento del 70 % en la producción de alimentos. La inseguridad alimentaria, la degradación del suelo y la contaminación del agua afectan la agricultura india. El desarrollo sociocultural y el cambio climático agregan incertidumbre a la seguridad alimentaria. El agricultor no puede vender su cosecha debido a la sobreproducción porque la cosecha no llega a tiempo. Es necesario mejorar los pronósticos de demanda y el desempeño.

Una de las alternativas y soluciones para el seguimiento del desempeño y uso del Sistema Global de Navegación por Satélite (GPS), que permite el posicionamiento preciso de puntos de medición en campo para crear mapas espaciales variables para la agricultura de precisión. (Agtech, 2017).

La agricultura sostenible está íntimamente relacionada con la gestión de los cultivos, la mejora de la gestión existente y las tareas de toma de decisiones. Con big data, existe la necesidad de almacenar y procesar grandes cantidades de datos para que puedan manipularse en tiempo real. Aquellos que registran y procesan datos de alimentos utilizando técnicas como algoritmos heurísticos y redes neuronales. (Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A., & Bhansali, S., 2019). Su objetivo es brindar a los agricultores asesoramiento y orientación sobre el rendimiento de los cultivos, la respuesta y el uso de fertilizantes. Mediante el uso de pronósticos meteorológicos, monitoreo de cultivos y detección de plagas, las empresas pueden administrar mejor, capturar valor económico y aumentar las ganancias generales.

(Bendre, M. R., Thool, R. C., & Thool, V. R., 2015), menciona que los datos de registro estaban sucios, interrumpidos, inexactos e incompletos. Estos datos se distribuyen en grandes datos y no son fáciles de ajustar a grandes conjuntos de datos. Para este propósito, se utiliza la limpieza de datos para mejorar la calidad de los datos, lo que ayudará en gran medida a las organizaciones a garantizar que los datos estén listos para la fase de análisis. El rápido crecimiento de los datos crea nuevas oportunidades para los negocios y los procesos. Los expertos en calidad de datos estiman que las empresas gastan entre el 40 % y el 50 % de su presupuesto en el proceso de limpieza de datos, que requiere mucho tiempo y esfuerzo. Los datos de mala calidad pueden afectar negativamente la eficacia de la organización. Un aporte importante es que un

buen proceso de limpieza conduce a mejores análisis y resultados que ayudan a determinar indicadores de producción y calidad para que la organización pueda tomar buenas decisiones sobre su proceso.
Según (Zhao, D., & Li, Y. R., 2015), la humanidad está confrontando un gran desafío de como incrementar la producción para lograr la seguridad alimentaria durante el siglo XXI y hacer una agricultura sostenible. Los problemas del cambio climático, agotamiento de recursos hídricos y el potencial de una mayor erosión ocasionan una pérdida de productividad.

Big data ayuda a integrar bases de datos ubicadas en diferentes redes para la gestión y sostenibilidad de campos y cuencas. La automatización y el uso de inteligencia artificial (IA), Internet de las cosas (IoT), drones, robots y big data están en el corazón de un "gemelo digital" global que ayudará a desarrollar la conservación y gestión específicas del sitio que aumentarán los ingresos. para sistemas agrícolas y prácticas de cultivo para la sostenibilidad global (Delgado, J. A., Short, N. M., Roberts, D. P., & Vandenberg, B., 2019).

AP se introdujo a fines de la década de 1990 cuando John Deere equipó sus tractores y máquinas con sensores GPS para la gestión de la información. Con la llegada del Internet de las Cosas al mundo de la tecnología, todos los dispositivos ahora están conectados e interactúan entre sí a través de una infraestructura de red inalámbrica. (Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.J., 2017). La AP es una extensión de este desarrollo y es la principal fuerza impulsora de la analítica de Big Data en la agricultura (Lesser, 2014). AP se centra en la recopilación, gestión y uso de datos para la toma de decisiones. Los puntos de acceso requieren múltiples tecnologías que trabajen juntas para permitir la recopilación y el análisis de datos. Estas tecnologías incluyen sistemas de posicionamiento geográfico (GPS), sistemas de información geográfica (GIS), sensores remotos, mapeo geográfico, sensores, equipos de comunicaciones electrónicas y tecnología de velocidad variable (VRT).

La teledetección en la agricultura genera gigabytes de datos recopilados de satélites, aeronaves y drones equipados con sensores e instrumentos de imágenes como el radar y el espectrorradiómetro de imágenes de resolución moderada (MODIS). Analice miles de imágenes utilizando algoritmos de procesamiento de imágenes y modelos de aprendizaje automático para resolver problemas de predicción y clasificación. Esta información se presenta luego en un formato fácil de entender. Un ejemplo de esto es NVDI (Índice de vegetación de diferencia normalizada), un índice

gráfico utilizado para estimar el verdor de la vegetación a partir de mediciones de detección remota.

La inspección visual de los campos de cultivo es una tarea muy importante en la agricultura de precisión, y su automatización implica la adquisición y el almacenamiento de grandes volúmenes de imágenes de diversas fuentes, como satélites, drones y equipos terrestres, así como el procesamiento de imágenes mediante técnicas de aprendizaje automático. para encontrar patrones, hacer medidas. y monitorear el comportamiento de los cultivos.

En (Saraiva, M., Protas, É., Salgado, M., & Souza Jr, C., 2020) Utilizando algoritmos basados en redes neuronales convolucionales profundas, las imágenes satelitales se procesan para ubicar, mapear y cuantificar automáticamente los sistemas de riego clave en diferentes escalas y etapas de crecimiento de los cultivos. Las ventajas de un sistema de este tipo son que se pueden mapear grandes áreas en un tiempo relativamente corto a un costo relativamente bajo y proporcionar una herramienta para monitorear el crecimiento de las áreas regadas y la cantidad de equipos de riego central instalados. La información satelital adquirida regularmente permite combinar la inteligencia artificial con métodos de series temporales, como la identificación de patrones de crecimiento de cultivos y la clasificación de tipos de cultivos en función de índices de vegetación derivados de información de banda multiespectral óptica. Las imágenes de alta resolución permiten tareas más complejas, como la estimación del rendimiento en función del conteo de frutas, utilizando algoritmos de detección de objetos en imágenes capturadas por drones.

Tomando como ejemplo a Perú, cada vez más empresas agrícolas están implementando tecnologías de Big Data e inteligencia artificial en sus operaciones para ayudar a incrementar la rentabilidad del sector, entre las que se pueden destacar los siguientes esfuerzos:

A. (Mayhua, 2016), propone una red de sensores inalámbricos con tecnología ZigBee en una topología de red para monitorear una finca en San Gabriel, ubicada en el distrito de Santa Rita de Ciguas en la provincia de Arequipa, con una superficie de más de 35 hectáreas. Utiliza sensores de conductividad, humedad y temperatura para medir las condiciones del suelo y válvulas solenoides de riego para controlar el flujo de agua.

B. (Palacios, 2017), presenta un sistema de monitoreo de cultivos de quinua que utiliza la plataforma de desarrollo de teléfonos móviles 2G/3G de Particle Electron, que se combina con la plataforma Arduino y crea una red en la niebla. Usa sensores para medir el pH, la temperatura ambiente, la humedad interior, la humedad del suelo y el movimiento PIR. La información recopilada por los sensores se almacena y recupera en la plaforma en la nube ThingSpeak. La información del sensor se muestra en el teléfono del usuario a través de la aplicación Pushbullet.

C. (Mora, H., & Rosas, J., 2019), describe un sistema que utiliza redes de Internet de las cosas y sensores inalámbricos para monitorear y controlar el cultivo de frijol canario en una finca en el distrito de Chincha Alta, provincia de Ica. Observe la temperatura y la humedad del ambiente, la humedad del suelo, la lluvia y la concentración de dióxido de carbono.

Por otro lado, (MINAGRI, 2020) monitorea la salinidad, la humedad del suelo, la planificación, la estimación del rendimiento, la detección de plagas, etc. utilizando la plataforma Google Earth Engine (GEE), que brinda una plataforma tecnológica de análisis de imágenes satelitales gracias a sistemas de acceso abierto. de Lansat o Copernicus Todo lo anterior es posible gracias a la amplia gama de imágenes de satélite disponibles.

Capítulo 3: Descripción de la Empresa

En el presente capítulo se describirá brevemente de cómo se inició la industria azucarera e inicios de las empresas azucareras en el Perú. Además de las actividades agrícolas y producción nacional del mercado azucarero.

3.1. Introducción

Se cree que los españoles introdujeron el cultivo de la caña de azúcar en el Perú en los primeros años después de la conquista, y gracias a las condiciones favorables de la costa, el cultivo se extendió rápidamente. En una publicación de 1945 de la Unión de Productores de Azúcar del Perú se menciona que en 1549 ya existían cuatro ingenios para moler tubos y el equipamiento necesario para la producción de azúcar, según HELFGOTT, 1977.

Hasta la reforma agraria de 1969, la industria azucarera peruana constaba de 12 empresas, 8 de las cuales eran grandes haciendas ubicadas en Lambayeque (Cayaltí, Pomalca, Pucalá y Tumán), La Libertad (Casa Grande, Cartavio y Laredo) , y Lima (Paramonga), mientras que las otras cuatro son pequeñas empresas ubicadas en Arequipa (Chúcarapi-Pampa Blanca), Ancash (San Jacinto) y Lima (Andahuasi e Ingenio). Las propiedades Casa Grande, Cartavio, Tumán y Paramonga se destacan como las unidades más importantes.

La mayoría (65%) de las acciones de las fábricas de azúcar están en manos de personas o empresas extranjeras. En Cartavio y Paramonga, casi el 100% de propiedad extranjera está en manos de la empresa norteamericana Grace and Company. En las fincas administradas por la familia Gildemeister, Casa Grande y Laredo, el capital extranjero representó el 73,51% y el 49,67%, respectivamente. En Tuman, el 60,81% de los fondos provino de fuentes extranjeras, principalmente de EE.UU., Suiza, Panamá y Portugal. Solo los complejos Pomalcas y Cayaltí y empresas menores como Andahuasi, Ingenio y Chucarapi son 100% propiedad de peruanos (Ver Tabla 1).

Tabla 1. Propiedades de los accionistas azucareros.

HACIENDAS	Extensión Total (ha)	Extensión con caña (ha)	Producción de Azúcar (t)	Propietarios	Acciones extranjeras (%)
LAMBAYEQUE					
Cayalti	6 565	5 035	63 428	Negociación Aspillaga Anderson Hnos. SA	—
Pomalca	10 107	8 426	83 663	Negociación Viuda de Piedra e Hijos.	—
Pucalá	35 887	8 487	94 824	Negociación Agric. Pucalá	30,64
Tumán	8 541	7 762	92 846	Negociación Tumán	60,81
LA LIBERTAD					
Casa Grande	107 716	15 339	178 346	Emp. Agrícola Chicama	73,51
Cartavio	23 648	11 853	122 654	Cartavio SA (W.R.Grace and Company)	99,99
Laredo	18 989	4 296	42 263	Negoc. Azucarera Laredo	49,67
LIMA					
Andahuasi	1 083	935	9 156	Andahuasi SCRL	—
Ingenio	7 64	385	3 312	Soc. Agric. Santiago Fumagalli Limitada	—
Paramonga	7 942	6 371	57 319	Soc. Paramonga Ltda	99,85
ANCASH					
San Jacinto	—	4 255	23 778	Negociación. Agric. Nepeña	96,21
AREQUIPA					
Chucarapi-Pampa Blanca	1 682	1 326	7 604	Azucarera Chucarapi SCRL Soc. Agrícola Pampa Blanca	_

Fuente: El cultivo de la caña de azúcar en la costa peruana. Salomón Helfgott. Universidad Nacional Agraria en 1997

Durante la implementación de las leyes de reforma agraria, todas las fábricas de azúcar de la costa fueron confiscadas, entregadas a los trabajadores y organizadas en cooperativas. Hasta la década de 1970, Perú era un exportador neto de azúcar, pero la disminución de la producción y la productividad de la caña de azúcar, combinada con el crecimiento de la población de Perú, hizo que Perú tuviera que importar más y más azúcar. Azúcar en los años 80. Para satisfacer una mayor demanda interna que no puede ser satisfecha por la producción nacional. En 1998, el volumen de importaciones fue extraordinario, ascendiendo las importaciones a unas 500.000 toneladas, equivalente a la mitad del consumo del país.

La privatización de las compañías azucareras comenzó en 1996 mediante la ley "Sobre el estado económico y financiero de las empresas agrícolas azucareras", y hasta el momento, la mayoría de ellas han optado por convertirse en sociedades anónimas. En la actualidad, la mayoría de estas compañías están prosperando y están invirtiendo significativamente en la modernización y expansión de sus fábricas, lo que ha permitido que la producción del país se recupere de su caída a uno de los niveles más bajos durante el fenómeno de 1998.

En la Tabla 2 muestra el desarrollo de la producción de caña de azúcar, que es la materia prima para la producción de sacarosa en el Perú. Debido a la ampliación de la frontera

agrícola (área cosechada), este cultivo crecerá sostenidamente en el tiempo hasta 2014 (11,3 millones de toneladas sobre 50). Por otro lado, el desempeño se ha deteriorado a pesar de hacerlo bien en comparación con otros países.

Tabla 2. Producción en el Perú

AÑOS	PRODUCCIÓN (t)	SUPERFICIE COSECHADA (ha)	RENDIMIENTO (kg/ha)
....	....	....	....
1955	6 097 566	35 898	169 858
1956	5 876 384	37 767	155 596
1957	6 077 792	39 353	154 443
1958	6 840 208	39 492	173 205
1959	6 543 824	41 367	158 189
1960	7 359 171	47 361	155 385
1961	7 288 136	47 075	154 820
1962	7 247 077	46 830	154 753
1963	7 697 310	49 160	156 577
1964	7 590 920	48 855	155 377
1965	7 498 940	46 520	161 198
1966	8 463 380	53 530	158 105
1967	7 942 800	49 670	159 911
....	....	....	....
2010	9 660 895	76 983	125 494
2011	9 884 936	80 069	123 455
2012	10 368 866	81 126	127 812
2013	10 992 240	82 205	133 717
2014	11 389 617	90 357	126 051
2015	10 211 856	84 574	120 744
2016 *	9 832 526	87 696	112 120

Fuente: DGESEP-DEA

3.2. Posición competitiva

El azúcar es un producto compuesto por un 82% de caña de azúcar y un 18% de remolacha azucarera. La caña de azúcar se cultiva principalmente en las regiones tropicales y subtropicales del hemisferio sur; la remolacha azucarera se cultiva en las regiones templadas del hemisferio norte. En general, el costo de producción de azúcar es más bajo que el del azúcar de remolacha.

A la fecha, Coazucar cuenta con cinco fábricas: Casa Grande, Cartavio, San Jacinto y Agrolmos en Perú y La Troncal en Ecuador. Cabe señalar que esta última instalación se encuentra en un país con calificación de riesgo internacional B-, mientras que la calificación de riesgo internacional de Perú es BBB+, por lo que se puede considerar que la instalación está operando en un estado precario.

Asimismo, al cierre de 2020, la empresa contaba con un total de aproximadamente 84.287 hectáreas de tierra cultivable con una superficie neta sembrada de caña de azúcar de 58.926 hectáreas, lo que corresponde al 69,9% de la superficie cultivable. La mayor parte del negocio se concentra en Perú, donde se encuentra el 73,6% de la tierra cultivable. La caña de azúcar de terceros representa aproximadamente el 30% del volumen total de caña de azúcar procesada. Una mayor uniformidad de suministro y calidad es garantizada por

la baja dependencia de terceros para el suministro de materias primas, lo que es una ventaja para la empresa.

Incluso en condiciones normales (es decir, sin sequía ni El Niño), la alta producción de la planta se debe en parte al clima favorable de Perú, que convierte al país en líder en términos de producción.

3.3. Derivados del azúcar

El azúcar se puede clasificar por su grado de refinación en:

i. azúcar rubia, contiene entre 96 y 98º de sacarosa
ii. azúcar blanca, presenta 99. 5º de sacarosa
iii. azúcar refinada, con 99.8 y 99. 9º de sacarosa

Además, existen otros derivados de la caña de azúcar, como la melaza, que se usa para producir ron y etanol, y el bagazo, que se usa para producir combustible y papel. El ciclo de producción de la caña de azúcar generalmente dura entre 6 y 8 años. La edad de cosecha o recolección, que oscila entre los doce y los ocho meses, está directamente relacionada con el nivel del 8% al 15% de sacarosa en la caña.

La temperatura, la humedad y la luz son los principales factores que afectan el crecimiento y desarrollo de la caña de azúcar, por lo que el abastecimiento de agua en el campo es muy importante. La caña de azúcar, por su parte, se caracteriza por una alta tolerancia a patógenos y plagas, por lo que puede considerarse un cultivo noble. Como se mencionó, Perú tiene las mejores condiciones agroclimáticas, lo que lo diferencia de otras regiones productoras de caña de azúcar en la región.

Según la Revisión Estadística del Minagri, la producción de caña de azúcar en 2020 será de 10.469 millones de toneladas, un 4,0% menos que en 2019 (10.903 millones de toneladas), principalmente por una disminución en la producción de la tierra. En Perú, los mayores cambios negativos en el rendimiento de la caña de azúcar (en términos de daños a los cultivos) estuvieron más relacionados con la sequía que con la aparición de El Niño.

Sin embargo, según los datos del Minagri, los ingenios azucareros controlan en promedio el 70% del área total cosechada, mientras que los cultivadores independientes controlan el 30%. Según él, principalmente 12 compañías azucareras en el Perú están a cargo de la producción.

De las empresas manufactureras, nueve son mayoritariamente propiedad de grupos importantes, mientras que las otras dos (Pomalca y Tumán) siguen siendo de propiedad de los trabajadores y del Estado, aunque siempre han estado a cargo de empresas privadas.

Según datos de 2020, el mayor productor de azúcar es Casa Grande (23%), seguido de Cartavio (14%), Laredo (11%), Paramonga (11%), Agrolmos (10%) y San Jacinto (9%).). %) %). La producción de estas empresas representa alrededor del 78% de la producción total del país. El procesamiento de azúcar blanca y refinada depende de la tecnología de cada ingenio azucarero y se produce únicamente en Laredo, Cartavio, Andahuasi y Casa Grande, lo que permite a Coazucar atender mejor el mercado industrial.

Capítulo 4: Diagnóstico de la empresa

4.1. La Empresa

4.1.1. Panorama mundial

Según el Departamento de Agricultura de los Estados Unidos (USDA, 2021), se espera que la producción mundial de azúcar aumente un 0,5%, o 181,1 millones de toneladas, durante la temporada 2021/2022. Las preocupaciones climáticas impulsaron la producción en India y Tailandia, mientras que la producción en Brasil disminuyó un 14,4%. Además, con el aumento de los combustibles derivados del petróleo, se reducirá la cantidad de cosechas disponibles para su uso en la producción de etanol.

Por otro lado, el consumo mundial crecerá un 2% hasta los 174,5 millones de toneladas en 2022, impulsado por la mayor demanda de India, la UE, China y EE. UU. En respuesta, se espera que las existencias en 2021/2022 disminuirá a 45,7 millones de toneladas, o 3,1 millones de toneladas menos que en 2020/2021. Tal situación puede conducir a una escasez de existencias de azúcar, que se reducirán más esta temporada en comparación con la temporada anterior.

Las exportaciones mundiales subirán un 0,7% respecto al trimestre anterior. Esto se debió principalmente a una mayor contracción en Brasil de 26 millones de toneladas (-19,1%). Por otro lado, se espera que las importaciones disminuyan un 1,7% en 2021/2022 a 54,2 millones de toneladas.

Los precios internacionales del azúcar para el Contrato N° 5 de Londres (blanco) y el Contrato N° 11 de Nueva York (oro en bruto) han seguido aumentando desde abril de 2020, siguiendo la tendencia de 2021, según la Ilutración 28. Como resultado, ha experimentado una disminución relativa en los primeros dos meses de 2022. Sin embargo, a partir de marzo de 2022, el precio ha vuelto a subir, llegando a $551/ton (contrato No. 5) y $439/ton (contrato No. 11), un aumento del 18 % y el 21%, respectivamente. para el mes en curso de 2021.

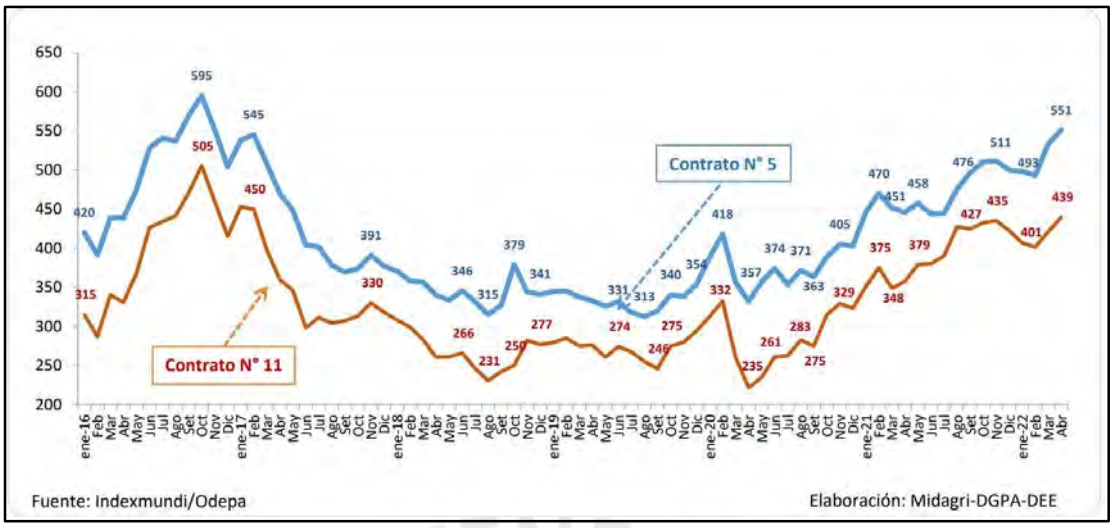


Ilustración 28. Precios Internacionales del azúcar

La oferta mundial de azúcar ha disminuido, principalmente debido a la disminución de la producción en Brasil. Esto se debe a la mayor demanda de azúcar de China e Indonesia, los principales importadores del mundo. El aumento de los precios del etanol contribuyó en gran medida al aumento de los precios internacionales del azúcar, lo que convirtió a Brasil en el principal exportador de azúcar del mundo.

Los precios altos pueden alentar la expansión de la producción de caña de azúcar, pero los costos también aumentan significativamente a medida que aumentan aún más los precios de la energía, como el petróleo y el gas, que alimentan los fertilizantes y el transporte en todo el mundo. Más por el conflicto entre Rusia y Ucrania.

A partir del 15 de abril de 2022, las variaciones diarias de precios internacionales del contrato no. 5 y núm. 11 se ubicaron en USD 576 por tonelada y USD 442 por tonelada respectivamente, mientras que los precios de la azúcar refinada y de la azúcar rubia sin refinar fueron de USD 576 por tonelada y USD 442 por tonelada. respectivamente.

4.1.2. Panorama nacional

Después de alcanzar una producción récord de 10,9 millones de toneladas en 2019 (+5,7% interanual), la producción disminuyó gradualmente durante los dos años siguientes debido al impacto de la pandemia sanitaria. A pesar del pequeño aumento de las áreas cosechadas (0,3%), la producción de caña de azúcar en 2021 ha disminuido un 6,1% a 9,83 millones de toneladas. Sin embargo, una de las razones de la caída de la producción fue un menor nivel de producción (-6,4%): 115,8 toneladas por año.

Tabla 3. Resumen Producción de Caña de azúcar según departamentos.

Departamentos	2010	2015	2 018	2 019	2 020	2021*	Var. 2021/2020	2021 (ene-)*	2022 (ene-)*	Var. 2022/2021 (ene)
Producción (t)										
Nacional	**9 660 895**	**10 211 856**	**10 336 178**	**10 902 906**	**10 468 800**	**9 827 808**	**-6,1%**	**772 440**	**768 940**	**-0,5%**
Lambayeque	2 824 848	2 022 870	2 648 009	2 566 492	2 184 189	2 267 691	3,8%	177 176	165 217	-6,7%
La Libertad	4 911 755	5 529 691	4 795 513	5 514 278	5 344 455	4 705 541	-12,0%	342 202	375 847	9,8%
Ancash	578 284	988 272	870 729	957 461	975 401	910 075	-6,7%	79 719	82 495	3,5%
Lima	1 293 061	1 614 043	1 528 325	1 525 064	1 378 391	1 525 491	10,7%	120 039	120 818	0,6%
Arequipa	52 947	56 980	55 859	64 633	64 801	55 598	-14,2%	5 500	4 500	-18,2%
Superficie cosechada (ha)										
Nacional	**76 983**	**84 574**	**84 838**	**86 473**	**84 590**	**84 852**	**0,3%**	**6 918**	**7 329**	**6,0%**
Lambayeque	26 773	23 430	27 600	26 362	23 382	25 595	9,5%	2 113	1 933	-8,5%
La Libertad	34 235	40 928	35 055	38 717	38 826	38 111	-1,8%	2 718	3 456	27,1%
Ancash	5 174	6 594	6 874	7 101	7 098	6 924	-2,4%	688	760	10,4%
Lima	10 163	12 992	11 707	11 847	10 899	10 949	0,5%	972	1 000	2,9%
Arequipa	638	630	545	605	561	593	5,6%	60	45	-25,0%
Rendimiento (kg/ha)										
Nacional	**125 494**	**120 744**	**121 834**	**126 085**	**123 760**	**115 823**	**-6,4%**	**111 663**	**104 912**	**-6,0%**
Lambayeque	105 511	86 337	95 941	97 356	93 412	88 600	-5,2%	83 838	85 457	1,9%
La Libertad	143 471	135 107	136 801	142 427	137 652	123 470	-10,3%	125 891	108 763	-13,6%
Ancash	111 761	149 874	126 666	134 839	137 424	131 431	-4,4%	115 893	108 611	-6,3%
Lima	127 234	124 236	130 552	128 735	126 472	139 329	10,2%	123 466	120 767	-2,2%
Arequipa	83 005	90 433	102 571	106 785	115 459	93 799	-18,8%	91 667	100 000	9,1%

Fuente: MIDAGRI-DGESEP-DEIA

De la estructura productiva, la producción de tableros de madera representa el 73% de la producción total del país, mientras que la producción de azúcar blanca representa el 27%. Cabe señalar que la producción agroindustrial de azúcar se concentra en empresas de mayor tamaño como Casa Grande, Cartavio (14%), San Jacinto (9%) y Agro Olmos (10%), que representan el 23% de la producción total, todas de los propios Aportado por Grupo Gloria, empresa del grupo Wong Paramonga (12%) y grupo Manuelita (de Colombia) Laredo (11%). El grupo de empresas incluye a Pomalca (5,5%) y Tumán (4,5%). Todas estas empresas concentran el 89% de la producción nacional.

La producción en 2020 alcanzará los 1,2 millones de toneladas, lo que significa que la producción será suficiente para cubrir el consumo. Sin embargo, la producción de azúcar en 2021 será de solo 1,1 millones de toneladas, un 8,2% menos que en 2020.

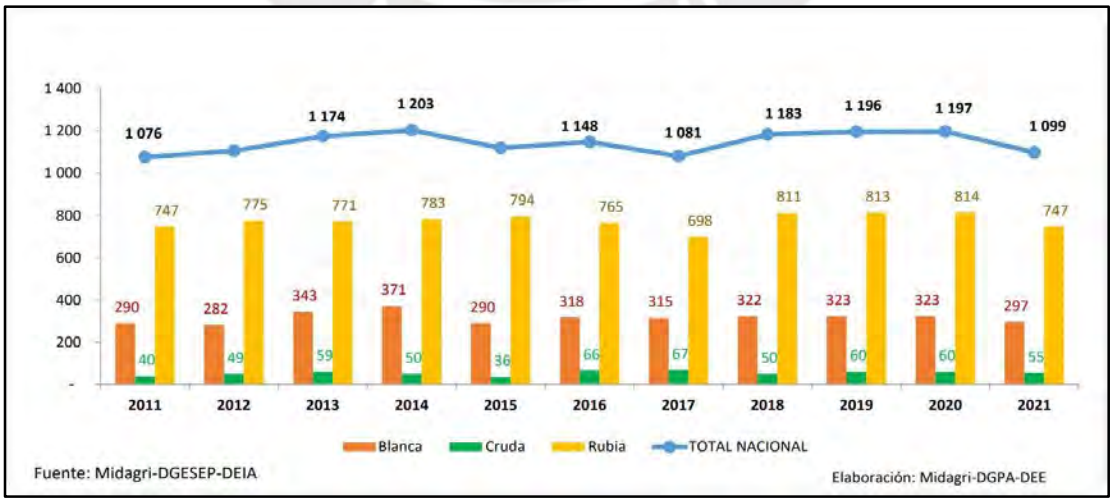


Ilustración 29. Producción por tipo de azúcar.

La pandemia de COVID-19 ha reducido la actividad en la industria azucarera al limitar la mano de obra para la cosecha de caña de azúcar y el funcionamiento de las operaciones industriales a capacidad completa. Sin embargo, los precios internos aumentan como resultado de una producción interna más baja y precios internacionales más altos. En ese contexto, se explicó que la principal industria azucarera demostró mayores ingresos por ventas a pesar de la caída en la producción de azúcar de caña este año.

De acuerdo con el Sistema de Información de Abastecimiento y Precios del Ministerio de Desarrollo Agropecuario y Riego (Midagri), los precios mayoristas de azúcar blanca y panela que se comercializan en el mercado productor de Lima (Santa Anita) vienen aumentando desde 2020 y seguirán aumentando; es decir, hasta 2021. Tampoco habrá escasez, ya que los ingresos del azúcar en marzo de 2022 son un 23% superiores a los del año anterior.

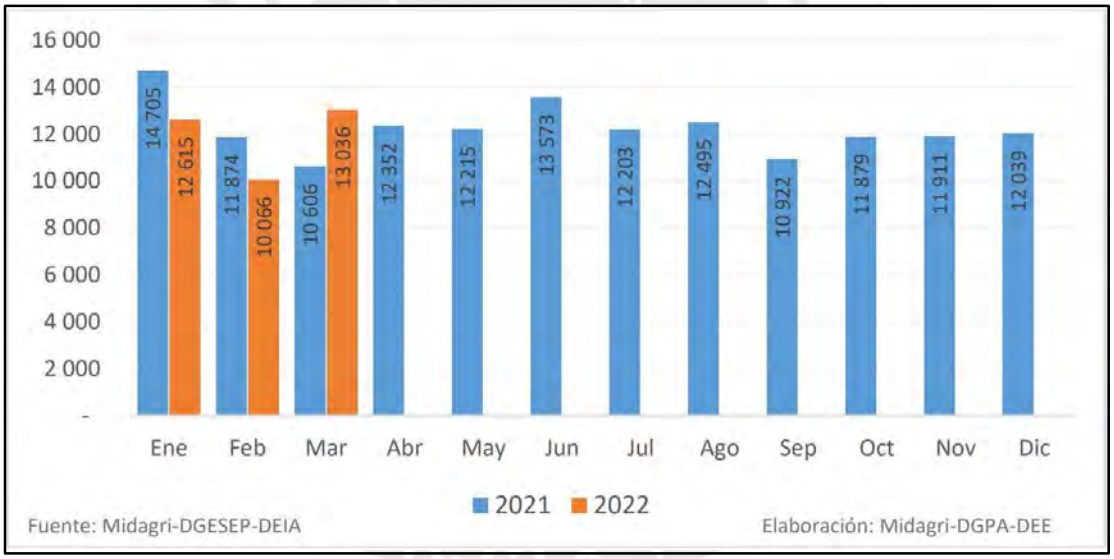


Ilustración 30. Ingreso Mensual de Azúcar de Comercial en Lima

El aumento especulativo de los precios del azúcar se vio alimentado además por factores como el conflicto entre Rusia y Ucrania. Los mercados mayoristas reflejan precios más altos de azúcar blanca y clara a USD 3,87 y USD 3,81 por kg respectivamente en abril de 2022; un aumento promedio del 72% en comparación con el mismo mes de 2021. Tal que, los precios internos aumentan directamente por el aumento de los precios internacionales.

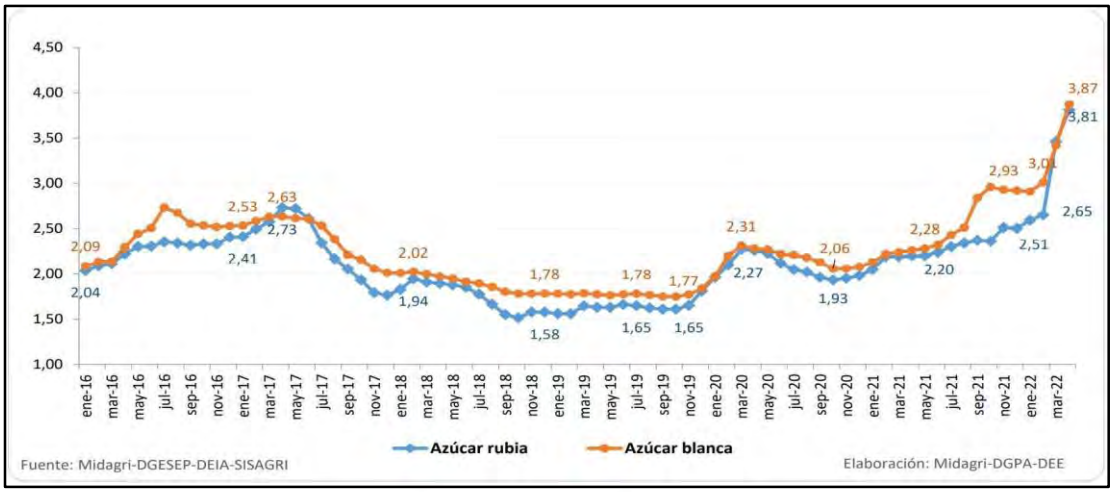


Ilustración 31. Evolución Mensual de los precios mayoristas del azúcar en Perú.

Las importaciones han sido bajas en los últimos años. En 2021, una importante disminución de 44,3%, con registros de 194.400 toneladas por un valor de $90,4 millones, se originó en Colombia (52,4%), seguida de Bolivia (19%), Guatemala (12,2%) y Brasil (11,1%). Esto significa que las necesidades de los consumidores nacionales ya están cubiertas por la producción nacional. En los tres primeros meses de 2022, las importaciones seguirán cayendo un 46,7%.

En cuanto a las exportaciones, en 2021 el volumen de exportación será de 85.300 toneladas y el valor de exportación será de 55 millones de dólares estadounidenses, una disminución del 30 % y el 14 %, respectivamente. EE.UU. fue el principal destino de las exportaciones (60% del total), registrando 51.100 toneladas. También continúa reportando una caída del 77% en las ventas en el primer trimestre de 2022.

La producción interna disminuye y los precios de las materias primas aumentan en el mercado interno. Las compañías azucareras dan prioridad a la comercialización en los mercados internos en lugar de los mercados externos.

4.1.3. Indicadores de desempeño Operativo

Coazucar posee fábricas en Casa Grande, Cartavio, Agrolmos y San Jacinto en Perú y La Troncal en Ecuador. En 2018 no estuvo disponible en Argentina con San Isidro. Así, la empresa cotizaba 84.287 hectáreas de tierra cultivable a finales de 2020 (correspondientes a datos de diciembre de 2019) y tenía una tasa de utilización del 69,9% (58.926 hectáreas de tierra cultivable). Asimismo, la empresa mantuvo su capacidad de procesamiento de caña de azúcar de 38.700.

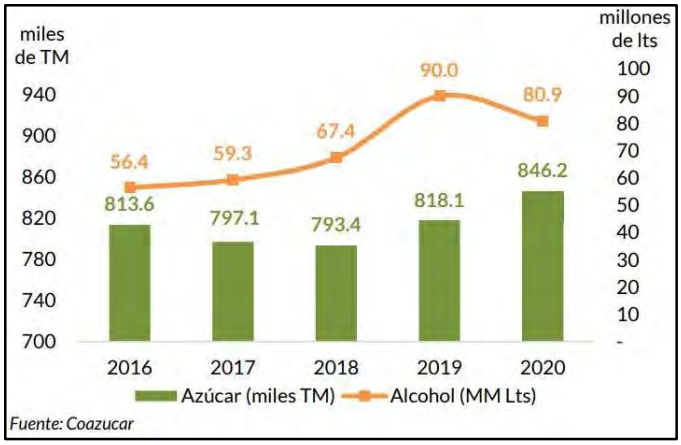
Tabla 4. Indicadores de Desempeño Operativo

Desempeño histórico	2018	2019	2020
Ha. Cultivables Totales	83,849	84,287	84,287
Ha. Cultivadas Neta	61,785	58,640	58,926
Ha. Cosechadas con caña	44,727	47,797	45,262
Capacidad Molienda (TM/día)	38,700	38,700	38,700
Rendimiento (TM de Caña / Ha)	118	120	127
% de Azúcar / TM de Caña	10.2%	10.1%	10.2%
Azúcar Producida (TM)	769,562	818,118	814,543
Producido de Importaciones (TM)	23,875	-	31,686
Total azúcar producido	793,437	818,118	846,229
Alcohol Producido (miles de Lt.)	67,440	90,023	80,941
Número de trabajadores	12,813	10,862	11,812
Precio Promedio (PEN x kg.)	1.56	1.49	1.77
Ventas (S/MM)	1,313	1,592	1,736

Fuente: COAZUCAR

En cuanto a la producción rural, la buena tendencia continúa luego de la severa sequía de 2016 y el fenómeno de El Niño costero (inundaciones y difícil acceso a la infraestructura vial) a principios de 2017, lo que indica una baja producción interna y un mayor número de importadores en el mercado interno. posibilidad de existencia del mercado. Como resultado, el rendimiento se fijó en 127 t/ha (120 t/ha en 2019 y 118 t/ha en 2018).

Asimismo, en 2020, Coazucar cosechó 45.262 toneladas y registró una producción de azúcar comercial de 814.500 toneladas (sin importar importaciones), un 0,4% menos que la producción de 2019 (818.100 toneladas). Si a la producción de azúcar se incluye la producción importada de 31.700 toneladas, la producción total es de 846.200 toneladas, lo que supone un aumento del 3,4% respecto a la producción registrada en 2019. En cuanto a la producción de etanol, alcanzó los 80.941 millones de litros en 2020, abastecidos por las plantas Casa Grande, Cartavio y San Jacinto en Perú y la planta Producargo en Ecuador.



Fuente: Coazucar

Ilustración 32. Producción de Azúcar y Alcohol

Del total de ingresos por ventas de azúcar en 2020, el azúcar moreno representó el 61,2 %, el azúcar blanco el 20,9 % y el azúcar refinado el 17,9 %. Cabe destacar que, gracias al inicio de operaciones de Casa Grande desde 2015, se logró una mayor participación de los ingresos de azúcar refinada. La refinería Casa Grande tiene una capacidad instalada de 11.000 toneladas por día.

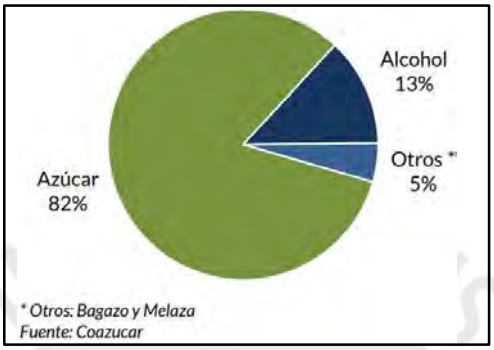


Ilustración 33. Ingresos por La línea de Negocio 2020

4.2. Situación Actual

4.2.1. Producción Nacional

En comparación con otros países productores de caña de azúcar en la región, el Perú cuenta con ventajas y altos rendimientos debido a su ubicación geográfica y condiciones agroclimáticas favorables. En 2020, nuestra producción nacional superará la de Brasil y Colombia en 124 toneladas anuales.

Actualmente representa el 3,6% de nuestro Producto Interno Bruto (PIB) en el sector agroindustrial y genera alrededor de 500.000 empleos directos e indirectos. Debido al clima, esta actividad se da principalmente en las zonas costeras, pero también se da en las montañas y selvas.

En la actualidad, el Perú posee una superficie cañera de más de 160.000 hectáreas, principalmente en las provincias de Piura, Lambayeque, La Libertad, Ancash, Lima y Arequipa. Los ingenios azucareros son los principales desarrolladores de la actividad. Casi diez de ellos son responsables del 65% de la producción nacional, mientras que el 35% restante es propiedad de pequeños productores.

La producción del azúcar en el país se concentra en las principales empresas del Grupo Gloria (Casa Grande, Cartavio San Jacinto y Agro Olmos), además de Paramonga, Laredo, Pomalca y Tumán, que en conjunto suman el 89% de la producción.

La producción de azúcar del país se acerca mucho a la demanda. Sin embargo, ha habido cierta especulación en torno al reciente aumento de los precios comerciales. Se observó que en el primer bimestre de enero-febrero el mercado mayorista de Lima registró menores ingresos por azúcar, pero en marzo el azúcar importado aumentó un 68% y se vendió a precios más altos. Aparentemente, las compañías azucareras que administran existencias con información distorsionada sobre la escasez de azúcar han limitado la oferta del mercado tardío y han elevado los precios especulativamente. El mayorista informó que una bolsa de azúcar de 50 kg se vendía a unos 130 dólares singapurenses la bolsa hasta principios de febrero, cuando salió a la venta en la primera semana de principios de febrero.

Los precios más altos de los combustibles, por otro lado, ejercen presión principalmente sobre los precios de los alimentos, mientras que la suspensión del transporte solo aumentará la especulación. Sin embargo, es difícil predecir cuánto durará la subida del precio del azúcar, así que depende. La oferta interna dependerá de la producción nacional y no de las importaciones, ya que los altos precios internacionales reducen las importaciones y las compensan con edulcorantes de producción nacional, así como de la evolución de los precios internacionales del petróleo, que están influenciados por muchos factores externos.

4.2.2. Desarrollo de las Operaciones en Azúcar Rica

La superficie total de la empresa al cierre del año era de 31.197 hectáreas, de las cuales 17.938 hectáreas estaban sembradas con caña de azúcar, un 5% menos que en 2019.

Molienda Anual

En 2020 molturaremos un 1,56% más de caña de azúcar que en 2019: se cosecharon 2.073.412 toneladas de caña de azúcar, de las cuales el 94,95% se procesó en nuestros ingenios. La superficie cosechada fue de 12.660 hectáreas, un 3,33% menos que en 2019.

Rendimiento de campos propios

Sobre la base del rendimiento, la edad de la cosecha de 2020 aumentó un 2,7 %, mientras que las toneladas por hectárea mes (TCHM) aumentaron un 2,3 %; aumentó un 5,1%. En comparación con el año anterior, la sacarosa aumentó un 2,9%.

Tabla 5. Rendimiento de campos propios

	2020	2019	Var. %
TCH	163.78	155.9	5.10%
TCHM	9.1	8.89	2.30%
Edad	18	17.53	2.70%
% Sacarosa	13.41	13.02	2.90%

Elaboración Propia

Recursos Hídricos

El caudal acumulado del río en 2020 en comparación con 2019 disminuirá en un 56%. Analizando el tráfico mensual, el mayor tráfico se concentra en los primeros 5 meses, como se puede observar en la Tabla 6:

Tabla 6. Recursos Hídricos

	2020	2019
Enero - Mayo	288.2	672.3
Junio - Diciembre	72.6	143.5
Acumulado	360.8	815.8

Elaboración Propia

La disponibilidad de agua alcanzó el 57 por ciento del plan de riego acumulado en comparación con el 86 por ciento en 2019. La tasa de cumplimiento del agua en los primeros cinco meses de 2020 fue del 71 % en comparación con el 97 % en 2019; sin embargo, de junio a diciembre de 2019, el plan de riego acumulado se cumplió en un 41% en comparación con el 75% durante el mismo período de 2019.

En el Año 2024, se realizó en Casa Grande S.A.A., el mantenimiento de la Infraestructura de Riego Comunal (Canales) con una longitud de 123 Km. y un cumplimiento del 75%; de los cuales el 44% fue manual y el 56% mecanizado

Fertilización

En el año 2024 se fertilizaron 11,608 ha, 0.08% menos que el año anterior (12,595 ha en el 2023) esto debido a una menor área cosechada (-20.36%) y menor área sembrada (-23.1%) con respecto al 2023.

Control de Malezas

En el año 2024 se aplicó 35,989 ha, 19% más que el año anterior (29,151 ha en el 2023), esto debido a que se mecanizo la labor de redondeos químicos bajando la cobertura de

malezas en acequias y girones de campo. Se tuvo mayor cobertura de aplicación en pre-emergente por moléculas que trabajan en húmedo facilitando la aplicación con drone.

Sanidad vegetal

En el año 2024 la plantación de caña ha tenido un % de Intensidad de Infestación por Diatraea de 2.16% menor al año 2023 de 2.34 %. Por el control biológico que se aplicó en campos susceptibles, además que el 86% de campos cosechados fueron socas, los cuales presentaron menor afectación por Diatraea saccharalis.

Preparación y Siembra

En el año 2024 se preparó 2,546 ha, siendo 4.2% menos que el año anterior (2,658 ha, en el 2023). En el año 2024 se sembró 1,729 ha, teniendo 23.1% menos que el año anterior (2,249 ha, en el 2023), esto debido al déficit hídrico presentado el año 2024.

4.2.3. Análisis del Problema

Este estudio se justifica porque conocer el correcto proceso y técnica de aplicación del cultivo de la caña de azúcar en la región liberteña permitirá adquirir prácticas agrícolas encaminadas a alcanzar mayores niveles de productividad y así mejorar la economía, no ignore la experiencia de los agricultores.

Las empresas azucareras del norte del país son las más grandes agroindustrias y utilizan entre 10 a 20 toneladas de caña de azúcar por día para su producción. Los problemas que afectan a estas empresas incluyen problemas con el monitoreo oportuno del crecimiento de los cultivos, escasez de nutrientes como nitrógeno, falta de agua o áreas afectadas por malezas y plagas.

El propósito de este trabajo es identificar las variables con mayor impacto en la productividad de los cultivos de la caña de azúcar, utilizando un sistema de recolección de datos y un modelo numérico. Se prueba con temperatura, humedad, índices de productividad y variables topográficas de la zona para diagnosticar mejor los cultivos.

Resolución del Problema

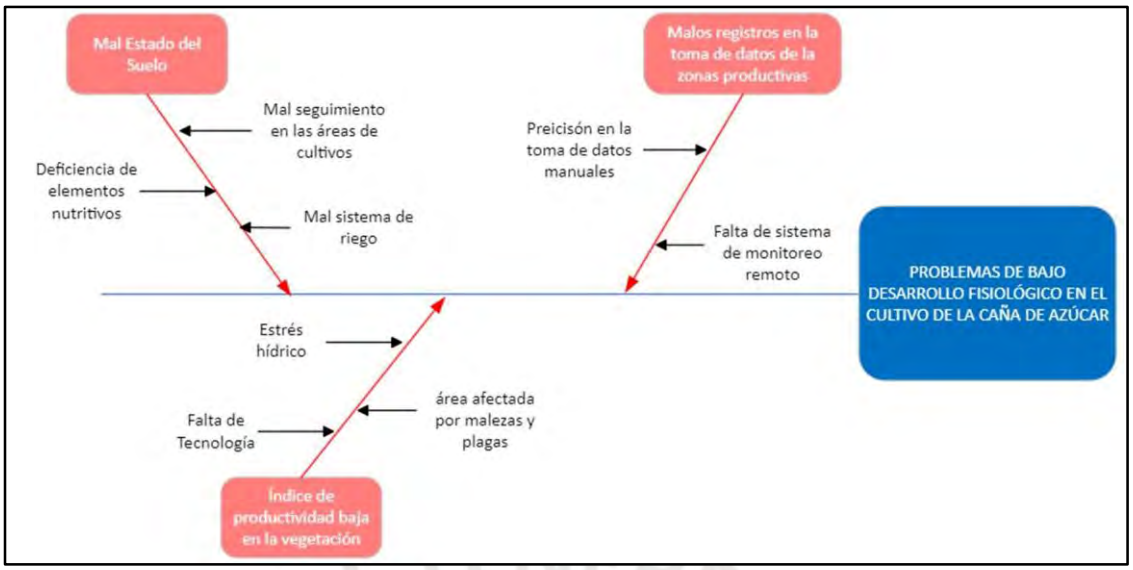


Ilustración 34. Diagrama de Ishikawa

Capítulo 5: Propuesta de Mejora

5.1. Diseño de la plataforma de Big Data

Para el diseño se utiliza un sistema de adquisición y arquitectura de datos. Primero se recolecta las señales de campo y fuente externas como sensores, drones e información de imágenes satelitales que permita tener datos del clima, estado del suelo y luego utilizar una solución y análisis de Big Data para tomar mejores decisiones sobre sus campos.

5.1.1. Adquisición de datos

A través de un sistema de adquisición se recolecta datos de sensores de humedad y temperatura para registrar el estado del suelo. Estos tienen tecnología de comunicación industrial y alta precisión de medición, y van conectados a un Gateway que concentra a todos los sensores distribuidos en el área enviando a tiempo real los datos y sincroniza la información al servidor de base de datos.

La humedad y temperatura son dos factores importantes en la agricultura, ya que puede ayudar a tomar decisiones par aún mejor hidratación de los cultivos y uso de fertilizantes.

En la agricultura de precisión, donde los UAV o drones son controlados de forma remota por un operador o autónoma a través de un piloto automático y pueden transportar equipos avanzados como cámaras, los agricultores se sienten cómodos con la interfaz de usuario de la plataforma. Además, la tecnología permite modelos digitales para estimar curvas de nivel, construir mapas de pendientes, modelos de superficie para medir la altura de las plantaciones de caña de azúcar.

La captura de imágenes satelitales nos permite conocer la extensión del área de producción y obtener las coordenadas.

En la Ilustración 35 muestra un diseño de adquisición el uso de diferentes tecnologías que registran datos en tiempo real. Toda la integración de datos vertical y horizontal hace un sistema flexible y fácil acceso.

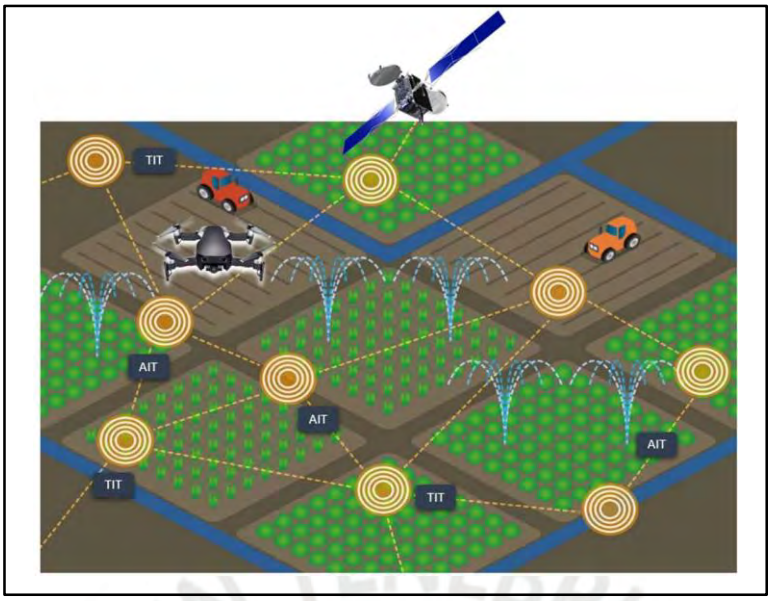


Ilustración 35. Diseño de Adquisición de datos

5.1.2. Arquitectura de datos

Para el diseño de la arquitectura de datos se utiliza OLTP y OLAP.

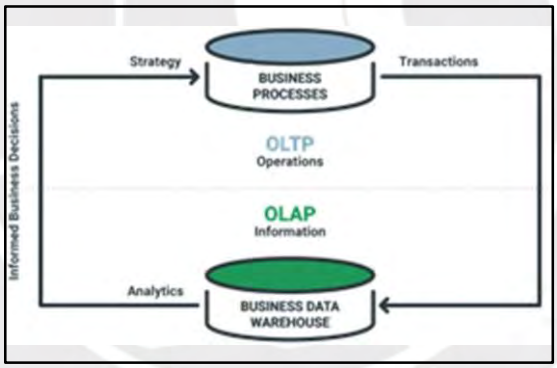


Ilustración 36. Diagrama de flujo del uso del OLTP y OLAP

El OLAP es el análisis de datos y OLTP es el procesamiento de datos. Puede realizar tareas frecuentes de lectura, escritura y acceder a datos optimizados con OLTP. Los formatos de los datos varían según la aplicación y la estructura. La mayoría de las veces, los datos históricos se limitan a datos actuales o recientes.

Con OLAP se tiene un sistema multiusuario, para que los usuarios tengan acceso rápido e interacción con los datos, para que estos se puedan agrupar, segmentar y desglosar según el interés. El sistema de recopilación proporciona una base de datos OLAP y se extrae, transforma y carga.

Para la arquitectura de base de datos extendida se utiliza Warehouse. Los datos en Warehouse están en esquema de formato estructurado que son procesados, formateados y diseñados para el rendimiento. Es un sistema no transaccional que está optimado para la lectura en un formato orientado a las columnas, a menudo en una amplia gama de filas. Se consulta a través de SQL.

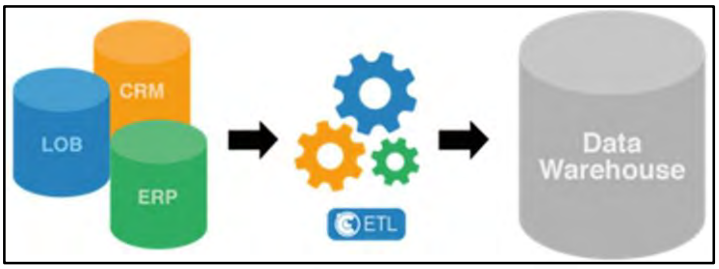


Ilustración 37. Arquitectura Warehouse

Para el sistema de ficheros utilizaremos Hadoop para almacenar grandes archivos de datos ya sea Megabytes, Gigabytes o Terabytes.

Dichos sistemas generalmente funcionan generando o replicando un conjunto de información de una fuente y luego realizando varios procesos analíticos en esos datos. Por lo tanto, el sistema pone más énfasis en leer grandes fragmentos de datos contiguos antes de escribir el primero.

Hadoop no requiere máquinas grandes a nivel de componentes, está diseñado para tolerar fallas altas en las máquinas. En el caso de tales fallas, HDFS puede continuar sin interrupciones percibidas por el usuario.

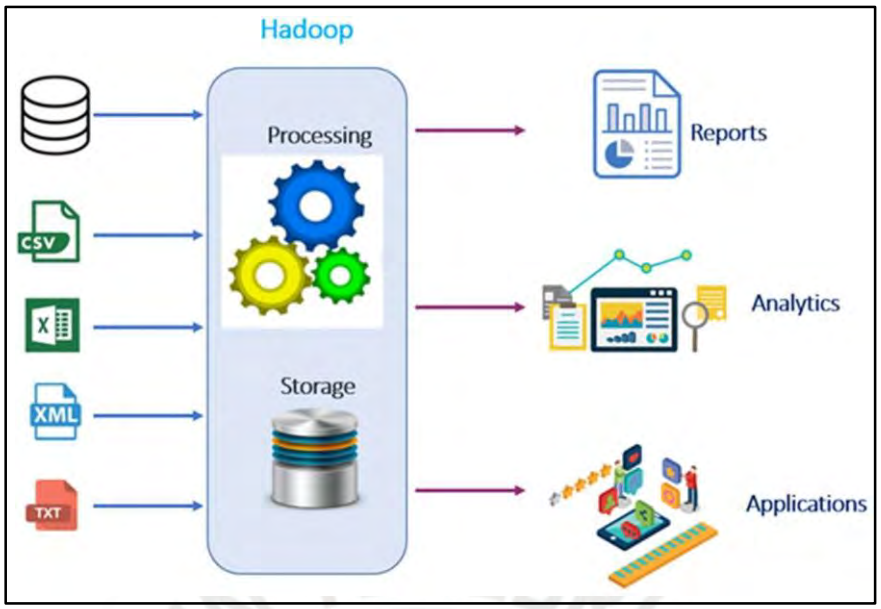


Ilustración 38. Arquitectura Hadoop

5.1.3. Modelo del Sistema

Se muestra el siguiente esquema del modelo a utilizar para la implementación de la arquitectura del Big Data.

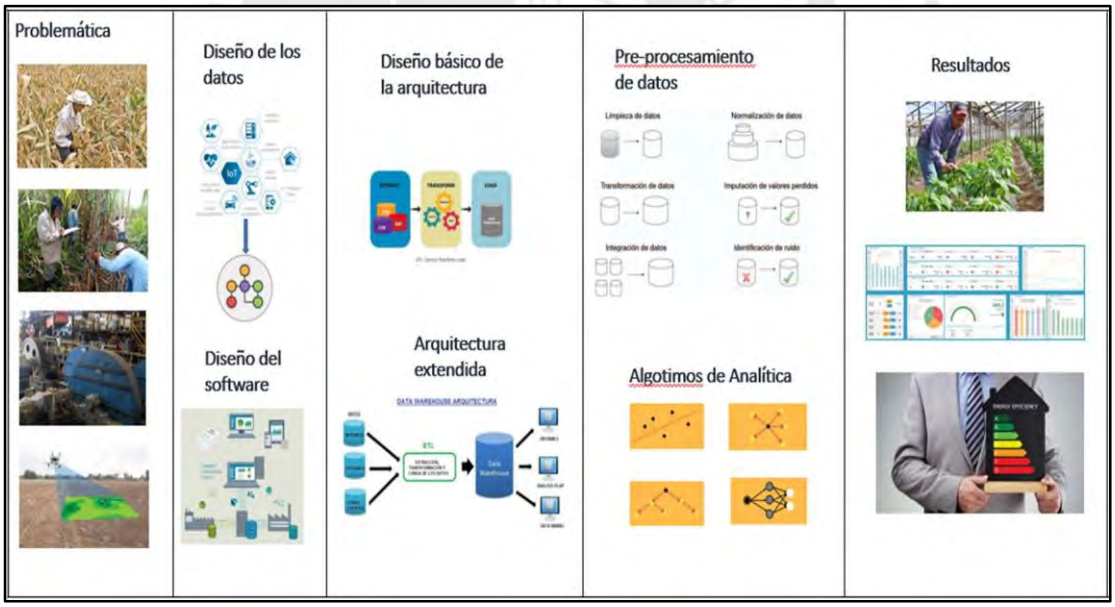


Ilustración 39. Modelo de diseño de solución

Se empieza analizando la problemática, observando las áreas de cultivos, el trabajo en el campo y el proceso del cultivo. Además, se revisan los reportes e informes del estado del suelo, variables biofísicas e índices de producción.

Luego cree una base de datos utilizando el gráfico de variables registradas. El software del sistema operativo recupera estas variables del sistema de minería de datos y las almacena en la computadora.

Para la arquitectura de Big Data, Hadoop se usa en un sistema de almacén para almacenar una gran cantidad de información. Utilice algoritmos predictivos y descriptivos para procesar datos para ayudar a analizar variables más fácilmente y tomar decisiones más rápido.

Los resultados que se obtienen son indicadores, tablas y dashboard que ayudan a tener una mejor presentación de KPI y curvas de tendencias que explican el comportamiento del estado de agricultura.

5.1.2. Desarrollo del Modelo

Para desarrolla el modelo se utilizan las variables biofísicas, características del suelo, índice de vegetación, imágenes fotográficas y satelitales.

Variables de entrada:

1. Características del suelo:
    - Temperatura
    - Humedad
2. Variables biofísicas:
    - Índice de área a foliar
    - Altura de tallos
    - Número de hojas verdes
3. Imágenes capturadas.
    - Detección de plagas y enfermedades
    - Detección de malezas.
4. Imágenes satelitales.
    - Mapa y área de cultivo.

Índices:

1. Variables biofísicas:
    - Índice de área foliar
2. Índice de vegetación:
    - CIG (Índice de clorofila verde)

- NDVI (Índice de vegetación de diferencia normalizada)

Todos estos datos adquiridos son pre-procesados, procesados y almacenados en la base de datos Warehouse. Para ello se realiza la limpieza y normalización de los datos obtenidos.

Una vez teniendo los datos más adecuado y comprensible para el algoritmo, se aplica métodos y modelos que ayudan a analizar mejor la problemática que afectan a los cultivos de la caña de azúcar. Se utilizan los siguientes modelos:

- Modelos de regresión para la estimación del rendimiento de caña de azúcar con los datos de variables biofísicas e índices de vegetación.
- Modelo de clustering para determinar que zonas de los cultivos tienen zonas afectadas y mejor productividad.

Para este diseño de arquitectura de big data se propone utilizar el lenguaje de programación R que permite:

- Manejo y almacenamiento de los datos.
- Crear visualizaciones de datos.
- Crear dashboards para visualizar y analizar datos.

Ilustración 40. Lenguaje R.

Después del modelamiento se realiza el análisis para discutir los resultados. Se realiza los siguientes análisis:

- Análisis estadístico para las variables biofísicas.
- Caracterización espectral de la caña de azúcar.
- Cálculo de índice de vegetación con respuesta a la fertilización.
- Analizar el nivel de productividad de cada zona de los cultivos.

Finalmente se toman acciones y decisiones para mejorar la productividad en el cultivo de la caña de azúcar como:

- Manejo del suelo

- Predicción del rendimiento
- Gestión de calidad del cultivo
- Manejo de riego
- Fertilización
- Control de malezas

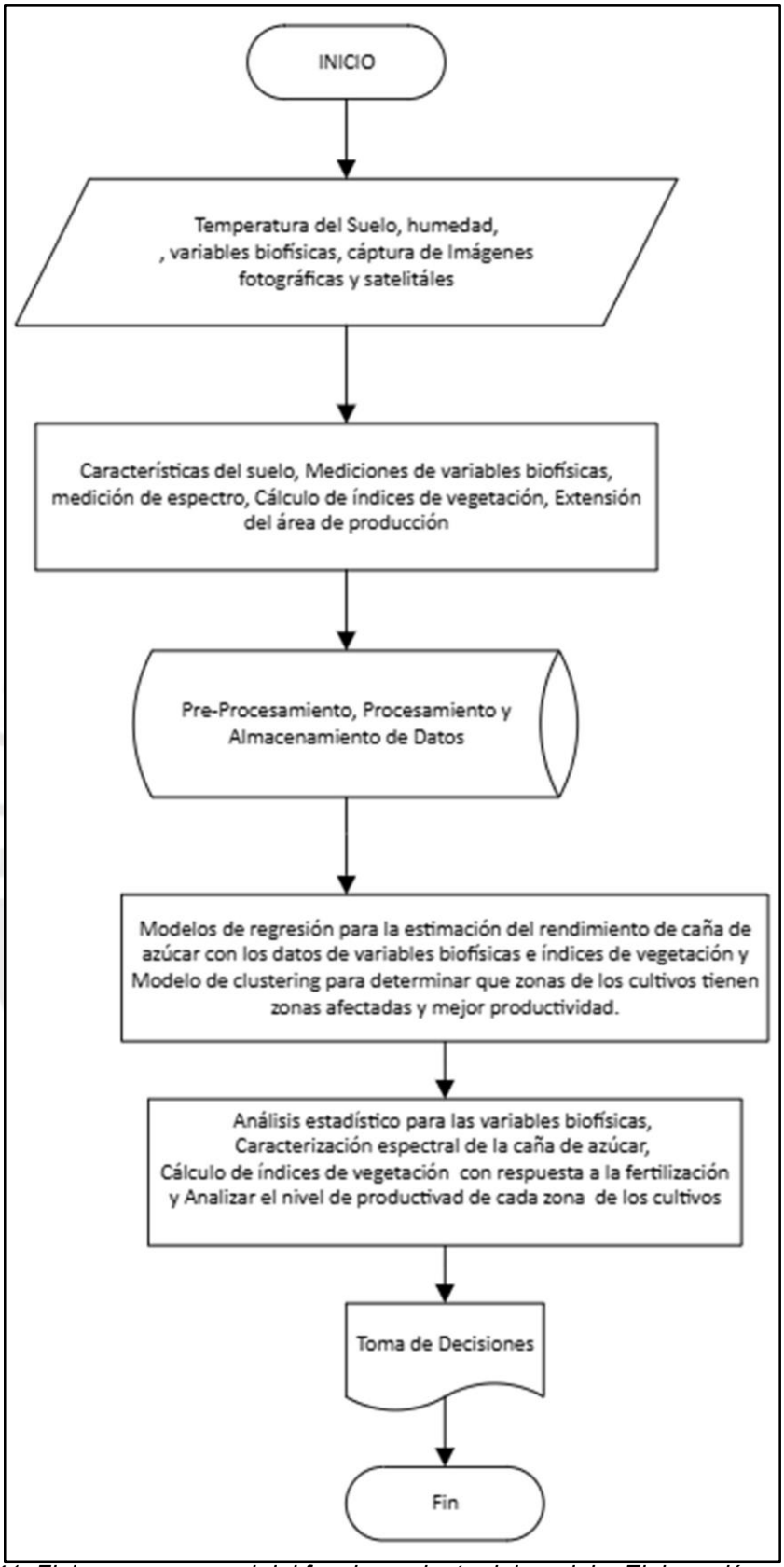


Ilustración 41. Flujograma general del funcionamiento del modelo. Elaboración Propia.

5.1.4. Generación de mapas de cultivo

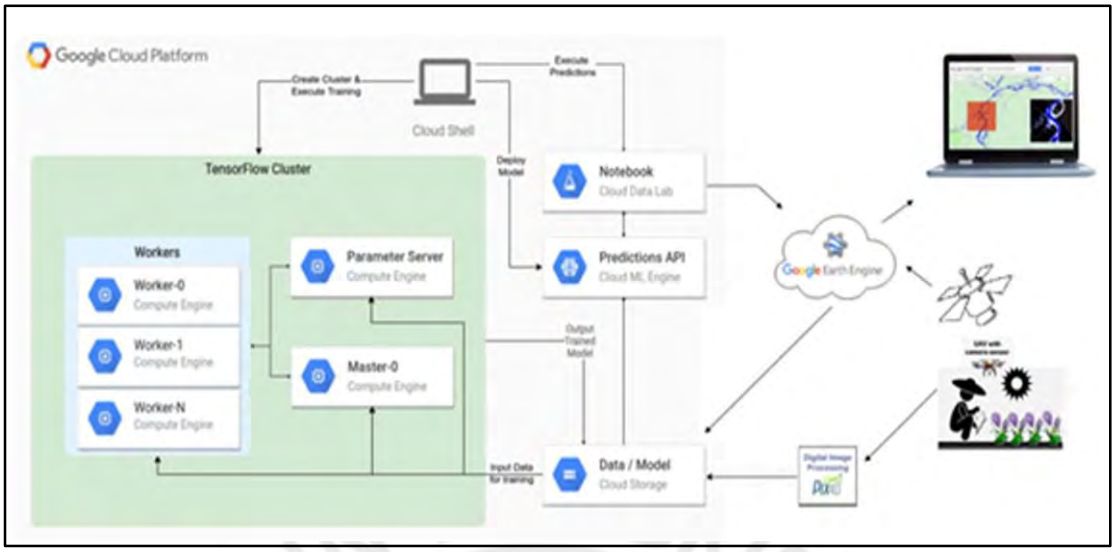


Ilustración 42. Arquitectura del Subsistema de Generación de mapas de cultivo

En la Ilustración 42 se presenta la arquitectura del subsistema de generación de mapas de cultivo. Las imágenes RGB, multiespectrales y térmicas del campo de cultivo obtenidas por vehículos aéreos no tripulados son procesadas generando ortomosaicos utilizando el software Pix4D Mapper y almacenadas en Cloud Storage de GCP (Google Cloud Platform). Las imágenes satelitales son adquiridas y procesadas en GEE (Google Earth Enginee) brindando acceso a datos científicos, que se actualizan y amplían diariamente. Mediante la API de GEE, disponible en Pythony y JavaScript, las imágenes son almacenadas temporalmente en Cloud Storage para luego ser procesadas y se genere modelos de predicción de patrones en el campo de cultivo. El entrenamiento de modelos predictivos basados en aprendizaje profundo requiere abundante poder computacional de hardware especializado como GPU (Graphics Process Unit), para esto una configuración distribuida de la plataforma TensorFlow ejecutándose en un clúster de máquinas virtuales de Compute Engine es necesario. La plataforma de IA de Google permite realizar el entrenamiento mediante código ejecutado por Cloud DataLab y la utilización de modelos entrenados mediante el API de predicciones de Cloud ML Engine. La imagen de los resultados se realiza en GEE para algún procesamiento adicional por expertos o su consumo como servicio en plataforma web o móvil.

5.1.5. Análisis Económico

Azúcar Rica al cierre del 2024 cuenta con 31,197 ha brutas, de las cuales 16655 ha se encuentran con cultivo de azúcar, teniendo una disminución del 8% respecto del 2023.

Tuvo una producción de molienda de caña de 14’25916 TM. La superficie cosechada alcanzó las 11,515 ha.

La edad de cosecha tuvo un crecimiento en un 18.86%. El Indicador de Caña por Hectáreas Mes (TCHM) tuvo un decrecimiento en 7.75%; ambos factores influyeron en el decrecimiento del aporte de toneladas de caña por hectárea (TCH) en 6.03% respecto al año 2023. La sacarosa tuvo un crecimiento en 6.02% respecto al año anterior.
Las toneladas de Caña Molida Diaria (TCMD) representa el 7.2% de toneladas de caña por hectárea

Tabla 7. Indicador de las Operaciones de Azúcar Rica al término del 2024.

INDICADOR	2023	2024	VARIACIÓN
TCH	131.78	123.83	-6.03%
TCHM	7.42	6.85	-7.75%
EDAD	17.76	18.09	1.86%
TCMD	9.49	8.92	-5.10%
TCA	0.95	0.89	-0.05%
% SAC	12.05	12.78	6.02%

Fuente. (Jonny Rocio Aquize Diaz, 2024)

En el 2023, la sanidad vegetal, la plantación de caña ha tenido una disminución de Intensidad de Infestación por Diatraea de 2.34%. Se fertilizaron 12,595 ha, 7.76% más que el año pasado, esto debido a una mayor área cosechada. Y se aplicó 29,151 ha de control de malezas, 8.54% menos que el año anterior, esto debido al incremento de área con variedades precoces y menos área para la aplicación de pre emergente

Con la propuesta del Diseño de una Arquitectura de Big Data se puede gestionar los datos de mediciones de variables biofísicas, índice de vegetación y extensión del área de producción para lograr mejorar los indicadores al cierre de cada año. Con esta propuesta se puede evaluar los resultados y aumentar el área cosechada a un estimado de 30 % anual, molienda de caña 15% y producción de azúcar de 5%.

La cadena de abastecimiento de caña de azúcar según sus operaciones de corte, alce y transporte de 10,000 toneladas de caña al día en una cosecha de 138.89 TCH. Permitiendo cubrir de forma sostenible y eficiente la demanda la fábrica. Así como los costos de operación, sistemas hídricos, fertilizantes, ingenieros de campo, personal de campo, transportes, maquinarias, combustible y abonos para el cultivo.

Para la evaluación económica se obtuvieron los datos comparativos de la molienda y producción de azúcar.

Tabla 8. Comparativo de la Molienda y Producción de Azúcar.

Molienda	UM	2024	2023	Variación (%)
Caña Total	TM	1,750,756	2,562,049	-31.67%
Azúcar Rubia	TM	71,803	163,795	-15.97%
Azúcar Refinada	TM	119,817	163,795	-26.85%
Azúcar Blanca	TM	0	82.45	-100.00%
AZÚCAR TOTAL (MP Caña + MP Crudo)	TM	191,620	249,330	-23.15%
Rdto. Comercial (MP Caña)	%	10.13%	9.36%	8.23%

Fuente. (Jonny Rocio Aquize Diaz, 2024)

Consideremos la producción del azúcar total de 191,620 TM y se tiene rendimiento comercial del 10.13% siendo la venta de 19,411.11 TM y mostramos algunos valores para el cálculo de rentabilidad.

- Costo de producción por bolsa de azúcar de 50 Kg a S/ 65.00
- Precio de venta bolsa de azúcar de 50Kg a S/ 85.00
- 20 bolsas de azúcar por tonelada

La rentabilidad del 2024:

$$Rentabilidad = Producción\ de\ azúcar * \frac{Bolsas\ de\ azúcar}{1\ Tn\ azúcar} * \frac{Utilidad\ Operativa}{1\ bolsa\ de\ azúcar}$$

$$Rentabilidad = 19{,}411.11 * \frac{20\ bolsas\ de\ azúcar}{1\ Tn\ azúcar} * \frac{85 - 65}{1\ bolsa\ de\ azúcar}$$

Rentabilidad = S/ 7,764,444.00

La rentabilidad 2023 total de azúcar en rendimiento comercial del 9.36% siendo la venta de

$$Rentabilidad = 23{,}337.29 * \frac{20\ bolsas\ de\ azúcar}{1\ Tn\ azúcar} * \frac{85 - 65}{1\ bolsa\ de\ azúcar}$$

Rentabilidad = S/ 9,334,916.00

Rentabilidad (2024 – 2023) = Rentabilidad (2024) – Rentabilidad (2023)

Rentabilidad (2024 – 2023) = S/ 7,764,444.00 – S/ 9,334,916.00

Rentabilidad (2024 – 2023) = - S/ 1,570,472.00

Comparando entre el 2024 y 2023 hay una pérdida económica de S/ 1,570,472.00 Si se logra completar la demanda requerida de la fábrica de 10,000 TM, se logra producir el 10% de azúcar de 1,000 TM. Lo óptimo sería 360,000 TM. Como se tiene 1 mes de mantenimiento de la fábrica y mejoras (parada de planta), 1 mes de pre-arranque y arranque toda la producción se tiene una producción bruta del 70% que es 270,000 TM del Azúcar Total. Considerando el 10% de Rendimiento Comercial se obtiene 27,000.00 TM en venta.

Realizando el mismo cálculo de rentabilidad se tiene:

$$Rentabilidad\ Óptima = 27{,}000.00 * \frac{20\ bolsas\ de\ azúcar}{1\ Tn\ azúcar} * \frac{85 - 65}{1\ bolsa\ de\ azúcar}$$

Rentabilidad Óptima = S/ 10,800,000.00

Se puede observar que hay una mejora y ganancia óptima comparando el último año de producción:

Rentabilidad (Óptima – 2024) = S/ 10,800,000.00 – S/ 7,764,444.00

Rentabilidad (Óptima – 2024) = S/ 3,035,556.00

Costo de la implementación del proyecto

Tabla 9. Costo del proyecto

ITEM	CANT	EQUIPO	COSTO UNIDAD (S/)	COSTO TOTAL (S/)
1	1	Sistema Adquisicón de Datos	61,466.73	61,466.73
2	8	BI-Sensor de Temperatura de Suelo	5,433.46	43,467.68
3	1	Drone DJI Agras T10	30,786.00	30,786.00
4	1	Aplicación Google Earth Enginee para imágenes satelitales	26,733.85	26,733.85
5	1	Servidor Warehouse	82,000.00	82,000.00
6	1	Implementación Arquitectura Hadoop	50,000.00	50,000.00
7	1	Software SQL	3,659.30	3,659.30
8	1	Software R (gratuito)	0.00	0.00
9	1	Ingeniero de automatización (6 meses)	48,000.00	48,000.00
10	1	Ingeniero de Ciencia de Datos (6 meses)	60,000.00	60,000.00
11	1	Ingeniero de Campo (6 meses)	36,000.00	36,000.00
12	2	Técnico de Campo (6meses)	21,000.00	42,000.00

13	1	Instalación de Equipos e Infraestrucutra	15,000.00	15,000.00
14	3	Otros Recursos	3,000.00	9,000.00
15	1	Otros Gastos	5,000.00	5,000.00
Total				**513,113.56**

Elaboración Propia

Utilizando el 20% de la rentabilidad óptima del útimo año se tiene un capital para invertir S/ 607,111.20 en el costo de la implementación de la Arquitectura de Big data para la Azucarera Azúcar Rica.

Calculamos la Tasa Interna de Retorno (TIR).

TIR = ((Capital Futuro/ Inversión Inicial) -1)*100

TIR = ((607,111.20/513,113.56) -1)*100

TIR = 18.32%

Considerando la Tasa de descuento del 12% como estudio del mercado, se tiene un Valor Actual Neto (VAN)

VAN = S/ 597,040.02

Si se invierte en este proyecto la Azucarera tiene recuperación y ganancia en el futuro.

Capítulo 6: Conclusiones

En el presente capítulo, se colocarán las principales conclusiones de este proyecto de investigación:

- El Big Data puede ayudar notablemente la productividad y la sostenibilidad en la agricultura. El análisis de grandes conjuntos de datos puede revelar patrones y tendencias importantes. Por lo tanto, es más sencillo tomar decisiones.
- Pese a la gran cantidad de trabajos realizados en los últimos años en relación a la implementación de Big Data en el cultivo de la caña de azúcar en Perú. No hay un diseño de arquitectura clara que explique los procedimientos de recopilación, almacenamiento y análisis de los datos.
- El diseño de una arquitectura de Big Data permite visualizar todas las variables bio-físicas y parámetros productivos en tiempo real. El uso de algoritmos de ciencia de datos mejora el proceso de toma de decisiones al incorporar datos sobre cultivos de caña de azúcar. Además, ayuda aumentar la productividad, actuar correctamente sobre los cultivos, optimizar las operaciones, ahorrar costos y proteger el medio ambiente.
- Para trabajo futuros se tiene desafíos de crear un modelo de negocio que ayude a las empresas y los agricultores el poder de saber qué plantar, cuándo y cómo plantar y qué insumos usar para maximizar los rendimientos y la rentabilidad. Además, tener una aplicación que realice clasificaciones de combinaciones de productos, zonas productivas, calidad y qué producto se consume más.
- Con el análisis económico se puede tener recuperación de la inversión y ganancia en el futuro. Y tener mayor equipamiento en las tecnologías de digitalización.

REFERENCIA BIBLIOGRÁFICAS

(s.f.).

Aboul Ella Hassanien, Ashraf Darwish Editors. (2021). *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges* (Vol. 77). (Springer, Ed.)

Agtech. (2017). *Cómo la Big Data está revolucionando la agricultura y la cadena de abastecimiento.*

Aguilar, N. (2017). *Perfila competitivo de la agroindutria azucarera de caña de azúcar*. Obtenido de Virtual Pro - Grupo Colombiano (INGCO): https://www.virtualpro.co/biblioteca/perfil-competitivo-de-la-agroindustria-azucarerade-cana-de-azucar

Amatya, S., Karkee, M., Gongal, A., Zhang, Q. & Whitting, M. D. (2016). Detection of cherry tree branches with full foliage in planar architecture for automated sweet-c herry harvesting. En *Engineering, Biosystems* (págs. 3-15).

Asociados&Apoyo. (2017). *orporación Acucarera del Perú SA. Lima: Asociados Y.* Obtenido de https://www.aai.com.pe/wpcontent/uploads/2018/05/Coazucar-Dic-2017.pdf

Balamurugan Balusamy, Nandhini Abirami. R, Seifedine and Amir H. Gandomi. (2021). *BIG DAT: Concepts, Technology and Architecture* (First ed.). (I. John Wiley & Sons, Ed.) NJ, USA: Editorial Office.

Barth, R., Ijsselmuiden, J. M. M., Hemming, J. & Van Henten E. J. (2017). *Optimising realism of synthetic agricultural images using cycle generative adversarial networks.* Proceedings of the IEEE IROS Workshop on Agricultural Robotics/ Kounalakis, Tsampikos, van Evert, Frits, Ball, David Michael, Kootstra, Gert, Nalpantidis, Lazaros, Wageningen: Wageningen University & Research. Obtenido de http:// library.wur.nl/ WebQuery/ wurpubs/ 533105

Bendre, M. R., Thool, R. C., & Thool, V. R. (September de 2015). Big data in precision agriculture: Weather forecasting for future farming. *In 2015 1st International Conference on Next Generation Computing Technologies (NGT)*, 744-750.

Biswas, S. K., Sinha, N., Purkayastha, B. & Marbaniang, L. (2014). Weather prediction by recurrent neural network dynamics. *International Journal of Intelligent Engineering Infor-matics*, 166– 180.

Corea, F. (2019). Everthing You Need to Know About AI, Big Data and Data Science. En Springer (Ed.), *An Introduction to Data* (Vol. 50, págs. 13 -420). Venice.

Craninx, M., Fieveza, V., Vlaeminck, B. & De Baets, B. (2008). Artificial neural network mod-els of the rumen fermentation pattern in dairy cattle. En *Computers and Electronics in Agriculture* (págs. 226– 238).

Datta, R., Smith, D., Rawnsley, R., Bishop- Hurley, G., Hills, J., Timms, G. & Henry, D. (2015). Dynamic cattle behavioural classification using supervised ensemble classifiers. En *Computers and Electronics in Agriculture* (pág. Computers and Electronics in Agriculture).

Delgado, J. A., Short, N. M., Roberts, D. P., & Vandenberg, B. (2019). *Big Data Analysis for Sustainable Agriculture on a Geosatial Cloud Framework* (Vol. 3). Frontiers in Sustainable Food Systems.

Ebrahimi, M.A., Khoshtaghaza, M.H., Minaei, S. & Jamshidi, B. (2017). Vision-b ased pest detection based on SVM classification method. En *Agriculture, Computers and Electronics in* (págs. 52-58).

FAO. (2009). La agricultura mundial en la perspectiva del año 2050. *Cómo alimentar al mundo en 2050*.

Fung, A. M. (2014). *Plan estratégico del sector agricultura.* CENTRUM Católica, Lima Perú: Tesis de Maestría.

Gandrud, C. (s.f.). Reproducible Research with R and RStudio.

Ghosh, S. & Koley, S. (2014). Machine learning for soil fertility and plant nutrient management using back propagation neural networks. En *International Journal on Recent and Innovation Trends in Computing and Communication* (págs. 292– 297).

Grinblat, G. L., Lucas, C. U., Mónica, G. L. & Granitto, P. M. (2016). Deep learning for plant identification using vein morphological patterns. En *Computers and Electronics in Agriculture* (págs. 418– 424).

Hansen, M. F., Smitha, M. L., Smitha. L. N., Michael, G., Salterb, M. G., Baxterc, E. M., Farishc, M. & Grieved, B. (2018). Towards on- farm pig face recognition using convolutional neural networks. En *Computers in Industry* (págs. 145-152).

HELFGOTT, S. (1977). *El Cultivo de Caña de Azúcar en la Costa.* Universidad Nacional Agraria, Lima.

Hinnell, A. C., Lazarovitch, N., Furman, A., Poulton, M. & Warrick, A. W. (2010). Neuro-D rip: estimation of subsurface wetting patterns for drip irrigation using neural networks. En *Irrigation Science* (págs. 535– 544).

Holmes, D. E. (2017). A Very Short Introducion BIG DATA. En *BIG DATA* (págs. 6-215). New Year: OXFORD University Press.

Huang, Y., Chen, Z. X., Tao, Y. U., Huang, X., Z., & Gu, X. F. (s.f.). Agricultural remote sensing big data: Management and applicatons. *Journal of Integrative Agriculture*(17), 1915-1931.

Jonny Rocio Aquize Diaz. (2024). *Estados Financieros Auditados.* CASA GRANDE, Trujillo.

Kurumatani, K. (2018). Time series prediction of agricultural products price based on time alignment of recurrent neural networks. *17th IEEE International Conference on Machine Learning and Applications*, (págs. 82-88).

Kussul, N., Lavreniuk, M., Skakun, S. & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosciences and Remote Sensing Letters*, (págs. 778–782).

Lesser, A. (2014). *Big data and big agriculture.* Gigaom Res.

Mannino, M. V. (2007). Diseño y desarrollo de aplicaciones. En M. Grawhill (Ed.), *Administración de base de datos* (3rd ed., págs. 249-293). Denver, Colorado.

Marin, R. (2008). Técnicas, métodos y aplicaciones. En Mcgraw-Hill (Ed.), *Inteligencia Artificial* (1 ed., págs. 32-85).

Mayhua, E. L. (2016). *Sistema de riego por goteo automático utilizando una red de sensores inalám-bricos.* Arequipa: Revista de Investigación.

Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A., & Bhansali, S. (2019). Machine Learning Techniques in Wireless Sensor Network Based Precision Agriculture. *Journal of the Electrochemical Society*, 167.

MINAGRI. (21 de 12 de 2020). Obtenido de Herramientas Tecnológicas Satelitales del MIDAGRI: https://siea.midagri.gob.pe/portal/gee/index.html

Mora, H., & Rosas, J. (2019). *Diseño, desarrollo e implementación de una red de sensores inalámbricos (WSN) para el control, monitoreo y toma de decisiones aplicado en la agricultura de precisión basado en internet de las cosas (IOT).* Universidad Ricardo Palma, Facultad de Ingeniería. Caso de estudio cultivo de frijol.

Morales, I.R., Cebrián, D. R., & Blanco, E. F. (2016). Early warning in egg production curves from commercial hens: an SVM Approach. En *Computers and Electronics in Agriculture* (págs. 69– 179).

Moshou, D., Bravo, C., Jonathan, W., Wahlen, S., Cartney, M. A. & Ramona, H. (2015). Auto-matic detection of 'ye llow rust' in wheat using reflectance measurements and neural net-works. En *Computers and Electronics in Agriculture* (págs. 173-188).

Palacios, G. (2017). *Diseño e implementación de un sistema para el monitoreo de cultivos nativos utilizando Internet del Todo y redes Fog.* Universidad Nacional del Altiplano, Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas.

Palangpour, P., Venayagamoorthy, G. K. & Duffy, K. (2006). Recurrent neural network based predictions of elephant migration in a South African game reserve. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, (págs. 4084– 408). Obtenido de The 2006 IEEE International Joint Conference on Neural Network Proceedings: https:// ieeexplore.ieee.org/ document/ 1716662

Pantazi, X. E., Tamouridou, A. A., Alexandridis, T. K., Lagopodi, A. L. & Kashefi, J. (2017). Evaluation of hierarchical self- organising maps for weed mapping using UAS multispectral imagery. En *Computers and Electronics in Agriculture* (págs. 224-230).

Pegorini, V., Karam, L.Z. & Pitta, L.S.R. (2015). In vivo pattern classification of ingestive be-havior in ruminants using fbg sensors and machine learning. En *Sensors* (págs. 28456– 28471).

Pierce, F. J., & Nowak, P. (1999). Aspects of precision agriculture. In Advances in agronomy. Academic Press.

Rai, A. (2021). Emerging Pedagogies of Deep Learning, Machine Learning and Internet of Things. En G. S. Patel (Ed.), *SMART AGRICULTURE* (págs. 12-158). London.

Ramos, P. J., Prieto, F. A., Montoya, E. C. & Oliveros, C.E. (2017). Automatic fruit count on coffe branches using computer vision. En *Computers and Electronics in Agriculture* (págs. 9-22).

Roberge, J. (2021). The Cultural Life of Machine Learning. Quebec.

Saraiva, M., Protas, É., Salgado, M., & Souza Jr, C. (2020). Automatic Mapping of Center Pivot Irrigation Systems from Satellite Images Using Deep Learning. En *Remote Sensing* (pág. 558).

Weiss, U., Biber, P., Laible, S., Bohlmann, K. & Zell, A. (2010). Plant species classification us-ing a 3D LIDAR sensor and machine learning. En *Ninth International Conference on Machine Learning and Applications* (págs. 12-14).

Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.J. (2017). Big data in smart farming-a review. Agric. Syst.

Yan, Y. (2023). *Learning R and Python for Business School Students.* (L. S. Library, Ed.) Newcastle, UK: Cambridge Scholars.

Zhang, M., Changying, L. & Fuzeng, Y. (2017). Classification of foreign matter embedded inside cotton lint using short wave infrared (SW IR) hyperspectral transmittance imaging. En *Computers and Electronics in Agriculture* (págs. 75-90).

Zhao, D., & Li, Y. R. (2015). *Climate change and sugarance production: potential impact and mitigation strategies.* International Journal of Agronomy.