Universidad Autónoma del Estado de Hidalgo

Instituto de Ciencias Básicas e Ingeniería

PROPUESTA DE CATEGORIZACIÓN DE LOS NIVELES DE SEQUÍA EN EL ESTADO DE HIDALGO UTILIZANDO MODELOS DE APRENDIZAJE NO SUPERVISADO E IMÁGENES MULTIESPECTRALES

TESIS

Que para obtener el título de Ciencias Computacionales

DIRECTOR

Dr. Aldo Márquez Grajales

CO-DIRECTOR

Dr. Fernando Salas Martínez

PRESENTA

Daniel Reyes Bistrain

Pachuca de Soto, Hidalgo., 2025

Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
School of Engineering and Basic Sciences

Mineral de la Reforma, Hgo., a 26 de noviembre de 2025

Número de control: ICBI-D/3024/2025
Asunto: Autorización de impresión.

MTRA. OJUKY DEL ROCÍO ISLAS MALDONADO
DIRECTORA DE ADMINISTRACIÓN ESCOLAR DE LA UAEH

Con Título Quinto, Capítulo II, Capítulo V, Artículo 51 Fracción IX del Estatuto General de nuestra Institución, por este medio, le comunico que el Jurado asignado al egresado de la Licenciatura en Ciencias Computacionales **Daniel Reyes Bistrain,** quien presenta el trabajo de titulación **"Propuesta de categorización de los niveles de sequía en el Estado de Hidalgo utilizando modelos de aprendizaje no supervisado e imágenes multiespectrales",** ha decidido, después de revisar fundamento en lo dispuesto en el Título Tercero, Capítulo I, Artículo 18 Fracción IV; dicho trabajo en la reunión de sinodales, **autorizar la impresión del mismo**, una vez realizadas las correcciones acordadas.

A continuación, firman de conformidad los integrantes del Jurado:

Presidente: Dra. Anilú Franco Árcega

Secretario: Dra. Guadalupe Carmona Arroyo

Vocal: Dr. Aldo Márquez Grajales

Suplente: Dr. Fernando Salas Martínez

Sin otro particular por el momento, reciba un cordial saludo.

Atentamente
"Amor, Orden y Progreso"

Mtro. Gabriel Vergara Rodríguez
Director del ICBI

GVR/YCC

Ciudad del Conocimiento, Carretera Pachuca-Tulancingo Km. 4.5 Colonia Carboneras, Mineral de la Reforma, Hidalgo, México. C.P. 42184
Teléfono: 771 71 720 00 Ext. 40001
direccion_icbi@uaeh.edu.mx, vergarar@uaeh.edu.mx

Resumen

La sequía presenta un problema a nivel global, debido a su creciente aparición y severidad en distintas partes del planeta, el estado de Hidalgo, tuvo recientemente la mayor crisis sobre este problema en 10 años. Por lo cual, la oportuna y eficaz identificación de este fenómeno es fundamental para mitigar sus consecuencias con medidas congruentes y acertadas.

El objetivo de esta investigación fue poder aplicar distintas técnicas de minería de datos con métodos de agrupamiento no supervisados, pobre imágenes multiespectrales para identificar y analizar patrones de sequía, así como evaluar la eficiencia de cada algoritmo sobre este problema. Con la finalidad de identificar si estos algoritmos pueden ser una herramienta para los expertos y así disminuir el margen de error humano que estos puedan tener.

Se siguió la metodología KDD (Knowledge Discovery in Databases) para el desarrollo de la investigación, obteniendo las imágenes multiespectrales mediante el programa Landsat 8 y 9 de la región de Hidalgo durante 4 periodos con niveles de sequía distintos. Estas fueron preprocesadas y transformadas con la finalidad de obtener una base consistente e integra. Posteriormente se aplicaron los algoritmos de agrupamiento, K Means, Jerárquico y DBScan para segmentar las zonas afectadas. Se realizo una evaluación basada en índices de validación a cada uno de los grupos y algoritmos, identificando el mejor resultado y recreando las imágenes basadas en las agrupaciones obtenidas por el mejor algoritmo.

La investigación arrojo una ligera ventaja de K Means sobre el método Jerárquico y una deficiencia de DBScan para la identificación de la sequía, puesto que la configuración de los parámetros y el coste computacional impidieron obtener resultados útiles para la investigación por parte de este algoritmo. La reconstrucción de las imágenes demostró una agrupación sensible a la severidad de la sequía, aunque hay que tomar en cuenta que, la forma de identificar sequía y asignar un clúster fue distinta dependiendo la fecha y el nivel de sequía que contenía la imagen multiespectral.

Se concluye que el uso de algoritmos de minería de datos en la identificación de la sequía es una herramienta bastante útil que puede permitir a los expertos minimizar el margen de error y algunas ambigüedades generadas por el ser humano. Además, se demostró que K Means tuvo un mejor desempeño sobre del método Jerárquico y DBScan para la identificación de la sequía en el Estado de Hidalgo siendo una herramienta útil en esta tarea.

Abstract

Drought is a global problem because its appearance and severity are growing. The state of Hidalgo recently experienced its worst drought crisis in 10 years. Therefore, the quick and effective identification of this phenomenon is essential to mitigate its consequences with correct measures.

The objective of this research was to apply different data mining techniques, using unsupervised clustering methods, on multispectral images. The goal was to identify and analyze drought patterns. We also wanted to evaluate the efficiency of each algorithm for this problem. We did this to see if these algorithms can be a tool for experts to reduce human error.

We followed the KDD (Knowledge Discovery in Databases) methodology. We got the multispectral images from the Landsat 8 and 9 program for the Hidalgo region during 4 periods with different drought levels. These images were preprocessed and transformed to get a consistent and complete database. After that, we applied the clustering algorithms: K-Means, Hierarchical, and DBScan to segment the affected zones. We did an evaluation using validation indices for each cluster and algorithm. We identified the best result and recreated the images based on the clusters from the best algorithm.

The research showed a small advantage for K-Means over the Hierarchical method, and a deficiency for DBScan. This was because the parameter configuration and the computational cost of DBScan stopped us from getting useful results. The reconstruction of the images showed clustering that was sensitive to drought severity. However, we must consider that the way to assign a cluster was different depending on the date and the drought level in the multispectral image.

We conclude that using data mining algorithms for drought identification is a very useful tool. It can allow experts to minimize human error and ambiguities. Also, we showed that K-Means had a better performance than the Hierarchical method and DBScan, making it a useful tool for this task in the State of Hidalgo.

Índice de Contenido

Índice de Formulas

Índice de Tablas

Índice de Figuras

CAPÍTULO 1 INTRODUCCIÓN

La sequía representa uno de los mayores desafíos que el medio ambiente enfrenta. En México el impacto de la sequía se ha intensificado en los últimos años debido a los efectos del cambio climático, las deforestaciones, una mala gestión de los recursos hídricos y las actividades humanas sobre los ecosistemas naturales. Tener herramientas para la identificación de gravedad de este fenómeno es indispensable para mitigar los efectos de la misma con mayor eficiencia. Actualmente la identificación de la sequía es un proceso rudimentario y dependiente de expertos que pueden hacer que el proceso pueda están influenciado por intereses, y en algunos casos, haya diferencias entre expertos. Por lo cual proporcionar una herramienta que permita a los expertos mitigar este fenómeno es fundamental.

Salas Martínez describe como a mediados del 2024 más del 70% del territorio nacional presentaba algún grado de afectación a causa de la sequía (Salas-Martínez, 2021), convirtiendo esta situación en una de las crisis hídricas más severas de las últimas décadas. Este escenario revela la necesidad de desarrollar herramientas tecnológicas que permitan comprender, medir y mitigar el impacto del fenómeno de la sequía, especialmente en regiones que se encuentra vulnerables ante este fenómeno como lo es el estado de Hidalgo.

Ante este escenario, la aplicación de técnicas de Minería de Datos y Aprendizaje Autónomo ofrece una alternativa viable para el análisis objetivo de grandes volúmenes de información climática y ambiental. Particularmente, los modelos de aprendizaje no supervisados dan la oportunidad de agrupar comportamientos similares en conjuntos de datos multivariables sin la necesidad de requerir etiquetas o categorías predefinidas. Esta característica resulta valiosa para el estudio de la sequía, dado que permite clasificar zonas con condiciones atmosféricas y ecológicas semejantes, facilitando la creación de mapas de intensidad y distribución con una mayor precisión.

Al día de hoy, la identificación de fenómenos ocasionados por la sequía y la sequía misma son relativamente pocos en comparación con otras áreas de estudios. Como punto de partida existen estudios capaces de identificar fenómenos relacionados con la falta de lluvias y sus afectaciones, los efectos causados en grupos específicos plantas y árboles, la perdida de volumen de los cuerpos de agua, la identificación de sequía por medio de imágenes multiespectrales o incluso, por medio de imágenes obtenidas a través de medios comunes de información como lo son Google Earth.

Muchos de estos trabajos se centran identificar estos fenómenos en regiones especificas a lo largo del mundo, lo que dificulta encontrar estudios aplicados a

México y aún más, estudios especializados en Hidalgo, donde su diversidad de biomas dificulta aún más la identificación de los niveles de sequía. Esto es importante, ya que, muchos de estos estudios aplican distintas métricas y algoritmos a los datos, resultando es estudios donde la exploración de algoritmos muchas veces se limita a interpretar los resultados y no a evaluar la calidad de estos, desconociendo si existe una mejor forma de generar agrupaciones. O creando confusión si existe una mejor forma para esta región especifica.

El presente trabajo de investigación propone la categorización de los niveles de sequía en el estado de Hidalgo mediante el procesamiento de imágenes multiespectrales Landsat y el uso de modelos de aprendizaje no supervisado. La metodología se sustenta en la aplicación de tres algoritmos de agrupamiento K Means, Clustering Jerárquico y DBScan, los cuales serán evaluados y comparados a través de índices de validación como Calinski Harabasz, Davies Bouldin y Silhouette, esta aproximación busca determinar cuál de los algoritmos ofrece una clasificación coherente con la realidad observada, reduciendo la dependencia de interpretación subjetiva y Contribuyendo a la automatización del monitoreo ambiental.

La presente investigación busca poder analizar el uso de algoritmos de agrupamiento no supervisado, para estimar la capacidad de estos en la identificación de distribución e intensidad de la sequía mediante imágenes multiespectrales de programa Landsat a lo largo de varios meses, en el estado de Hidalgo, México.

De la misma forma, explora distintos objetivos a la hora de la realización de dicha investigación como lo son, los diferentes tipos de sequía en el estado de Hidalgo mediante algoritmos de agrupamiento no supervisado evaluando los resultados de las agrupaciones frente a resultados oficiales. Validar la efectividad de cada uno de los algoritmos mediante índices de validación para demostrar la diferencia lógica del desempeño de cada uno en la identificación de niveles de sequía. Y finalmente, comparar los resultados mediante los índices y datos previamente evaluados a través de organismos expertos acerca de la sequía en los mismos periodos de tiempo para validar la veracidad y calidad de los resultados obtenidos.

Esto le lograra a través de diversos capítulos en los cuales de aportar información de forma estructurada y conveniente para la correcta interpretación. A través de 5 capítulos se desarrolló comenzando con el marco teórico en el cual se desarrollan cada uno de los conceptos básicos e imprescindibles para la investigación, estos se desarrollan es tres etapas correspondientes al marco conceptual, tecnológico y metodológico. En el capítulo 2 se desarrolla el are4a de aplicación explorando temas referentes al campo estudiado, el cual es la sequía, y toda aquella información

referente a su estudio y obtención de datos. En el tercer capítulo se estudian aquellos trabajos con afinidades a este trabajo evaluando sus características, campos de estudio, objetivos y conclusiones de cada uno. En el siguiente y cuarto capitulo, se ejecuta libremente la metodología elegida conforme a la definición y estructura, obteniendo y almacenando los resultados y análisis de la misma. Por último, se genera información relevante y las conclusiones acerca de los resultados de la investigación, plasmando los hallazgos y todo el conocimiento obtenido.

PLANTEAMIENTO DEL PROBLEMA

A lo largo de los años, México ha afrontado diferentes fases de sequía. En el último año, el territorio nacional se encontró afectado de manera crítica. Según el North American Drought Monitor (NADM) a finales de mayo del 2024 aproximadamente el 76% del país se encontraba afectado por condiciones de sequía.

En el mismo periodo, datos oficiales (Conagua, 2025) indican que más del 65% de México presentaba condiciones de sequía por lo que se ha considerado la peor crisis hídrica de los últimos años.

Ante este escenario, se vuelve imprescindible analizar e interpretar los patrones de sequía a nivel nacional. Dado que es un problema que persiste a lo largo de los años, el ser humano ha tenido la necesidad de automatizar y simplificar procesos que ayuden a monitorear y predecir el clima, esto con amplios beneficios, que van desde una mejora en la agricultura moderna, hasta la prevención de desastres naturales y fenómenos atmosféricos.

La búsqueda por la automatización y simplificación de estos procesos, ha llevado al desarrollo de instrumentos altamente complejos y técnicos. Dichos procesos requieren de especialistas capaces de leer e interpretar información y convertirla en conocimiento útil.

Para ser específicos, el área analizada en este trabajo será el estudio de los niveles de sequía. Este proceso es rudimentario y muy complejo, debido a que los datos presentan problemas en su estructura y temporalidad, así como las actuales metodologías cuentan con el aporte subjetivo de expertos del clima. En el contexto meteorológico, la ambigüedad y subjetividad entre los expertos locales, pueden generar opiniones divididas y deficiencias en los resultados, así como, una mala interpretación puede llevar a pérdidas de recursos en la agricultura, turismo, ecología

o todos aquellos campos con relación al medio ambiente y el clima que se ven afectados por la sequía.

Como ya se ha estudiado, el poder definir si existe la sequía, requiere implementar monitores capaces de identificar muchos factores como, por ejemplo: el aire, temperatura, velocidad del viento, humedad, el color de la vegetación, formación y presencia de las nubes, los cuerpos de agua y sus niveles, etc. Todos estos aspectos ayudan a identificar la sequía.

La sequía es un fenómeno real y en este está definido por el "Intergovernmental Panel of Climate Change" como un periodo excepcional de escasez de agua para los ecosistemas y la población humana debido a múltiples factores (Edenhofer, 2013), estudios como el *Methodological estimation to quantify drought intensity based on the NDDI index with Landsat 8 multispectral images in the central zone of the Gulf of Mexico*. Describen que la sequía tiene efectos colaterales que incrementan la temperatura y condiciones climáticas de la atmósfera ayudando a formar otros fenómenos naturales como *"El Niño"* (Salas-Martínez, 2023).

Existen diversos estudios sobre cómo la sequía afecta estados o regiones de México, como por ejemplo el *"Analysis of the Evolution of Drought through SPI and Its Relationship with the Agricultural Sector in the Central Zone of the State of Veracruz, Mexico"* (Salas-Martínez, 2021). El cual estudia cómo es que la sequía afecta al sector agrícola y ganadero en una región donde dichas actividades son principales. Este trabajo es un ejemplo del por qué un estudio preciso de la sequía puede ayudar a mitigarla.

JUSTIFICACIÓN

Actualmente, el Servicio Meteorológico Nacional implementó una metodología para la detección de la sequía en México, mediante información institucional, análisis geoespacial y el aporte de expertos locales. Lo cual sugiere una subjetividad en la interpretación de la información y que hace que dicha metodología no sea reproducible completamente. Adicionalmente, esta información es realizada cada 15 días, presentando un desfase en la información, es decir, en una quincena se presentan los datos de la anterior.

Además, el problema de un error al identificar un tipo de sequía se puede traducir en una agricultura deficiente, escasez de agua para actividades humanas, actividades dependientes del medio ambiente y/o disponibilidad de agua (Salas-Martínez, 2021).

Existen algunos índices que tratan de describir la presencia de la sequía en diversas regiones de México, cómo el “Normalized Difference Drought Index (NDDI)”, el cual describe al fenómeno mediante el uso de la vegetación y presencia de agua sobre imágenes multiespectrales, haciéndola una herramienta eficiente para dicho fin (Salas-Martínez, 2023).

Este proyecto puede presentar algunas dificultades como el nivel de asertividad que los algoritmos presentan en contraste a los resultados de los expertos debido a toda la experiencia y la identificación de características únicas de los datos en este campo de estudio. Es claro, que el trabajo que dichos expertos han realizado por años, no puede ser sustituido, lo que se busca es proporcionar una herramienta que disminuya aquellas ambigüedades de los datos.

Un resultado positivo de este estudio traerá múltiples beneficios en distintas disciplinas, por ejemplo, los expertos ya no dependen únicamente de sus criterios para deliberar en sus resultados sobre los índices de sequía, tendrán una herramienta confiable que los guíe en sus decisiones reduciendo el margen de error y las ambigüedades generadas por una equivocación humana.

Cabe destacar que el obtener resultados positivos permitirá eliminar esa brecha tecnológica entre los expertos del clima y la aplicación de nuevas tecnologías, proporcionando una herramienta que no requiera conocimiento a priori en computación ni apreciaciones subjetivas.

OBJETIVO GENERAL

Analizar el uso de algoritmos de agrupamiento K Means, Agrupamiento Jerárquico y DBScan, para estimar la identificación de distribución e intensidad de la sequía mediante imágenes multiespectrales Landsat 9 durante los meses de febrero, julio y octubre del 2024, en el estado de Hidalgo, México.

OBJETIVOS ESPECÍFICOS

- Identificar la distribución e intensidad de la sequía en el estado de Hidalgo mediante algoritmos de agrupamiento no supervisado para evaluar la efectividad de estos frente a los expertos.

- Validar la efectividad de cada uno de los algoritmos mediante índices de validación para identificar la eficiencia de cada uno en la identificación de niveles de sequía.

- Comparar los resultados mediante los índices y datos oficiales acerca de la sequía en los mismos periodos de tiempo para validar la veracidad y calidad de los resultados obtenidos.

CAPÍTULO 2 MARCO TEÓRICO

2.1 Marco Conceptual

En esta sección del documento se explorarán los diversos conceptos claves para poder entender las bases de esta investigación. Estos conceptos darán un panorama amplio acerca de los conceptos, algoritmos, organizaciones y demás temas clave implicados en este documento.

Se abordarán temas clave como la minería de datos y la importancia de la misma en el descubrimiento de información y automatización de procesos. Sus distintas formas de obtener información y la clasificación de los algoritmos de agrupamiento. Además de sus distintos índices de validación y su importancia en la evaluación de calidad de grupos.

2.1.1 Minería De Datos

La Minería de Datos (MD) desde sus inicios ha sido un proceso muy importante en cualquier área donde se trabaje con grandes volúmenes de información, es el proceso de descubrir patrones y relaciones significativas en grandes conjuntos de datos, utilizando técnicas de análisis estadístico y algoritmos de aprendizaje automático (Han, 2012). Este conjunto de técnicas se utiliza para extraer información útil de grandes conjuntos de datos, lo que puede ser útil en áreas como el marketing, la investigación de mercado, la medicina, la ciencia y la tecnología (Rubiños, 2024).

De manera general, existen dos modelos que pueden generarse a través del uso de la MD, el descriptivo y el predictivo (Han, 2012). El modelo descriptivo se basa en la identificación de patrones y relaciones en los datos, obteniendo así información oculta dentro de los datos. Por su parte, el predictivo tiene el fin de predecir comportamientos desconocidos considerando un conjunto de variables conocidas.

En el modelo descriptivo, uno de sus objetivos es encontrar grupos naturales dentro de los datos, lo que se conoce como agrupamiento o clustering. El algoritmo de clustering agrupa los datos en función de su similitud (Han, 2012). Por ejemplo, la Minería de Datos descriptiva para identificar grupos de clientes con comportamientos similares (Rubiños, 2024).

Esta tecnología ha ido evolucionando por desde las formas más primitivas hasta llegar a ser un proceso muy complejo del cual se requiere estudio, comprensión y experiencia no sólo en sí misma, sino también su campo de aplicación, como ya se vio

tiene muchas aplicaciones, pero de la misma forma, tiene muchas formas en las que puede desarrollarse, emplearse y organizarse. Las formas en que esta tecnología ha ido evolucionando son las siguientes:

2.1.2 Análisis Manual

Muy similar al proceso de identificación de sequía de la CONAGUA este modelo se basa en una serie de expertos capaces de obtener resultados basados en experiencia, estudio y complejos cálculos que pueden ser manuales o computados, estos resultados requieren de una compleja recolección de datos y varios barridos manuales para poder obtener un resultado medible para los expertos, este análisis comprende varios problemas, uno de ellos es el conflicto de intereses de los expertos que pudiera alterar el resultado, al no existir un proceso que elimine cualquier tipo de predilección por los resultados, es muy común que estos se vean influenciados por factores externos a ellos (Salas-Martínez, 2021). El segundo problema más común entre este método, es la imposibilidad para identificar patrones de forma automática, cuando se busca un resultado de forma manual, se ignoran los patrones que los datos pueden o no sugerir de estos mismos. Por lo cual una tarea humana dificulta obtener información más allá del objetivo.

Aun así, este método fue el precursor para obtener resultados favorables y muchas veces acertados. también creó la necesidad de automatizar estos procesos para un bien mejor.

2.1.3 Sistemas Operacionales

Los sistemas operacionales buscan procesar grandes cantidades de información en tiempo real, su objetivo es obtener resultados con la información actualizada y real. Esto se logra de forma estructurada y estos se dividen en dos partes:

OLTP (Online Transaction Processing): Este modelo se basa únicamente en procesar información real, las grandes empresas lo usan para así poder procesar pedidos, facturas, llevar contabilidad, o generar estadísticas del estado actual de los datos. Normalmente estos están montados en entornos de producción para agilizar su velocidad y la obtención y recolección de los datos, este sistema prioriza la escritura de los datos sobre la lectura, su estructura normalmente es unidimensional donde cada fila es un objeto y cada columna es un atributo (Han, 2012).

OLAP (On-Line Analytical Processing): Se utiliza para generar reportes o informes basados en el universo histórico de la misma empresa, estos buscan poder realizar predicciones, pronósticos, resúmenes o descubrir información oculta de las grandes

cantidades de datos que la empresa disponga. Su estructura es multidimensional donde se pueden manejar los datos como atributos en forma de cubo donde cada celda puede representar desde un valor hasta una medida multidimensional (Han, 2012).

2.1.4 Sistemas De Consultas De Bases De Datos

A pesar de no ser el sistema de manejo de información predilecto por las pequeñas y medianas empresas, la facilidad para organizar información en sistemas, SQL y NoSQL lo hace perfecto para el manejo, organización y consulta de información. Estos sistemas permiten añadir reglas, restricciones, disparadores y un excelente manejo de copias de seguridad. Estos sistemas son ideales para almacenar registros unidimensionales y tratar la información, pero carecen de la capacidad para procesarla y descubrir patrones dentro de la misma información, estos sistemas se limitan a poder ejecutar complejas consultas creadas por el mismo usuario y mostrar solo la información solicitada (Wegmann, 2021). Son una herramienta con la capacidad de guardar y leer información, no incluyen la capacidad de análisis y/o la capacidad de descubrir nueva información a partir de la ya existente.

2.1.5 Sistemas Informativos

Estos son los más completos, pero también los más complejos de los sucesores, por decirlo de alguna forma, comprenden un conjunto de todos los anteriores, británica lo describe como un conjunto integrado de componentes de recolección, almacenamiento, y procesamiento de datos; del cual se obtendrá conocimiento, información y productos digitales (Jovic, 2014).

Para ampliar esta definición se comprende que los sistemas informativos son aquellos sistemas que implementan robustos sistemas con objetivos distintos para un mismo fin, un sistema informativo puede conjuntar desde complejos sistemas informativos hasta robustos componentes de hardware, pudiendo o no estar centralizados o diferidos (Jovic, 2014). Estos deben permitir obtener datos, informes y estadísticas de los datos para su fácil interpretación.

Este tipo de herramienta es multidisciplinar y no solo es para el uso de la informática al ser su rama base. Esta herramienta fue hecha para crear entornos multidisciplinares, con ella se puede obtener, almacenar y procesar información de cualquier rama mientras esta esté organizada debidamente.
Por ejemplo, una tienda de autoservicio por medio de estrictos registros de entradas de almacén, salida en caja, registros de merma y devoluciones por mercancía

defectuosa; puede obtener predicciones sobre cómo mejorar las ventas agrupando mercancía que está en tendencia llevarse junta desde las compras de otros usuarios. También pueden organizar los pedidos sobre aquellos productos que podría no venderse y abastecer menos o incluso que la tendencia apunta a una venta superior al promedio y así tener un mejor abastecimiento y así minimizar el desabasto o merma en la mercancía. Otro uso será el de poder identificar momentos cruciales de la disertación de los clientes, prevenirlos y volverlos a traer con promociones hechas para ese grupo de clientes.

Como se puede observar en este sencillo ejemplo, los sistemas informáticos pueden tener múltiples beneficios, únicamente se analizó un mercado de tiendas de autoservicios, pero estas tendencias pueden ayudar en empresas multinacionales o incluso en el desarrollo científico y así obtener avances en innumerables campos como el medio ambiente, microbiología, medicina, obtención de nuevas formas de energía, desarrollo de nueva tecnología etc.

Sus objetivos pueden ser, obtener información inferible a través de los millones de datos que se manejan, obtener patrones ocultos en los datos con complejos algoritmos, descubrir tendencias de los datos para poder predecir datos o resultados de acuerdo a un complejo análisis del histórico de los datos (Jovic, 2014).

Uno de los aspectos más importantes de esta técnica, es el uso, gestión y manejo del almacenamiento, esta tarea es muy delicada, un solo atributo mal formado en los datos, puede generar resultados erróneos y variantes. El manejo de la información requiere de distintas técnicas y tecnologías. El aspecto más básico para el almacenamiento de datos son las bases SQL y las NoSQL, estas herramientas permiten tener una gestión de datos estructurada, veloz, confiable y con acceso a copias de seguridad que ayudan a tener información robusta y óptima (Han, 2012).

Un sistema informático puede tener una o múltiples fuentes de datos, estas pueden manejar datos con diferentes normalizaciones, estructuras y formas, pero, para poder ser utilizadas como una forma eficiente, se requiere de un componente fundamental para los sistemas informativos, el cual es un Data Warehouse (WD), este es un sistema capaz de reunir información de múltiples bases de datos, y otras fuentes de información, sus modelos de datos se basan en los sistemas OLAP, la información debe estar lista para realizar consultas complejas, realizar informes estadísticas y detectar tendencias en los mismos datos (Han, 2012) .

El uso de este componente se basa en el modelo ETL (Extract, Transform, Load), en el primer parte, la información debe pasar por complejos, robustos y relativamente

pequeños gestores de bases de datos para obtener información sólida y eficiente, en el segundo paso esta debe ser transformada a primitivos datos que deben ser limpiados, validados y libres de ruido para evitar guardar información sucia o errónea. En el último paso, esta información debe ser cargada a los robustos sistemas llamados Data Warehouse, en formas multidimensionales, si los pasos anteriores se realizaron correctamente, ahora se tiene información limpia, útil y especializada, lista para ser analizada y procesada (Rubiños, 2024).

Una vez que se tiene información confiable, es hora de integrar complejos sistemas de minería de datos, se pueden aplicar desde algoritmos capaces de agrupar datos con la finalidad de encontrar patrones o información oculta entre ellos, de igual forma aplicar algoritmos capaces de predecir tendencias, resultados o acciones de los mismos datos, basados en grupos de entrenamiento y modelos predictivos altamente complejos (Rubiños, 2024).

Adicional a esto se pueden incluir modelos de redes neuronales o usar los datos para poder entrenar modelos de inteligencia artificial y así poder crear procesos donde se creará nueva información y generar agentes basados en los datos (Jovic, 2014). Esto ya permite obtener predicciones, agrupamientos, informes, estadísticas, estados actuales etc.

El uso de esta técnica es altamente eficiente de la mano de un experto de los mismos datos, ya que, aunque existen modelos muy avanzados, aún se requiere de expertos capaces de apreciar los resultados, organizar los datos y tomar las decisiones finales para mejorar cada uno de los propósitos de esta herramienta.

También requieren una costosa y compleja implementación, el costo computacional de analizar toda la información representa un costo considerable para las empresas, aun así, esta técnica es ampliamente utilizada por grandes empresas tales como: Amazon, Microsoft, Apple, Google, etc.

2.1.6 Técnicas De Minería De Datos

La minería de datos es una materia con múltiples usos y beneficios, así como de un largo estudio de conceptos, definiciones y algoritmos. Estas definiciones y categorías permiten al usuario poder escoger el enfoque y caminos correctos para llegar al resultado deseado.

Dentro de la minería de datos existen dos grandes vertientes, el análisis de información para la verificación de datos y el uso de datos orientado al descubrimiento de información. Estas vertientes son aquellas que hacen que la minería de datos sea una

herramienta muy útil para el descubrimiento de información y análisis de la misma. La diferencia entre ambas radica en que la primera busca patrones dentro de la misma información y por medio de nuevos datos, con la finalidad de reconocer patrones de información donde se desconoce algún posible comportamiento normal en los datos y así crear nuevos conocimientos de los mismos datos. La segunda busca reconocer dichos patrones y descubrir información oculta dentro de la misma, poder clasificar los datos y de una misma forma, predecir valores booleanos, numéricos o categóricos únicamente usando los datos ya existentes y nuevos objetos. La minería de datos está en una constante evolución con el desarrollo de nuevas tecnologías tales como la inteligencia artificial, redes neuronales, sistemas de almacenamiento de datos, computación cuántica, etc. (Wegmann, 2021).

Dadas las descripciones anteriores los grupos existen dos tipos de análisis: supervisados y no supervisados. Estos análisis permiten manejar la información en enfoques muy diferentes, los sistemas supervisados son aquellos que realizan algoritmos de identificación de patrones, estos permiten a las empresas poder predecir agrupaciones en los datos dados los datos previos, de la misma forma estos son capaces de predecir datos numéricos, booleanos o categóricos. Esto se basa en obtener los datos de grandes repositorios y conocer las entradas y las salidas deseadas, para que así el algoritmo etiquete los datos de forma deseada (Wegmann, 2021).

Un ejemplo clásico de esta disciplina es el de poder identificar posibles fraudes en un banco en la autorización de un crédito, en este ejemplo, cada solicitud está llena de variables para analizar, como es sabido, antes de ser autorizado un crédito, es cliente debe llenar una registro sobre datos de él, como lo son: domicilio, fuentes de ingreso económico, tiempo en esa fuente de empleo, historial crediticio, uso que se le dará al crédito, estado de salud, edad, incluso acciones. Todo esto se transforman en datos categóricos, numéricos y booleanos que se almacenan y se comparan con todo el histórico de clientes donde cada crédito se concluye en una variable, es o no fraude, para lograr identificar si el cliente puede o no ser factible para entregar dicho crédito, se puede hacer un modelo de regresión logística y así determinar qué tan propenso es al pagar o no el crédito en un futuro. Esto como es sabido es una aproximación y no siempre es 100% verídico y real, pero, usando las técnicas adecuadas y con una base de datos lo suficientemente vasta y correcta, se llega a una aproximación que ayuda al banco a tomar una decisión.

Por otro lado, las técnicas de minería no supervisadas, buscan obtener información oculta en los datos, gracias a diferentes algoritmos que permiten detectar anomalías en los datos, permitir obtener puntos de vista sobre los mismos e incluso agruparlos y

así obtener información sobre sus comportamientos como un conjunto. Un ejemplo de las técnicas de minería no supervisadas es el como una empresa de internet y telefonía aplica un algoritmo de agrupamiento para identificar el uso de servicio de sus usuarios, tomando en cuenta aspectos como el consumo de internet, uso de llamadas, uso de la televisión y edad de los usuarios. Todos estos datos permiten a la empresa crear grupos por medio de algoritmos de agrupación como el conocido K Means, el cual permite generar grupos dado un número esperado de grupos y este los agrupa obteniendo conjuntos etiquetados tomando en cuenta las afinidades de sus datos o usando algoritmos basados en densidad como DBScan el cual además de generar los grupos que el algoritmo encuentra necesarios, también es capaz de encontrar el ruido o datos atípicos entre los mismos, con esto la compañía es capaz de generar paquetes y/o publicidad enfocada a las necesidades específicas de cada grupos, como crear planes con llamadas reducida y aumento en los megas o vice versa.

Cada uno de estos grupos contienen diferentes técnicas de las cuales solo se abordan dos de las técnicas más famosas de cada grupo las cuales son:

2.1.6.1 Supervisados

Clasificación: este método consiste en colocar una etiqueta a un nuevo dato, para esto requiere conjuntos de datos ya previamente clasificados y etiquetados por algún algoritmo o los mismos datos, este ayuda a prever en qué grupo pertenece el nuevo dato, de acuerdo a sus atributos, este modelo puede utilizar diversas herramientas como arboles de decisión, K means, redes neuronales etc. (Jovic, 2014).

Regresión: Este busca convertir un conjunto de datos y variables predictor en una única variable independiente comprendida como el resultado, siendo esta numérica, este método permite hacer cálculos del sistema financiero, poder predecir una tendencia en los datos o incluso calcular la probabilidad de un suceso, este método tiene como algoritmos la regresión lineal, polinómica o logística (Jovic, 2014).

2.1.6.2 No supervisados

Reglas de asociación: Esta técnica de aprendizaje no supervisado permite descubrir relaciones entre variables de un conjunto de datos, puede calcular la probabilidad de que un evento ocurra si se desencadena otro. Aunque se escuche muy útil, esta puede tener deficiencias como la correlación espuria donde los datos no tienen nada que ver, aunque así se dicten. Estos algoritmos llegan a ser muy útiles para poder unir dos o más variables distantes de si, son útiles para establecer reglas de mercado y

recomendaciones de productos sin relación aparente. Sus algoritmos pueden ser Apriori o FG-Growth (Jovic, 2014).

Agrupamiento: La última técnica importante del aprendizaje no supervisado, esta organiza los datos en grupos (clusters), donde a de acuerdo a sus atributos pueden o no ser afines entre sí, existen múltiples técnicas para realizar esta tarea, desde algoritmos capaces de identificar el número óptimo para agrupar los datos, o algoritmos que automáticamente dividen los datos en el número esperado de grupos. Estas son muy útiles para obtener información oculta entre los datos e identificar aquellas afinidades ocultas entre los datos. Estos grupos no siempre suelen ser óptimos ya que se desconoce el resultado esperado, por lo cual se requiere de un experto o algoritmos de validación que verifiquen las uniones y viabilidad de estos grupos entre sí. Dentro de estos métodos obtenemos algoritmos como: K Means, DBScan, Agrupamiento jerárquico, etc. (Jovic, 2014).

Para fines objetivos, únicamente se utilizarán algoritmos basados en agrupamiento o clustering, puesto que, dado el problema presentado, estos son aquellos que permiten un desarrollo óptimo de este trabajo.

2.1.7 Algoritmos Descriptivos

Esta es una técnica que busca formar o segmentar los datos en grupos, esta técnica no requiere de una variable de salida, puesto que no busca predecir ninguno de los datos. Estos buscan descubrir información oculta, o explicándolo de otra forma, buscan entender los datos y sus interacciones consigo mismos. Descubriendo patrones, estructuras o relaciones difíciles de identificar a simple vista. Esta poderosa herramienta ayuda a identificar anomalías y a poder simplificar grandes volúmenes de datos para de esta forma, ayudar a tomar decisiones informadas y coherentes (Wegmann, 2021).

2.1.8 Algoritmos Basados En Partición

El agrupamiento de los datos, permiten generar segmentos de los datos más conocidos como clusters, dando por entendido que datos que pertenezcan al mismo clúster compartirán semejanzas o alguna relación cercana dependiendo el algoritmo a utilizar. Estos dividen a los datos en un grupo predefinido de grupos, estos se basan en la elección de centroides y a los datos se le asigna una etiqueta midiendo la distancia entre ellos moviendo los centroides hasta lograr agrupar a todos de una forma matemáticamente correcta como se muestra en la figura 1. Estos algoritmos se

utilizan cuando se espera que los datos tengan divisiones medianamente uniformes o cuando se espera que los datos sean esféricos (Wegmann, 2021).

2.1.8.1 K Means

El algoritmo K Means, o también llamado por el nombre de k-medias, este algoritmo fue creado en 1967 por MacQueen y se convirtió en el algoritmo predilecto para el aprendizaje de la minería de datos, ya que su implementación suele ser simple, pero, eficaz, siendo un algoritmo aún vigente. Este algoritmo es ampliamente utilizado para la minería de datos y el aprendizaje automatizado, permite aplicaciones como la segmentación de datos, análisis de mercado y el procesamiento de imágenes para poder segmentar los colores.

Para poder iniciar con este procedimiento, previamente se debe identificar el número deseado de centroides. Al ser un algoritmo basado en centroides, tiene la facilidad de ser fácilmente verificable y visual. Cada clúster obtiene un centroide que lo identifica y se encuentra en el centro del mismo.

Su procedimiento es el siguiente

1. Identificar el número de grupos *k* deseados, esta identificación puede darse por un experto en los datos a analizar o buscando un número deseado a buscar en los datos.
2. Inicializar los centroides de acuerdo al número de grupos deseados, estos objetos con aquellos que identificaran a los grupos colocándose en su centro.

   $\{\mu_1, \mu_2, \dots, \mu_k\}$

3. Re-Calcular los centroides. Se asigna un clúster a cada objeto de acuerdo a la distancia más corta a cada centroide, para eso el algoritmo más utilizado es la distancia euclidiana, formula mostrada en la ecuación 1. Donde n es el número de dimensiones del espacio, x es el primer punto, así como y la coordenada al segundo punto.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Ecuación 1 Distancia Euclidiana

4. Reasignar los centroides según la media de los objetos, según la media de la distancia, para estos se usan algoritmos como la distancia euclidiana para medir cada punto.

5. Repetir la operación anterior hasta que los centroides se estabilicen, normalmente se establece un número de iteraciones extras para asegurarse que no se están repitiendo.

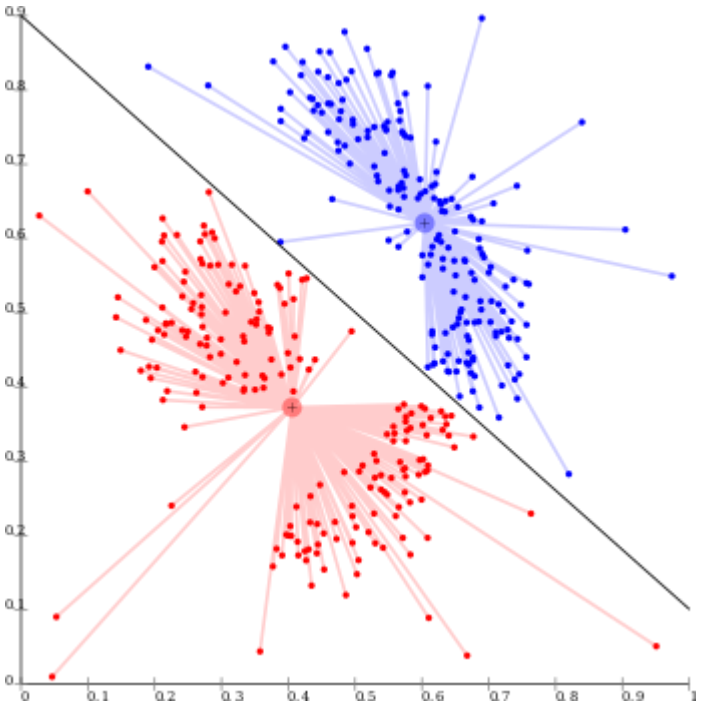

Figura 1 Representación Gráfica del Algoritmo K Means

El número de iteraciones dependen únicamente de los datos. Muchas veces las iteraciones pueden tender al infarto con conjuntos de datos muy unidos, otras veces, con un número pequeño de iteraciones finitas se pueden encontrar los centroides correctos sin cambios. Para impedir las iteraciones infinitas se puede asignar una variable de iteraciones máximas para evitar el bloqueo computacional (Wegmann, 2021).

La complejidad de este algoritmo requiere medir la distancia de cada uno de los n puntos de los k centroides y actualizar cada centroide, hasta la convergencia. Por lo tanto, la complejidad del mismo es de orden de $O(I * K * N * d)$ donde d es la dimensión de los datos. Este algoritmo suele ser muy eficiente y por lo tanto escala con grandes volúmenes de datos dado que sus cálculos son sencillos, así como sus actualizaciones vectoriales (Wegmann, 2021).

Las ventajas de este algoritmo son:

1. Este algoritmo es fácil de implementar puesto que existen muchas herramientas que ofrecen este algoritmo como parte de su repositorio de algoritmos de minería de datos
2. Su costo computacional es bajo por lo que es fácil escalar con grandes conjuntos de datos, pues el costo no escala exponencialmente si no, linealmente.
3. Produce particiones donde se minimiza la varianza interna.
4. Cuenta con variantes dependiendo los objetivos específicos de los datos que mejoran la elección inicial de los centroides.

Las desventajas con:

1. Para poder ejecutarlo se requiere saber el número de clusters a crear, si este número fuera desconocido, el número de grupos finales puede no significar nada realmente.
2. Es sensible a la elección de centroides iniciales, aunque lo más común es escoger los primeros centroides, de escoger otros, los resultados no podrían ser repetibles a grandes conjuntos.
3. La forma de los datos puede influenciar mucho el resultado, pues si los datos tienen formas no esféricas, estos pueden no agruparse correctamente.

2.1.9 Algoritmos Jerárquicos

Estos tienen como objetivo, construir una jerarquía entre los clústeres, esta jerarquía forma un diagrama en forma de árbol (dendrograma), observado en la figura 2, este permite ver de forma visual la interacción de todos los datos con sí mismos. Este algoritmo, al igual que el anterior, permite al usuario seleccionar el número de clusters k antes de su ejecución, pero la forma en que opera dicho algoritmo, permite aun después de la ejecución, poder seleccionar el número de k de interés. Este método se puede basar en dos principios:

Aglomerativo, este une dos clústeres por medio de alguna medida de enlace como pueden ser (Single, Complete, Average, Ward etc.). Esta también conocida como bottom-up, este enfoque parte desde todos los datos con su propio clúster, llamados hojas, conforme los datos van uniéndose mediante el método de enlace, con el objetivo de disminuir cada vez más los clusters, estos van formando clusters más pequeños hasta llegar a un solo clúster conocido como raíz (Rubiños, 2024).

Divisivo, este separa los clusters desde la raíz hasta las hojas del mismo, por medio de una partición interna. Este proceso es básicamente el inverso del método agrometrativo. El primer clúster contiene todos los datos, y busca dividirlo, en datos

cada vez más pequeños por medio del método de enlace. Conforme las iteraciones pueden proceder, estas buscan separar los datos hasta llegar a las hojas, que corresponden a todos los datos crudos y sin tratar (Rubiños, 2024).

Este método es especialmente apreciado en estudios biológicos donde el histograma de unión o división de los datos son muy importantes para el desarrollo de las investigaciones, estas permiten mostrar una división o unión jerárquica de todo el conjunto de datos.

2.1.9.1 Algoritmo De Aglomeración

El algoritmo "Agglomerative Clustering", es parte de los métodos aglomerativos, por lo que su uso se basa en comenzar de las hojas e ir escalando hasta llegar a la raíz. Muy a pesar de la complejidad descrita, este algoritmo requiere información muy general para su ejecución, esta configuración se caracteriza por el uso de Ward como método de enlace, este método minimiza la suma de las diferencias al cuadrado dentro de todos los clusters, es decir, la varianza. Permitiendo así, tener uniones por su punto más cercano (Rubiños, 2024).

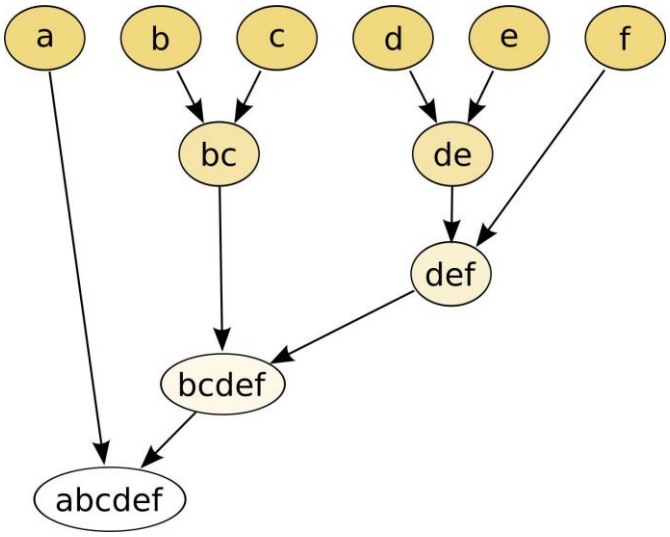

Figura 2 Representación Gráfica del Algoritmo Jerárquico

El procedimiento se basa en un nivel inicial donde $K = 0$ clusters, cada uno es formado por un individuo u objeto, en la siguiente fase se buscarán dos datos que tengan una mayor similitud, la cual se calcula mediante la distancia entre todos con ecuaciones como la distancia euclidiana, esto termina en $n - 1$ clusters. De la misma forma se seguirá iterando en los datos hasta encontrar otros dos nodos con distancias mínimas hasta obtener que cada nivel es igual a $n - L$ grupos formados. Hasta llegar

al nivel $L = n - 1$ donde finalmente todos los datos convergen hasta ser un solo clúster conformado por todos los datos de muestra (Rubiños, 2024).

Este método tiene una característica y es que una vez que dos clústeres se unen para formar uno solo, estos ya contienen a todos los datos anteriormente formados dentro de ellos y ya están formados jerárquicamente para el resto de niveles.

Una vez que este algoritmo ha finalizado todas las operaciones, puede formar el árbol de clasificación, mejor conocido en la materia como dendrograma, este gráfico permite poder seguir la línea de cada dato y observar cada una de sus agrupaciones de forma jerárquica, mostrando cómo se agrupan estos y en qué nivel se unen como se muestra en la figura 2. También permite observar el nivel o valor de la fusión (Rubiños, 2024).

Este a pesar de todas las ventajas gráficas que este método presenta, su nivel de dificultad computacional es del orden $(n\text{^}2)$ o incluso llega a ser de orden $(n\text{^}3)$ en implementaciones muy complejas, esto quiere decir que el costo computacional aumenta en forma exponencial entre más datos puedan existir. Lo cual lo hace muy poco práctico para conjuntos grandes de datos y muy difícilmente escalable sin técnicas de mejora computacional.

Ventajas:

1. Este no requiere tener un número de clúster k fijado con anterioridad, permitiendo decidir la partición luego de construir el árbol.
2. Este algoritmo permite poder utilizar casi cualquier algoritmo para medir distancias entre objetos.
3. Su principal ventaja es la creación del dendrograma, el cual permite observar de manera gráfica las uniones y valores de los datos y cómo forman clúster entre ellos.

Desventajas

1. La principal desventaja de este algoritmo es su alto costo computacional y su difícil escalabilidad para grandes conjuntos de datos, puesto que su orden es exponencial.
2. Este método es irreversible, una vez que dos grupos se unen, no habla una segunda iteración para cambiar de clúster, pues esta decisión es definitiva, por lo cual, se presta para divisiones erróneas y hace que el ruido pueda afectar los enlaces.
3. Aunque es resistente a los algoritmos para medir distancias, los métodos de enlace pueden cambiar drásticamente los resultados del algoritmo,

dependiendo del enfoque del problema, estos pueden dar diferentes resultados que deben ser estudiados con más detenimiento antes de tomar una decisión.

2.1.10 Algoritmos Basados En Densidad

Estos algoritmos buscan formar clusters acorde a la ubicación de los datos, es decir busca por todos los datos aquellas zonas más pobladas y estas las ubica en un clúster, este clúster es bueno para identificar clusters en los datos con formas irregulares, como se muestra en la figura 3, a diferencia de métodos como K Means donde los datos deben tener preferentemente forma esférica, además, este método es capaz de identificar datos atípicos también llamados ruido, estos datos no tienen la característica de pertenecer a ningún clúster, por lo cual no pueden ser agrupados (Han, 2012).

Otra diferencia clara con los demás métodos, es que estos no requieren de conocer cuántos clústeres se requieren, por el contrario, este algoritmo elegirá cuántos grupos pueden o existen en los datos para esas configuraciones.

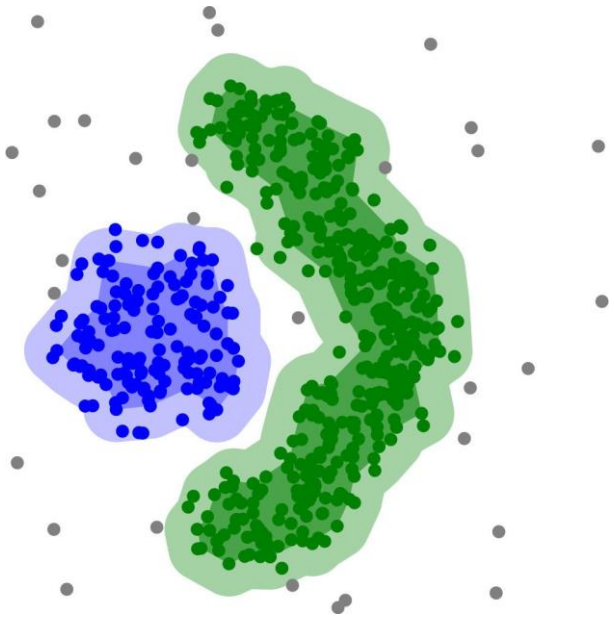

Figura 3 Representación Gráfica del Algoritmo DBScan

Este tipo de algoritmos se utilizan en principios geográficos, para poder agrupar formas arbitrarias en datos especiales, o simplemente para poder detectar el ruido de los datos, esta última utilidad del algoritmo es especialmente útil cuando se requiere detectar los comportamientos irregulares en los datos y trabajar únicamente sobre

ellos, como, por ejemplo: compras irregulares en clientes, datos muy alejados de un uso normal, comportamiento de microorganismos no predecibles, etc. (Han, 2012).

2.1.10.1 DBScan

Este algoritmo es el principal exponente de los algoritmos basados en densidad, este define clusters a partir de las regiones densamente pobladas en el espacio de los datos. Como parámetros esenciales para este algoritmo, se tienen dos:

ε (Radio de vecindad): este parámetro, permite saber cuál es la distancia mínima entre los puntos vecinos para poder identificarlos como un grupo o como parte de un mismo clúster. Este algoritmo se basa en medir una distancia por medio de alguna ecuación como la euclidiana ya mencionada. Y una vez que se miden las distancia entre los puntos se asigna clusters a aquellos que están juntos (Rubiños, 2024).

MinPts (Mínimo de puntos): Este parámetro permite saber que grupo es considerado como clúster o solo como ruido, una vez que las distancias han sido medidas y se identifican grupos dentro de los datos, este parámetro ayuda al algoritmo a saber si es un grupo útil o si solo es ruido para el algoritmo.

Este algoritmo puede ser configurado con más parámetros, pero únicamente con estos dos, puede dar resultados extraordinarios para los campos donde su implementación es requerida. Este algoritmo fue creado especialmente para tratar datos geográficos o imágenes satelitales, donde es comúnmente usado. También se utiliza por su capacidad única para identificar el ruido y anomalías, así como identificar patrones especiales e identificación de clusters para datos no uniformes (Rubiños, 2024).

La complejidad de este algoritmo puede ser desde el orden $O\ (n\ log\ n)$ hasta, $O(n^2)$ todo depende de cómo es que se accede a los datos, si se utiliza algún índice eficiente para iterar entre los datos, se obtiene una complejidad baja, aunque sin el uso de índices eficientes para acceder a los datos, estos dan un orden exponencial lo cual lo hace muy difícil para escalar en grandes conjuntos de datos. Este método dentro de complejidad toma un lugar neutro comparándolo con k means y el jerárquico (Rubiños, 2024).

Ventajas:

1. Este identifica automáticamente la cantidad de clusters sin necesidad de indicarlo como parámetro.
2. Puede identificar datos del ruido de los mismos datos, este, permite identificar datos atípicos en los conjuntos lo cual permite un estudio más profundo para

identificar estos datos y sus comportamientos alejados de los conjuntos grandes.

3. La simplicidad de sus parámetros que permite una sencilla ejecución, puesto que estos son muy claros y fáciles de entender aun para personas no matemáticas o expertas en el tema.

Desventajas:

1. Encontrar una configuración de ambos parámetros, muchas veces es difícil, requiere de muchas iteraciones y pruebas para encontrar una respuesta realmente convincente. Una mala configuración muchas veces culmina en agrupar todos los datos en un mismo clúster, todos los datos convertidos en ruido o un mega clúster con muchos clusters apenas contenidos en el mínimo de puntos.
2. Este algoritmo es sensible a la forma de los datos, si los datos están muy juntos o muy separados puede que la configuración de los parámetros termine siendo una amalgama de un solo clúster o solo ruido.
3. Cuando se habla de datos con múltiples atributos, se vuelve muy difícil enfocarlo como un sistema de densidad puesto que esto ya no podría tener sentido
4. Su complejidad permite poder ejecutar conjuntos de datos medianos sin la necesidad de ajustes, para conjuntos de datos grandes, este algoritmo puede requerir utilizar índices eficientes para el manejo de las distancias.

2.1.11 Técnicas de Evaluación o Validación

Los índices de validación de un clúster son algoritmos que evalúan la unión y eficiencia de los datos entre sí de cada grupo. En la minería de datos no supervisada esto funciona para poder medir la distancia de cada elemento con otros clusters y entre cada clúster de cada uno de los grupos formados (Wegmann, 2021).

Para estas validaciones existen dos clasificaciones: la validación externa y la validación interna, como el nombre lo indica, la validación externa requiere de datos previamente etiquetados para validar la cohesión de los nuevos clusters creados. Los internos buscan la cohesión dentro de los mismos datos ya etiquetados, estos son los más comunes cuando se está buscando el algoritmo correcto a utilizar (Wegmann, 2021).

Estos son mayor mente utilizados para poder identificar la eficiencia de algoritmos de clustering, tales como k means, agrometrativo o DBScan y comparar los resultados. Su beneficio radica en que, al desconocer cómo es que los datos se comportan o que,

al no haber un estudio previo de los datos, se beben iterar a través de múltiples algoritmos para poder escoger el algoritmo y los parámetros correctos para el trabajo. Un buen resultado en los índices de validación dirá cuál fue la propuesta óptima para: Algoritmo, parámetros, número de clusters, algoritmos para medir distancias, enlaces, variación, (según corresponda) entre otros datos.

2.1.11.1 Índice De Calinski Harabasz

Este índice de validación también es conocido como razón de varianza, este mide que tan densos y separados entre sí están los clusters, este índice define la razón de la dispersión inter clusters e intra clusters.

Este algoritmo se caracteriza por ciertos aspectos, como ejemplo, valores grandes, significan mejores resultados. Este índice baja cuando existe dispersión entre clusters y aumenta cuando la separación entre clusters es mayor. Su uso mayormente se basa en poder encontrar un número óptimo de clusters, normalmente se grafican los resultados con números diferentes de clusters hasta encontrar la mayor altura de los resultados obtenidos (Fayyad, 1996).

La fórmula de este índice se basa en k como número de clusters en un conjunto de datos D, en esta fórmula nk y ck son el número de puntos y centroide del grupo k, c es el centroide global, N es el número total de puntos de datos esta fórmula se muestra en la ecuación 2.

$$CH = \left[\frac{\sum_{k=1}^{k} n_k ||c_k - c||^2}{K-1}\right] / \left[\frac{\sum_{k=1}^{k}\sum_{k=1}^{n_k} ||d_i - c_k||^2}{N-K}\right]$$

Ecuación 2 Índice De Calinski Harabasz

En pocas palabras, este algoritmo busca encontrar la separación a partir de la suma de las distancias entre el centroide de cada grupo y el centroide global del conjunto de datos. Su coste computacional dice que este algoritmo es fácilmente eficiente ya que pertenece a la orden $O(N)$, esto es gracias a que únicamente suma distancias al centroide global y locales, lo cual es bastante apto para conjuntos de datos grandes (Fayyad, 1996).

Este algoritmo suele caracterizarse por ser fácil de calcular además de ser muy eficiente, su coste es lineal orientado al número de datos. Es independiente del algoritmo con el cual se hizo la agrupación y no requiere ningún dato adicional al conjunto ya etiquetado. Aunque, hay que considerar que este índice depende se compararse con otros resultados ya obtenidos, si solo se tiene un conjunto, este índice no puede ser tomado como punto de referencia para obtener una conclusión.

Este también puede verse afectado con el aumento de clusters, por lo que no es recomendable para datos muy irregulares o muy densos, pues la curva puede solo ir en aumento y afectar los resultados e interpretaciones (Fayyad, 1996).

2.1.11.2 Índice Davies Bouldin

Este índice basa su funcionamiento en la similitud media entre clusters, en términos de su radio interno relativo a la separación entre ellos. Esto significa que, en promedio, cada clúster tiene la menor proporción posible frente a la distancia del vecino más cercano. Este índice considera mejor partición aquella con un índice bajo sobre las demás.

$$DB = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Ecuación 3 Índice Davies Bouldin

En esta fórmula N es el número de grupos, σi es la dispersión promedio dentro del grupo i, medida, mediante la distancia media entre los puntos del grupo y su centroide, ci es el centroide del grupo i y, por último, $d\ (ci, cj)$ es la distancia entre los centroides de los grupos (Zwass, 2025). Formula representada en la ecuación 3

Este número indica qué tan densamente formados están los clusters y que tan separados de los demás clusters están. Estos dos factores son fundamentales para Davies Bouldin para indicar que los grupos son óptimos. Por eso es que un número alto en este algoritmo indica que no existe calidad de ellos.

Este algoritmo es fácil de interpretar pues toda su fórmula se basa en medir la unión de los clusters y la separación entre ellos. Esta fórmula no requiere datos previamente etiquetados ni predefinir algún tipo de parámetro antes de ser ejecutado. Al ser un algoritmo lineal donde únicamente se miden distancias, este tiene un coste computacional bajo de orden $o(n)$ donde se demuestra su bajo costo. Pese a esto este índice puede tener problemas al medir clusters esféricos homogéneos, si se tienen clústeres muy dispersos, estos darán una medición errónea, alta. También se ha demostrado que tiene preferencias por clusters compactos y es sensible al ruido o clusters con tamaños dispersos (Zwass, 2025).

2.1.11.3 Índice Silhouette

Este índice funciona asignando un valor a cada dato evaluando que tan cercano es un objeto de su clúster y de un clúster lejano. Su objetivo es cuantificar qué tan bien definidos y separados están los grupos en función de la distancia entre los objetos de

datos. Para mostrar su eficiencia únicamente se cuenta con resultados de -1 a 1, en el cual, un número grande indica una mejor relación entre ellos (Fayyad, 1996).

$$Silhouette(i) = \frac{b(i) - a(i)}{max\{\, a(i), b(i)\}}$$

Ecuación 4 Índice Silhouette

Su fórmula se calcula para cada objeto de datos i, donde, $a(i)$ es la distancia promedio entre el objeto i y los demás objetos del mismo clúster. $b(i)$ es la distancia promedio entre el objeto de datos i y los objetos en el clúster más cercano y diferente al suyo. La fórmula es mostrada mediante la Ecuación 4. Este también tiene la característica de identificar no solo un agrupamiento deficiente, si no, que, si se obtiene un valor debajo de 0, indica que el agrupamiento podría estar mal realizado ya que los grupos están superpuestos (Fayyad, 1996).

Este método es muy interpretativo el cual permite dibujar gráficos des Silhouette para cada clúster y evaluar la solidez de la partición, este no depende del algoritmo de clustering utilizado y su escala de [-1,1] facilita comparar soluciones de distinto k, este método baja su puntaje equitativamente con la cohesión e inter separación, lo que ayuda a identificar un k óptimo. Y aunque este método es muy útil para identificar la eficiencia de los clusters, su mayor problema es el coste computacional ya que pertenece a la orden $o(N\text{^}2)$, lo que es difícil de escalar para grandes volúmenes de datos (Piatetsky-Shapiro, 2012).

CAPÍTULO 3 Marco Tecnológico

Para hacer uso de estas técnicas se tiene software dedicado a realizar estas tareas, algunos ejemplos son Python, KNIME y Weka. Estas herramientas suelen ser complejas y requerir un conocimiento previo de todas las variables y características de los algoritmos. Por lo tanto, no están al alcance de todos los usuarios que podrían beneficiarse de la Minería de Datos. La elección de la herramienta adecuada depende de la tarea específica y las necesidades del usuario, lo que subraya la importancia de desarrollar una herramienta integral que simplifique el proceso de Minería de Datos para usuarios no expertos (Wegmann, 2021).

Existen diversas plataformas y herramientas similares que ofrecen capacidades de Minería de Datos. Además de Python, que es altamente versátil y ampliamente utilizado en este ámbito, existen alternativas como MATLAB, conocido por su potencia en cálculos numéricos y procesamiento de datos. Weka es otra opción popular, una plataforma de código abierto con una interfaz gráfica intuitiva para la Minería de Datos. R, por otro lado, es un lenguaje de programación y entorno estadístico robusto, ampliamente preferido por los estadísticos y analistas de datos. Cada una de estas plataformas tiene sus propias ventajas y características, y la elección dependerá de los requisitos y preferencias del usuario (Jovic, 2014).

3.1 Python y el Ecosistema scikit-learn

Python Es un lenguaje de programación de alto nivel, orientado a objetos y de propósito general. Su valor en la minería de datos radica en sus librerías y no en el lenguaje mismo; librerías como NumPy, SciPy y Pandas permiten que el lenguaje pueda ser usado en campos de minería de datos de una forma simple, siempre y cuando se tenga conceptos de programación entendidos (Pedregosa, 2011).

La principal biblio teca para la realización de algoritmos de minería de datos es scikit-learn, la cual, integra una amplia gama de algoritmos de última generación para tareas tanto supervisadas como no supervisadas. El diseño de esta librería sumado a su amplia documentación y rendimiento, la hacen ideal para desarrollos rápidos y eficientes con un gran escalado (Pedregosa, 2011).

3.1.1 Algoritmos

Como se mencionó anteriormente, este lenguaje y scikit-learn, pueden ejecutar una amplia cantidad de algoritmos no supervisados, incluyendo los necesarios para la

realización de esta investigación, como lo son: K Means, Jerárquico y DBScan. También destaca por su gran capacidad para modificar los parámetros de configuración de cada algoritmo. Pudiendo controlar aspectos como las métricas de las distancias, métodos de enlace, procesamiento procedural, parámetros especiales de cada algoritmo, etc. (Pedregosa, 2011).

Dentro de los índices de validación también tiene un amplio repertorio, simplificando la ejecución de estos, puesto que se sigue la misma curva de aprendizaje que los algoritmos.

3.1.2 Ventajas y Desventajas

Ventajas:

- Versatilidad. Este lenguaje es muy versátil gracias a sus librerías permitiendo generar aplicaciones altamente complejas y con muchas funcionalidades sin la necesidad de implementar otro tipo de tecnologías o lenguajes.
- Ecosistema unificado. Python permite una aplica ejecución de los algoritmos necesarios sin la necesidad de buscar librerías fragmentadas, la curva de aprendizaje de scikit-learn es muy reducida, una vez que se ejecuta cualquier algoritmo por primera vez.
- Listo para producción. Gracias a su soporte para programación orientada a objetos, Python permite código, mantenible, simple y de grado de producción.

Desventajas:

- Requiere saber programación. Utilizar Python como herramienta para minería de datos sin tener con0cimientos solidos sobre programación puede ser muy complicado. Esto representa una barrera de entrada para usuarios nuevos y no expertos en la programación debido a su falta de GUI
- Selección de parámetros. A pesar de contar con índices de validación, configurar métricas de algoritmos como K means, que requieren ya conocer el número de clusters a generar, puede ser una tarea difícil si no se conocen previamente los datos.
- Gestión de memoria. El análisis de Big Data es un problema debido a su alto coste computacional como lenguaje de alto nivel. Utilizar esta herramienta para tratar grandes volúmenes de datos, puede ser catastrófico si estos superan a la memoria en algoritmos de alta complejidad como lo son el método Jerárquico, ocasionando un desbordamiento de memoria y en la caída del sistema.

3.2 R: Lenguaje Estadístico

R es un lenguaje y entorno de código abierto, diseñado para la computación estadística y visualización de datos. Esta en una herramienta orientada a los estadísticos, investigadores y académicos para la exploración y experimentación de datos (Charrad, 2014).

Sus casos de uso son mayormente académicos en investigaciones de vanguardia como la bioinformática, la cual usa datos de alta dimensión. También se usa en estudios clínicos con grandes conjuntos de datos (Charrad, 2014).

3.2.1 Algoritmos

Los algoritmos que maneja este lenguaje tienen un soporte amplio con implementaciones especializadas y estadísticamente rigurosas.

K Means y DBScan son dos algoritmos ampliamente utilizados y ejecutados en este lenguaje debido a su implementación especializada (Charrad, 2014). El método Jerárquico en R es particularmente fuerte gracias a su librería fastcluster que proporciona una compilación en C++ de cada uno de los métodos de enlace, este método a menudo supera el rendimiento a aplicaciones como Python y MATLAB.

R proporciona un conjunto de paquetes de validación con herramientas complejas y de alta calidad. Desde paquetes como NBClust, que proporciona algoritmos con la capacidad de identificar el numero optimo de clusters (Charrad, 2014). Hasta en paquete de índices de validación aun mayor a scikit-learn en Python. Su característica principal de NBClust es poder ejecutar distintos índices y métricas de validación simultáneamente a los algoritmos, proporcionando un esquema de agrupamiento optimo basado en la votación mayoritaria de los índices.

3.2.2 Ventajas y Desventajas

Ventajas:

- Rigor estadístico. Esta herramienta es predilecta para el aprendizaje estadístico profundo y la exploración.
- Validación Masiva. Proporciona múltiples herramientas y esquemas de validación de clusters, especialmente con el paquete NbClust, el cual permite ejecutar cerca de 30 índices.

- Visualización. R es considerado superior debido a poder crear visualizaciones de datos complejas y de calidad publicación gracias a sus paquetes especializados.

Desventajas:

- Fragmentación. El ecosistema de R radica en sus múltiples paquetes pequeños y altamente especializados, lo cual dificulta la curva de aprendizaje, por sus API's inconsistentes en comparación con scikit-learn en Python, el cual, presenta un ecosistema unificado.
- Curva de aprendizaje. Este entorno requiere una mentalidad estadística y es difícil de operar aun en las personas especializadas o familiarizadas con la programación.
- Integración con producción. Al no ser un lenguaje de propicito general, integrar un módulo de R en sistemas de producción o aplicaciones web resulta más complejo.

3.3 Weka (Waikato Environment for Knowledge Analysis)

Weka es una herramienta considerada workbench, ya que este software implementa una colección de algoritmos de aprendizaje automático y herramientas de procesamiento previamente cargadas y listas a utilizar con una GUI (Hall, 2009).

Desarrollado en la universidad de Waikato, este sistema está escrito únicamente en java lo que garantiza su portabilidad (Hall, 2009).

Su principal modo de interacción con el usuario es a través de una GUI que permite la carga de archivos en formato csv o ARFF y aplicar los algoritmos visualmente sin la necesidad de escribir código.

3.3.1 Algoritmos

Weka ofrece un amplio catálogo de algoritmos de agrupamiento estándar a través de su pestaña clúster en el explorador. Dentro de estos se encuentran algoritmos como SimpleKMeans, el cual es el K Means estándar, DBScan y Jerárquico con sus nombres más comunes. Proporcionando de estos resultados estadísticos y gráficos sobre cada algoritmo (Hall, 2009).

Este algoritmo cuenta con distintos algoritmos de validación como el método de codo, el cual ayuda a identificar el número optimo de clusters para los datos, pero tiene un punto débil, el cual es que carece de índices de validación como Silhouette, Calinski-Harabasz o Davies-Bouldin, siendo que, para utilizar estos índices, se requieres buscar otras herramientas especializadas en esto (Hall, 2009).

3.3.2 Ventajas y desventajas

Ventajas:

- Facilidad de uso. Su principal característica es que cuenta con una interfaz intuitiva, fácil de usar para aplicados en materia de la minería de datos y no programadora a la hora de cargar datos y ejecutar algoritmos complejos.
- Herramienta educativa. Esta herramienta es muy útil para la enseñanza de la minería de datos a estudiantes en entornos académicos.
- Integral. Proporciona un flujo de trabajo completo desde el procesamiento de datos hasta la visualización en una sola aplicación.

Desventajas:

- Validación de agrupamiento. Esta herramienta no ofrece índices de validación de clusters, las métricas que proporciona son validación por el método de codo para encontrar números óptimos de grupos o el uso de clases a clústeres que depende de tener datos etiquetados.
- Escalabilidad. Weka no está diseñado para el manejo de Big Data, por lo cual, es mejor con conjuntos pequeños de datos que no superen la memoria.
- Flexibilidad: Esta herramienta está limitada a los algoritmos previamente instalados e implementados, puesto que no es tan flexible como las herramientas basadas en lenguajes de programación.

3.4 WebMinerX

WebMinerX es una aplicación web desarrollada con el objetivo explícito de reducir la barrera de entrada a la minería de datos para usuarios no expertos. La principal diferencia con las demás aplicaciones, es su entorno accesible desde un navegador web y una interfaz sencilla e intuitiva en todo momento (Guzmán, 2024).

Para su desarrollo se utilizaron múltiples tecnologías y enfoques. Su interfaz esta desarrollada sobre HTML, CSS y JavaScript. Mientras que, el Backend está hecho sobre Python con scikit-learn y FastApi para la conexión por medio de protocolos https,

utilizando también la librería Matplotlib para poder mostrar gráficos e ilustraciones acordes a los resultados obtenidos por los modelos (Guzmán, 2024).

Este sistema está hecho para guiar al usuario a través de una carga de datos en formatos CSV y ARFF, selección de algoritmo y la visualización interactiva de resultados.

3.4.1 Algoritmos

A diferencia de las anteriores, esta aplicación está fuertemente intencionada para la ejecución de únicamente 3 algoritmos de aprendizaje no supervisado los cuales son: K Means, Agrupamiento Jerárquico y DBScan, los algoritmos propuestos para esta investigación (Guzmán, 2024).

Una característica importante para esta aplicación en diferencia con Weka, es su integración nativa con los algoritmos de agrupamiento junto con índice de validación internos, como lo son: Silhouette, Calinski-Harabasz y Davies-Bouldin. Esta herramienta está diseñada para presentar estos índices de una manera comprensible permitiendo a un usuario no experto tomar una sección informada sobre la calidad de cada una de sus agrupaciones hechas (Guzmán, 2024).

3.4.2 Ventajas y Desventajas

Ventajas:

- Accesibilidad y simplicidad. Su objetivo es simplificar el proceso de Minería de datos para usuarios sin conocimientos extensos de programación o estadística.
- Flujo de trabajo de validación integrado: Su ventaja más grande es la inclusión de los índices de validación en una interfaz intuitiva y fácil de manipular, esto permite un verdadero análisis no supervisado.
- Interfaz de usuario intuitiva. Al ser una aplicación web guiada, simplifica la carga de datos, configuración de parámetros y la interpretación de resultados. Las pruebas de usabilidad realizadas con 20 usuarios no expertos arrojaron una puntuación de 4.43/5, confirmando que el sistema es fácil de usar, intuitivo y no requiere asistencia técnica.

Desventajas:

- Alcance limitado. El sistema está actualmente limitado en cuanto a los algoritmos e índices que ofrece, siendo una limitante para usuarios con diferentes necesidades sobre sus datos específicos.

- Falta de persistencia. Esta aplicación carece de una integración con bases de datos, dando como resultado una capacidad nula para almacenar o recuperar sus conjuntos.
- Escalabilidad. Esta herramienta está diseñada para la usabilidad, lo cual afecta al rendimiento, siendo no útil en la Big Data, carente de optimizaciones y protección en el sobre desbordamiento.

3.5 Elección de la herramienta para el procesamiento de datos.

En un análisis exhaustivo acerca de las ventajas y desventajas que ofrece cada herramienta y una recopilación de sus características principales, se logró recabar la tabla 1. La cual demuestra principales problemas y virtudes de cada tecnología.

Tabla 1 Evaluación de herramientas de Minería de Datos

Característica	Python (scikit-learn)	R	Weka (Waikato Environment)	WebMinerX
Plataforma	Lenguaje de Programación	Lenguaje de Programación	Aplicación de Escritorio (Java GUI)	Aplicación Web (Python Backend)
Usuario Objetivo	Expertos, Desarrolladores, Científicos de Datos	Estadísticos, Académicos, Investigadores	Estudiantes, Principiantes, Académicos	Usuarios No Expertos, Analistas
Algoritmos de Agrupamiento	Muy Extenso (K-Means, DBSCAN, Jerárquico, Spectral, etc.)	Muy Extenso (K-Means, DBSCAN, Jerárquico optimizado)	Básico (K-Means, EM, DBSCAN, Jerárquico)	Básico (K-Means, DBSCAN, Jerárquico)
Índices de Validación Internos	Silhouette, Calinski-Harabasz, Davies-Bouldin (scikit-learn)	Silhouette, Calinski-Harabasz, Davies-Bouldin (NbClust, clusterCrit)	No	Silhouette, Calinski-Harabasz, Davies-Bouldin

En base a la tabla 1, se concluyó que, la mejor herramienta para este caso de estudio especifico es WebMinerX, su versatilidad y la capacidad de integrar los índices de validación junto con los algoritmos, sin la necesidad de escribir código, lo hace la herramienta ideal para este proyecto permitiendo así un veloz desarrollo e implementación con la investigación.

3.6 United States Geological Survey (USGS)

El USGS es una de las agencias científicas del gobierno de Estados Unidos dedicada al estudio de la tierra. Esta agencia funciona como un departamento de investigación oficial que proporciona datos y conocimientos sobre los recursos naturales, sus peligros y la salud de los ecosistemas (U.S. Geological Survey, n.d.).

El objetivo de esta agencia es la de monitorear, analizar y predecir las interacciones del sistema terrestre para poder entregar información útil y oportuna. Esta información debe ser útil para describir y comprender la tierra, poder predecir fenómenos y desastres, minimizar cualquier tipo de muerte humana o biológica, gestionar recursos naturales como cuerpos de agua, flora, fauna, minerales, y hablando de una forma más general, proteger la calidad de vida (U.S. Geological Survey, n.d.).

La USGS busca liderar la investigación, las evaluaciones y predicción de los recursos y procesos naturales para así poder satisfacer necesidades que la humanidad pueda tener en el ahora.

Esta agencia se caracteriza principalmente por tener imparcialidad científica, destinada únicamente a la investigación, sus tareas no deben ser de regulación de ningún tipo. Su principal función es la de proveer información fiable, objetiva y sin ningún tipo de inclinación ya sea, política, social y moral (U.S. Geological Survey, n.d.).

No se ubica solo en el área geológica, puesto que, su área de estudio es multidisciplinar, abarcando áreas de estudio muy amplias como lo son la biología, geología. hidrología y por supuesto, geografía.

La USGS es responsable de monitorear fenómenos continuos como el volumen de los cuerpos de agua, actividad sísmica y los cambios en la superficie terrestre. Esto lo logra gracias a sus imágenes satelitales de acceso público y gratuito. Además de complejos mapas topográficos (U.S. Geological Survey, n.d.).

CAPÍTULO 4 Marco Metodológico

En esta sección se explorarán las metodologías orientadas al descubrimiento de información mediante la minería de datos, estas metodologías proveen disantos puntos de vista para la identificación de información oculta en un conjunto de información verídica y fiable. Se explorarán conceptos clave para su análisis y estudio como lo son sus características generales, las ventajas, desventajas y fases de cada uno de los procesos que las distintas metodologías conlleven.

4.1 Metodología KDD (Knowledge Discovery in Databases)

La metodología KDD es un proceso basado en iteraciones con la finalidad de descubrir conocimiento a través de grandes conjuntos de datos. Se define como un proceso no trivial para poder identificar patrones válidos, novedosos, buscando que los datos sean útiles y comprensibles. Descritas de una forma general. Estas actividades incluyen: selección de datos, preprocesamiento y limpieza, transformación de las variables, el uso de la minería de datos con algoritmos de aprendizaje y la evaluación e interpretación de los datos y/o patrones descubiertos. Normalmente, estos nuevos patrones requieren de alguna validación externa para poder afirmar que sean correctos y útiles. Esta metodología busca convertir datos grandes en conocimiento mediante el uso de la minería de datos para así poder descubrir y encontrar patrones en los datos (Piatetsky-Shapiro, 2012).

El concepto Knowledge Discovery in Databases (KDD) nació en 1989 durante el primer taller de descubrimiento de conocimiento y su nombre fue otorgado por Gregory Piatetsky-Shapiro, a partir de esto, cada año se realizaban conferencias internacionales de KDD hasta 1996 donde Fayyad y otros, lo consolidaron como un proceso no trivial de extracción de patronal útiles en los datos. Y desde entonces se consolidó como una metodología de minería de datos académica de referencia (Piatetsky-Shapiro, 2012).

Esta metodología es útil para cualquier conjunto grande de datos donde se requiera obtener conocimiento o identificar patrones para su estudio, su mayor campo de utilidad lo es el marketing y el comercio en la segmentación de clientes y análisis de tendencias en los productos, aunque las aplicaciones para esta metodología son muy amplias extendiéndose a ramos como: las finanzas para la identificación de fraude en la aprobación de créditos, también, en el área de salud donde ayuda con la detección

predictiva de enfermedades y/o el avance de las mismas. En la seguridad informática para identificar el uso anormal de las redes o tecnologías e identificar peligros potenciales. Y estos solo son algunos ejemplos de las utilidades de esta metodología. Todas estas tareas están orientadas a la minería de datos en materia de clasificación, clustering y reglas de asociación, a través de las cuales, todo este tipo de conocimiento puede ser adquirido de la mano del analista (Fayyad, 1996).

4.1.1 Ventajas

El uso de esta metodología permite obtener conocimiento nuevo, descubrir patrones, tendencias y/o relaciones que serían muy difíciles de identificar manualmente en las bases de datos grandes. Esto permite poder encontrar información útil, novedosa y concisa sobre los datos.

Esta herramienta también permite automatizar tareas complejas de exploración de los datos, el objetivo de KDD es automatizar al máximo el proceso de descubrimiento de información, lo cual mejora el procedimiento de manejar volúmenes masivos de datos.

Una de sus grandes ventajas es la de su gran capacidad para adaptarse a cualquier disciplina, ya que combina enfoques de diferentes materias para ser aplicada en enfoques diversos. Lo cual le da una gran versatilidad a la forma de enfrentar problemas de diferentes ramas. Además, la característica de la metodología de ser iterativa, permite al analista refinar las tareas de la minería de datos, conforme se van haciendo descubrimientos nuevos.

Por último, esta metodología permite evaluar la calidad de los datos ya descubiertos. Lo que aumenta el grado de validez y comprensibilidad de la información obtenida. Se pueden utilizar diferentes métricas para evaluar la calidad de los resultados además de permitir la colaboración humana para la revisión de estos mismos. Esto da como resultado información útil, eficaz y novedosa.

4.1.2 Desventajas

La principal desventaja es su alto coste computacional, cuando se manejan grandes conjuntos de datos no solo el tiempo de procesamiento es un factor, puesto que, muchas veces, sin las optimizaciones correctas, los algoritmos pueden terminar con un desbordamiento de memoria o un sobreuso de la CPU, lo que termina con equipo de cómputo en colapso. Para esta tarea (Dependiendo del volumen de datos) Se requieren equipos robustos, y algunas veces, costosos lo que implica un alto costo de

tiempo y recursos para su ejecución. Además, todo esto requiere ir de la mano de expertos para que puedan supervisar, planificar y evaluar el proceso.

Los resultados de esta herramienta dependen altamente de la calidad de los datos, ya que, si los datos no vinieran de una fuente confiable o existiera ruido entre ellos, pueden resucitar en información deficiente o errónea. Por ello, la realización de procesos de preparación y limpieza de los datos es fundamental para la realización del proceso.

4.1.3 Fases del Proceso KDD

Selección de Datos:

En la primera fase del proceso de la metodología KDD consiste en revisar, validar y obtener información consistente y robusta sobre el problema a analizar. El objetivo de esta base es obtener información confiable e íntegra, estas fuentes de información pueden ser bases de datos, almacenes de datos, estadísticas, o métricas almacenadas (Fayyad, 1996).

Para esto se debe establecer las delimitaciones de la investigación y delimitar la información de acuerdo al problema, pueden delimitarse por uno o más tipos como, zona geográfica, fecha, delimitación de un espacio, población etc. Esto con la finalidad de no ingresar información irrelevante que pudiera generar ruido en la investigación (Piatetsky-Shapiro, 2012).

Preprocesamiento:

Esta fase busca reducir la mayor cantidad de ruido de los datos, eliminando o rellenando datos faltantes (dependiendo el problema se pueden llenar huecos usando medias, medianas o modas en los datos, aunque no es recomendable). Dentro de este procedimiento también se deben integrar los datos de los distintos orígenes de los datos creando relaciones entre los datos y de ser posible, normalizar los datos para un correcto procesamiento con la mayor optimización posible en los datos (Piatetsky-Shapiro, 2012).

Esta fase es fundamental para poder procesar los datos de una forma eficiente y muchas veces, si esto se hace bien, aumenta la calidad y disminuye el tiempo de procesamiento.

Transformación:

Esta es la última fase de la preparación de los datos, esta fase consiste en reducir la dimensionalidad de los datos y crear nuevas variables a partir de esto, se emplean

técnicas de reducción de dimensionalidad las cuales buscan reducir los datos hasta crear conjuntos convenientes y específicos, los cuales únicamente tienen información concreta y relevante para las pruebas (Piatetsky-Shapiro, 2012).

Su objetivo es obtener un conjunto de datos capaz de ser procesado con el mínimo de recursos posibles, pero sin perder información fundamental para el mismo sistema, este conjunto también debe ser capaz de poder hacer operaciones básicas de conjuntos sin tener algún tipo de problema de datos, las operaciones que esté conjunto deben permitir pueden ser, sumas, promedios, modas, análisis estadísticos etc. Obteniendo así un conjunto homogéneo, claro y simplificado de datos (Piatetsky-Shapiro, 2012).

Minería de datos:

de acuerdo al problema y a la solución que se desea buscar, en esta fase se debe aplicar alguno de los algoritmos ya antes descritos, se debe elegir en técnicas de aprendizaje automático sobre los datos ya transformados. Algunos de los ejemplos de técnicas aplicables en este paso pueden ser: clasificación, regresión, clustering, reglas de asociación, etc. Este paso puede permitir manejar los datos desde, usar k means o DBScan para poder identificar grupos ocultos en los datos, hasta, entrenar modelos de redes neuronales o árboles de decisión para predecir una etiqueta (Jovic, 2014).

Esta fase debe dar como resultado el poder obtener conjuntos de patrones, etiquetas, reglas, modelos, variables predictivas etc. Para poder ser analizadas más adelante, este paso puede no realizarse una única vez, y, por el contrario, se deben aplicar numerosas iteraciones de los datos, para así, poder conseguir un resultado que pueda satisfacer el problema o encontrar la solución.

Evaluación e Interpretación:

En esta fase es donde finalmente se obtiene información relevante, aunque primero se debe evaluar la calidad de los resultados obtenidos en el paso anterior, es esta fase se aplican procedimiento como índices de validación de los datos, porcentajes de asertividad, usar métricas de exactitud y confianza de las variables obtenida, todo lo la finalidad de poder decidir interpretar los datos obtenidos o si se requerirán nuevas iteraciones sobre los datos (Jovic, 2014).

Una vez que la integridad y calidad de los datos es la correcta y permitida, un experto en los datos debe extraer el conocimiento de los resultados, se analizan patrones, estructuras nuevas, asociaciones encontradas y se conjunta todo es nuevo conocimiento como conocimiento válido. Este conocimiento se da a conocer al

usuario y se puede permitir ajustar para futuras iteraciones del procedimiento KDD o retroalimentar los algoritmos con esta información descubierta (Jovic, 2014).

4.2 Metodología CRISP-DM (Cross-Industry Standard Process for Data Mining)

La metodología CRISP-DM se traduce como el proceso estándar inter-industrial para la minería de datos. Definida en la literatura como un modelo de estándar abierto y no propietario para guiar los procesos de minería de datos. Su principal característica es su neutralidad, ya que, fue diseñado para ser independiente de la industria, software y de la aplicación especifica. Es decir, no depende de las herramientas y objetivos para si implementación (Saltz, 2021).

Este modelo tuvo lugar en el año 1996 y publicado de manera formal en el 2000, su construcción estuvo a cargo de un grupo de entendidos en materia de la minería de datos, una de las personas que encabezaban este proyecto era Daimler Crysler, quien ya aplicaba la minería de datos en procesos de negocio (Saltz, 2021). Durante esta apoca la minería de datos era inmadura y muchas veces, su éxito dependía de la experiencia y habilidades del usuario, puesto que no existían formulas fiables para poder ser repetidas o duplicadas.

Desde el comienzo este proyecto buscó demostrar que la minería de datos era ya lo suficientemente avanzada para ser desarrollada como una ciencia formal y estaba lista para ser incluida en procesos de negocio. Y de esta forma, crear una partición ordenada del complejo procesos de minería de datos y se fácilmente reproducible sin depender del contexto.

4.2.1 Ventajas

Su principal fuerte desde un comienzo fue el estatus de “estándar” y “probado por la industria”, el cual creo estructuras familiares y una estructura fácilmente reconocible por los entendidos de la materia y los nuevos usuarios (Saltz, 2021).

Esto, sumado a que no es dependiente del contexto como la herramienta utilizada, campo o aplicación, la hizo fuertemente aceptada dentro de la comunidad. Esta naturalidad lo convirtió en un marco perfecto adaptado a cualquier situación y diferentes contextos, lo cual le añade tanto estructura como flexibilidad lo cual es esencial en un campo creativo.

Su utilidad real estaba en haber sido un producto creado mediante la experiencia en el campo aplicado, lo cual proporciona un modelo integral que cubre todo el ciclo de vida del proyecto. Ayudando desde la planificación, la comunicación del proyecto y la documentación.

4.2.2 Desventajas

Una de sus grandes limitaciones como objeto de estudio de la minería de datos es que no realiza actividades de gestión de proyectos. Puesto que se define que tareas deben hacerse, pero, no el cómo deben de hacerse lo cual, afecta a la coordinación del equipo y requiere ser combinada con alguna metodología ágil como SCRUM para abordar los desafíos que esta metodología enfrenta (Saltz, 2021).

Otra clara desventaja es el hecho sobre que para el año 2000 la tecnología no tenía las dimensiones actuales. Por lo cual, muchos problemas no pudieron ser previstos y actualmente se enfrenta a una necesidad de extensiones o remodelaciones para acoplarse a entornos modernos y no quedar obsoleta, estos problemas van desde el manejo de la big data, baja robustez en el modelo en la utilización de datos con gran dimensionalidad, la falta de mecanismos explícitos para integrar aspectos legales, éticos o relacionados con la confianza y las adaptaciones de dominio para campos específicos como la industria manufacturera o la atención médica.

Como ya se mencionó, esta presenta un fuerte problema a la hora de organizar los procesos y tareas, lo cual se nota principalmente en la parte de despliegue, puesto que muchas de las implementaciones de este modelo, tienen un despliegue muy deficiente, o no incluyen estas nuevas soluciones con modelos de negocio o sistemas de TI. Lo cual indica una brecha entre el rendimiento teórico y práctico de este modelo. Donde los proyectos se detienen en la parte de la evaluación aún en entornos de investigación (Saltz, 2021).

4.2.3 Fases de la Metodología

Comprensión del negocio:

Esta fase debe dominar los objetivos del negocio, evaluar la situación y determinar los objetivos del proyecto a lograr desde la minería de datos con la finalidad de producir un proyecto.

Dentro de esta fase de busca comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial. Su función es poder traducir las necesidades de la empresa en una definición del problema de minería de datos y un plan de proyecto

preliminar, a través de cual se generarán las tereas iniciales y se organizara al equipo (Saltz, 2021).

Comprensión de datos:

Esta fase busca metas preparatorias sobre los datos, entre sus faces se incluye el poder recopilar los datos, describirlos, explorar la utilidad y uso de cada uno, así como verificar la calidad de estos (Saltz, 2021).

Esta fase comienza con la recopilación de los datos iniciales, esto incluye cualquier fuente confiable de la cual obtener los datos, estos no deben tener ruido a causa de una mala recolección o huecos dentro de los mismos, siendo una base consistente y útil. Esta fase se centra en el conocimiento de los datos, busca saber que datos existen con la finalidad de aportar conocimiento a la investigación y cuáles no.

Preparación de los datos:

Esta fase tiene tareas muy marcadas como son la limpieza de los datos, selección de registros y atributos, la construcción de los datos, integración y homogenización de los mismos.

Siendo una fase fundamental para el proceso de la minería de datos, esta fase busca obtener una base de datos altamente eficiente capaz de poder obtener estructuras de datos n direccionales útiles, compactas, concisas y lejos de todos aquellos datos que podría alterar una correcta aplicación de los algoritmos. Esta fase se define como una de las más largas de todo el proceso (Saltz, 2021).

Modelado:

La aplicación de los modelos de minería de datos se ejecuta en esta fase, se deben seleccionar las técnicas a aplicar y se genera un modelo de prueba para hacer una evaluación técnica de dicho modelo con datos y recursos reducidos, una vez que los n modelos pasan las pruebas de calidad se debe generar un modelo donde se puede ejecutar el algoritmo con todos los algoritmos y recursos necesarios para su ejecución. Por último, se evalúa este modelo mediante técnicas o índices y se almacenan los resultados obtenidos.

Esto con cada uno de los modelos seleccionados en etapas anteriores y se procede a su ejecución, este proceso es uno de los más rápidos de todo el proceso. Con cada iteración de los modelos, los parámetros deben ser ajustados y calibrados para evaluar las técnicas entre sí (Azevedo, 2008).

Evaluación:

Aquí se deben evaluar los resultados obtenidos contra los objetivos de negocio de la fase 1, re revisa de nuevo todo el proceso seguido hasta ahora y se determina si las iteraciones deben parar o continuar con el despliegue de la información y el modelo ya obtenido.

Todos los modelos generados deben pasar por una rigurosa evaluación y generar conocimiento útil para la aplicación y para el campo estudiando antes de pasar al despliegue.

Despliegue:

Esta fase implica una puesta en producción de los entornos empresariales finales de la aplicación, no en todas las aplicaciones de esta metodología se logra un despliegue, aunque este es recomendado para un proceso de vida completo y practico.

Este debe planificar y monitorear todas las fases de la implementación, generando un informe sobre cómo crear este proceso repetible en toda la empresa y sobre cómo mantenerlo. Así como una informe final y revisión retrospectiva del proyecto (Azevedo, 2008).

4.3 Metodología SEMMA

SEMMA es un acrónimo que representa las 5 fases de un proceso minería de datos: Sample, Explore, Modify, Model, and Assess, que se traduce como: Muestrear, Explorar, Modificar, Modelar y Evaluar.

Esta metodología fue desarrollada por el SAS Institute, su propósito principal es organizar metodológicamente las funciones de su software estadístico y de inteligencia de negocios específicamente la herramienta SAS Enterprise Miner. Algo que hay que tener en consideración, es que, dentro de su documentación se explica que SEMMA no es una metodología de minería de datos, en el sentido de un marco de gestión de proyectos de ciclo de vida completo. Lo describe como una organización lógica del conjunto de herramientas funcionales de SAS Enterprice Miner (Inc., 2024).

Esto quiere decir que no está diseñado para gestionar un proyecto de principio a fin. Esta tecnología, más bien, es un flujo de trabajo de desarrollo de modelos o un proceso centrado en los datos. Describe la secuencia lógica de pasos que un analista de pasos realiza dentro de un entorno de software para construir y validar un modelo predictivo (Inc., 2024).

4.3.1 Ventajas

Las fortalezas de esta metodología radican en su definición como flujo de trabajo, su principal ventaja es la eficiencia de procesamiento y su velocidad. En su fase inicial de muestreo de mina una muestra representativa en lugar del volumen completo, reduciendo así el tiempo de procesamiento. Este paso hace que sea tan fácil, rápido y preciso como codear toda la base de datos, permitiendo que el analista pueda producir rápidamente un modelo validado (Inc., 2024).

El modelo obtenido proporciona un enfoque lógico y permite aplicar técnicas estadísticas y de modelado. Este enfoque integrado con las herramientas SAS le permite al analista obtener resultados mediante un GUI que lo guía durante el flujo de todo el proceso (Inc., 2024).

A pesar, de ser un modelo aplicado, este enfoque permite resolver una amplia gama de problemas de negocio incluyendo la detección de fraude, marketing, retención de clientes y análisis de riesgos (Inc., 2024).

4.3.2 Desventajas

Las limitaciones de SEMMA son consecuencia de sus alcances deliberadamente definidos. Su depende3ncia del proveedor es una de las desventajas clave para su ejecución en la búsqueda de investigaciones no aplicadas a negocios. Este es un modelo de proceso especifico de proveedor y al ser dependiente de las herramientas SAS, es poco práctico usarlo por fuera de ellas lo que limita su aplicabilidad en diferentes entornos que utiliza otras herramientas de software (Inc., 2024).

Este modelo omite una etapa muy importante para la obtención de datos, la cual es la comprensión de negocio, enfocándose únicamente en el flujo de trabajo asumen que los objetivos ya han sido previamente definidos lo cual puede ser incorrecto y corren el riesgo trabajar sin metas definidas y llegar a modelos sin una calidad deseada.

Otra etapa muy importante que ignora por completo, es el despliegue de los modelos previamente obtenidos. Consiguiendo un flujo de trabajo completo sin la tarea que implementa en un modelo de negocio este nuevo conocimiento. Lo que termina en un ciclo de vida incompleto y con una evaluación sin propósito (Inc., 2024).

Por último, hay que tomar en cuenta que el método SEMMA no es un modelo típico de minería de datos. Por lo cual, el omitir pasos fundamentales para otros modelos es solo muestra de sus aplicaciones específicas dependientes de sus herramientas a las cuales fue construido este modelo.

4.3.3 Fases

Muestreo (Sample):

Esta fase consiste en extraer una porción de un gran conjunto de datos. El objetivo es obtener una gran cantidad de datos lo suficientemente consistente como para poder ser procesada pero tan reducida como se pueda para poder obtener un procesamiento rápido. El sistema SAS indica que se busca una estrategia de muestreo para reducir el tiempo de procesa miento y en algunas ocasiones este paso se omite por la baja calidad de los resultados (Inc., 2024).

Exploración (Explore):

Se exploran los resultados obtenidos en el paso anterior buscando relaciones, hipótesis y resultados anticipados buscando anomalías y tendencias que posteriormente deban ser confirmadas y así comprender un poco mejor la naturaleza de los datos. Su objetivo es comprender ideas sobre el conjunto de datos muestreado.

Modificación (Modify):

Se crean y se modifican las variables para poder generar modelos orientados a los distintos algoritmos escogidos. La finalidad de esto es obtener variables predictivas más significativas y de la misma forma, orientar los resultados hacia las hipotesos previamente definidas (Inc., 2024).

Modelado (Model):

Esta es la aplicación de los algoritmos, con todo el trabajo anterior, se aplicarán los distintos algoritmos y se busca confirmar o desmentir las hipótesis previamente marcadas. Buscando combinaciones de datos o patrones que puedan demostrar de manera fiable resultados y conclusiones (Inc., 2024).

Evaluación (Assess):

Esta fase está orientada completamente a la evaluación del modelo obtenido en la fase anterior, por medio de diferentes métricas se evalúa la viabilidad y fidelidad del modelo anterior. Obteniendo finalmente una conclusión sobre el modelo de datos o tomando la decisión de volver a comenzar con una nueva aplicación de dicha metodología con un entendimiento de los datos distinto (Azevedo, 2008).

4.4 Selección de la Metodología

En el presente estudio, se busca el descubrimiento es conocimiento por medio de la aplicación de algoritmos no supervisados para la identificación de la sequía, como marco de referencia metodológico se seleccionó a Knowledge Discovery in Databases (KDD), en virtud a su enfoque integral flexible, y orientado al descubrimiento de conocimiento. Esta elección ya cumple con las características solicitadas por el problema mismo y se ajusta a las limitaciones y necesidades. A diferencia de sus contrapartes estudiados como SEMMA y Crisp-DM.

La metodología KDD aborda de una forma sistemática todo el proceso del descubrimiento del conocimiento, comprendiendo las etapas de selección, preprocesamiento, transformación, minería de datos e interpretación del conocimiento. Esta estructura resulta idónea para el tratamiento de imágenes multiespectrales, los cuales requieres procesos complejos para la identificación y corrección de información útil y significativa. Esta metodología ofrece un marco metodológico coherente con la naturaleza de la investigación y con los objetivos de un estudio exploratorio basado en algoritmos no supervisados (Azevedo, 2008).

En comparación, la metodología CRISP-DM constituye un modelo amplio y flexible, altamente utilizado en contextos industriales. Aunque este fue el punto más débil para su elección, esta metodología está orientada a la evaluación e implementación en entornos meramente industriales o empresariales (Saltz, 2021). Su enfoque está centrado en la comprensión de un negocio y la validación de los resultados a través de métricas de desempeño. A pensar de poder adaptarse en contextos científicos, su estructura enfatiza un modelo supervisado y meramente validado, lo cual no resulta correcto para la realización de esta investigación de carácter exploratorio (Saltz, 2021).

La metodología SEMMA desarrollada por SAS Institute, está orientada principalmente al desarrollo de modelos predictivos dentro de los entornos de análisis estadístico y de negocio. Su estructura asume que los datos se encuentras previamente depurados y listos para el modelado omitiendo etapas esenciales para una investigación de carácter exploratorio como lo es esta investigación (Inc., 2024). Las limitaciones del programa Landsat en sus imágenes multiespectrales como lo son el ruido, variaciones radiométricas, y diferencias espectrales significativas, hacen que esta metodología no sea meramente útil a la hora de desarrollar este trabajo (Inc., 2024).

Una vez teniendo todo esto en cuanta, se puede concluir que KDD facilita la integración de los resultados en un contexto científico y ambiental, contribuyendo a la

toma de decisiones en materia de monitores o gestión de recursos naturales (Azevedo, 2008).

La elección de KDD como metodología base, está basada mayormente en los siguientes puntos:

- Su enfoque permite un ciclo de vida completo desde la adquisición hasta la interpretación del conocimiento
- Su adecuación a datos con gran volumen y dimensionalidad como lo son las imágenes multiespectrales.
- Su orientación es de carácter científico y exploratorio, ideal para trabajar con algoritmos no supervisados.
- Su capacidad para generar conocimiento significativo en contextos de análisis ambiental y geoespacial.

Estas razones permiten observar por que KDD es la metodología escogida y apropiada para el desarrollo de este trabajo, esta permite abordar de manera rigurosa y estructurada el proceso de descubrimiento de patrones relacionados con la sequía a partir del análisis de información multiespectral.

CAPÍTULO 5 ESTADO DEL ARTE

Ahora se deben evaluar los trabajos hechos con anterioridad sobre este tema, sobre estos trabajos se analizarán diversos aspectos, que deben ser fundamentales para ser tomados en cuenta.

Los aspectos más importantes serán analizar el objetivo, este debe ser relacionado con la meteorología y/o la sequía, este objetivo debe incluir el poder generar grupos de sequía o generar análisis sobre la vulnerabilidad a la sequía, así como poder correlacionar fenómenos atmosféricos con la sequía.

Contar con una delimitación espacio temporal de estudio, cada estudio debe estar delimitado tanto en años de estudio como en área geográfica específica.

Cada estudio debe incluir métodos de minería de datos para así, poder obtener conocimiento nuevo, de preferencia haciendo uso de algoritmos de aprendizaje no supervisado.

Por último, cada investigación debe obtener resultados describiendo la información obtenida, estas pueden ser nuevas agrupaciones o el descubrimiento de patrones ocultos en las variables que se tomaron en cuenta para el estudio.

Los estudios presentados cumplen con la mayoría de estos aspectos y pueden aportar información importante sobre cómo es que esta investigación es relevante en el campo de estudio y futuramente un marco de referencia para investigaciones futuras.

Análisis Espacio-Temporal De La Sequía

Spatial And Temporal Characteristics of Drought and Its Correlation with Climate Indices in Northeast China. Fue publicado en el año 2021, el objetivo de esta investigación es analizar las características especiales y temporales de sequía, en China, este estudio se realizó en tres regiones: Noroeste (R1), Occidental (R2) y Sureste (R3). Para esto se basaron en el índice estandarizado de precipitación (SPI) entre 3 y 12 meses, para, así poder determinar la relación entre la sequía y patrones climáticos (Wang, 2024).

Los algoritmos utilizados en este trabajo fueron analizados por el método jerárquico, se enfatiza en el uso del método Ward como enlace y el uso de la distancia euclidiana. También se realizó una prueba de tendencia Mann-Kendall y un estimador de pendiente Sen para determinar tendencias en las series anuales y estacionales (Wang, 2024).

Este trabajo arrojó una reducción de las sequías entre primavera e invierno en la región R1, identificó un aumento de sequía en el otoño en las regiones R2 y R3. Y por último identificó, a escala anual que en las regiones R2 y R3 son más propensas a la sequía, ya que, en estas regiones las sequías se intensificaron y fueron severas, mientras que en R1 la sequía se mitigó (Wang, 2024).

Este estudio también identificó que al noreste de China la sequía tiene intervalos de 2 a 6 años en las 3 regiones, aunque en las regiones R2 y R3 la sequía se da con mayor intensidad (Wang, 2024).

Análisis Espacio-Temporal De La Sequía En España Peninsular. Influencia De Los Principales Patrones De Tele conexión. Es un estudio realizado en el año 2018, realizó un análisis espacio temporal sobre la sequía en España en una región peninsular y su influencia de los principales patrones de telecomunicación. Para esto analiza información dada de forma semanas que comprende desde 1962 hasta 2014 para el índice SPEI y datos con resolución temporal mensual en el mismo periodo para los patrones de telecomunicación. realizando el análisis a partir de una descripción espacial y temporal. La zona de estudio comprende a España Peninsular y con este estudio también se buscó el desarrollo de nuevas metodologías en la predicción estacional de sequía (Salvador, 2018).

La información se obtuvo a través del índice SPEI con una configuración de 3 meses. Y patrones de telecomunicación donde se analizaron la oscilación del atlántico norte, este mediterráneo este entre otros. Con el método de agrupamiento K means se dividió la región en grupos para identificar las zonas con evolución temporal del índice SPEI homogénea (Salvador, 2018).

Para analizar toda esta información se utilizaron técnicas de análisis estadístico, tales como el valor promedio del SPEI en cada clúster y se comparó con los patrones de telecomunicación, Se obtuvieron distribuciones espaciales y se calculó la correlación entre el índice SPEI y los patrones de telecomunicación. Se generaron contrastes de hipótesis y transformadas de Wavelets para validar la viabilidad del estudio (Salvador, 2018).

Este estudio logró concluir que la oscilación del atlántico norte y el patrón de oscilación del ártico son aquellos con mayor viabilidad hidrológica. La oscilación del mediterráneo oeste es aquella con mayor importancia en la sequía del sureste de España. Según los datos, se encontró una correlación entre el SPEI y los patrones de

telecomunicación donde la oscilación del mediterráneo oeste y la oscilación del atlántico norte se relacionan entre enero y mayor mayormente (Salvador, 2018).

La oscilación del atlántico norte produce correlaciones negativas en todo el territorio peninsular siendo más importantes en la vertiente atlántica. Una fase positiva produce un SPEI negativo debido a que bloquea la precipitación muestras que, de forma inversa, se producen precipitaciones intensas en la vertiente atlántica (Salvador, 2018).

Por otro lado, la oscilación del mediterráneo oeste presenta correlaciones bipolares dando resultados negativos en la vertiente mediterránea y positivos en el área cantábrica (Salvador, 2018).

Análisis De La Sequía Meteorológica En Cuatro Localidades Agrícolas De Venezuela Mediante La Combinación De Métodos Multivariados. Es un trabajo realizado en el año 2018, analizó la ocurrencia de la sequía climatológica por medio de series temporales del índice SPI para cuatro localidades venezolanas. Este estudio también buscó servir de base para una correcta planificación del espacio y fortalecimiento de las estrategias para la seguridad alimentaria de dichas localidades. Para esto se obtuvieron registros de entre 1980 a 2014. Las localidades estudiadas fueron: El Tigre (Anzoátegui), Banco de San Pedro (Guárico), Tapipa-Padrón (Miranda) y El Guayabo (Zulia) (Olivares, 2018).

Las técnicas utilizadas en este trabajo fueron el índice SPI para cuantificar las condiciones de déficit o exceso de precipitación. Métodos multivariados (análisis de coordenadas principales y análisis de clúster) para reducir los patrones de sequía a unos pocos. Se determinó el número de grupos óptimo utilizando el índice de Silhouette. Por último, se categorizaron los años dependiendo el nivel de humedad obtenido (Olivares, 2018).

Tras el análisis se determinó que, para las zonas de El Tigre y Banco de San Pedro hubo 3 grupos de años húmedos, normales y con déficit hídrico. Mientras que para Tapipa-Padrón y el Guayabo solo hubo dos grupos, húmedos y secos (Olivares, 2018).

Se determinó que la comunidad de El Tigre que, aunque normalmente se mantuvo en una media, tuvo consecuencias importantes para los cultivos frutales el déficit hídrico. San Pedro fue el grupo más seco donde la precipitación disminuyó significativamente. En Tapipa-Padrón hubo una mayor humedad y tuvo mayor cantidad de lluvia, aunque, la sequía también afectó a cultivos como el cacao. Por último, El Guayabo también fue una localidad con predominancia por la humedad en los años de estudio, aunque las

épocas de sequía causaron estragos en cultivos como la palma aceitera, plátanos, cacao y pastos (Olivares, 2018).

Clasificación, Tipificación Y Agrupamiento De La Sequía

Unsupervised Clustering of Forest Response to Drought Stress in Zululand Region, South Africa. Es un estudio realizado en el año 2019 buscó evaluar la utilidad de enfoques de mapeo no supervisado para agrupar comportamientos de árboles de eucalipto con características de sequía similares, todo esto basándose en el índice normalizado de diferencia de agua (NDWI). Y de esta forma demostrar que herramientas cotidianas como lo son Google Earth Engine puede ser muy valiosas a la hora del monitoreo de sequía a nivel de paisaje. La región de estudio de este trabajo es la región de Zululand en Sudáfrica, se estudió una región de plantación forestal de cerca de 20,000 hectáreas en Kwabonambi, esta plantación forestal se caracteriza por mantener mayormente variaciones de la planta de eucalipto (Xulu, 2019).

Como ya se mencionó, la forma de obtención de los datos fue mediante el índice (NDWI), este índice fue implementado con el propósito de estudiar el contenido relativo de agua de la región. También se utilizó la herramienta Google Earth Engine, de esta herramienta se obtuvieron imágenes Landsat de forma gratuita.

Los algoritmos implementados para este trabajo fueron la matriz de proximidad de random forest para poder medir distancias no lineales entre píxeles de las imágenes obtenidas. Medidas de distancia lineal como la euclidiana o la Manhatan (Xulu, 2019).

Los métodos de clustering fueron el algoritmo jerárquico utilizando Ward como método de enlace y K Means utilizado para separar comportamientos en clusters afectados y no afectados en la sequía. Adicionalmente se utilizaron métricas de evaluación de precisión como la precisión general, la precisión del productor y la precisión del usuario para reafirmar los resultados (Xulu, 2019).

Este estudio concluyó que la matriz de proximidad de random forest produjo mejores resultados para el estudio, obteniendo la mejor calificación sobre todas las métricas, dando un total de 87.7% de precisión. Seguido de esto la mejor métrica fue de la distancia Manhattan y euclidiana (Xulu, 2019).

Dentro de las conclusiones referentes a los grupos, se encontró que existieron dos grupos principales que difieren de sus resultados a la sequía, el primer grupo con comportamientos afectados por la sequía y un índice NDWI bajo y el segundo grupo no afectado por la sequía (Xulu, 2019).

Observó que, durante el paso de los años, el contenido de agua de hojas en todos los árboles se redujo, coincidiendo con el fenómeno “El Niño”.

Estos resultados confirmaron que el uso de la herramienta Random Forest y la capacidad de los enfoques de mapeo son esenciales para el monitoreo automático del estrés por sequía en plantaciones forestales además de destacar el uso de la herramienta Google Earth Engine como medio de obtención de imágenes satelitales para el monitoreo de sequía a gran escala (Xulu, 2019).

Drought Vulnerability Assessment and Cluster Analysis of Island Areas Taking Korean Island Areas at Eup (Town) And Myeon (Subcounty) Levels as Study Targets

Es un estudio realizado en el año 2021, este trabajo busca realizar una evaluación de vulnerabilidad de sequía y generar grupos sobre las zonas denominadas como EUP (ciudad pequeña) y myeon (sub condado), este estudio se realizó para 90 zonas insulares de Corea (Shim, 2021).

Las herramientas utilizadas en este estudio fueron el análisis factorial para poder producir la dimensionalidad de los datos y así poder seleccionar los 22 indicadores de vulnerabilidad, estos incluyen exposición climática, sensibilización y capacidad adaptativa. Método de re escalado, estandariza los indicadores entre 0 y 1. Cálculo de índice de vulnerabilidad de la sequía, mediante una fórmula que incluye factores obtenidos en el análisis factorial. Para la formación de las agrupaciones se utilizó el algoritmo k means para clasificar las 90 áreas insulares en grupos por medio de la vulnerabilidad. Utiliza el método de la curva del codo para definir el número óptimo de clusters. Por último, utilizó ANOVA y HSD para verificar la diferencia y significancia estadística entre los clusters (Shim, 2021).

Los resultados obtenidos durante la investigación revelaron que los indicadores con mayor impacto en la evaluación de la sequía según el método de entropía fueron la precipitación invernal, los días sin lluvia, tasa de población agrícola, área de cultivo, suministro de agua y capacidad subterránea. Estos factores se relacionaron con la agricultura obteniendo un alto peso en la sostenibilidad (Shim, 2021).

La evaluación reveló que las zonas más vulnerables fueron: Seodo-myeon (Ganghwa-gun), Seolcheon-myeon (Namhae-gun) y Samsan-myeon (Ganghwa-gun). Siendo Ganghwa-gun la zona mayor vulnerabilidad a la sequía en relación a las áreas insulares cercanas (Shim, 2021).

Las agrupaciones realizadas revelaron que de las 90 zonas insulares la mejor combinación resultó en 3 clusters. El primero es vulnerable a la exposición climática con buena sensibilidad y capacidad adaptativa, el segundo es vulnerable a la sensibilidad, tiene un alto impacto a la población y agricultura/pesca y el último es vulnerable a la capacidad adaptativa con buena exposición climática y sostenibilidad (Shim, 2021).

Diagnostic Classification of Flash Drought Events Reveals Distinct Classes of Forcings and Impacts. Este trabajo se realizó en el año 2022 buscando examinar si el término en inglés conocido como "Flash drought" comprende múltiples clases distintas de eventos, lo cual implicaría la necesidad de más de un estudio para su comprensión y pronóstico. Para ser más específico, busco clasificar estas sequías repentinas basándose en las condiciones precursoras meteorológicas y de la superficie. Esta investigación se realizó basados en el país de Estados Unidos de América abordado en este trabajo como Contiguous United States (CONUS). Con datos que comprenden los periodos de 1979 a 2018 para las estaciones de primavera a otoño (Osman, 2022).

Dentro de la realización de esta investigación se utilizó el índice Soil Moisture Volatility Index (SMVI) para definir la sequía repentina como el promedio de humedad en el suelo cae en un radio de 5 días y se mantiene por debajo de la media de 20 días atrás manteniéndose por aproximadamente 20 días más por debajo de este. Se utilizó la severidad para cuantificar una escala de 0 a 5 basada en el déficit integrado estandarizado de RZSM por debajo del percentil 20 durante el evento (Osman, 2022).

Para la clasificación de los grupos se utilizaron técnicas de clustering no supervisadas de severidad mayor a 2. Dentro de este grupo de técnicas se utilizó el algoritmo K Means para realizar las agrupaciones. Para determinar el número óptimo de grupos, se utilizó el método del codo, determinando así un número óptimo de 3 clusters. Las variables para la calificación utilizadas fueron las anomalías estandarizadas de la partición de inicio, las cuales comprenden nueve variables: temperatura (TMP), precipitación (PRCP), humedad del suelo en la zona radicular (RZSM), evapotranspiración real (EVP), evapotranspiración potencial (PEVP), presión superficial (SPRES), cobertura total de nubes (TCC), velocidad del viento (WS) y déficit de presión de vapor (VPD) (Osman, 2022).

Los resultados obtenidos refieren que los 3 grupos obtenidos comparten características entre sí. El primer grupo son las sequías sigilosas, estas se caracterizan por carecer de lluvias en un área evaporativa lo que las hace difíciles de identificar siendo el tipo más común en las altas llanuras occidentales y las más extendidas a lo

largo del año. El segundo grupo son las secas y demandantes, estas se caracterizan por tener mucha demanda evaporativa antecedente, baja humedad del suelo y baja evapotranspiración real, lo que muestra una mayor severidad promedio y son dominante en el sur de las grandes llanuras y Texas. Por último, es grupo 3, tiene anomalías positivas en la evapotranspiración real y alta demanda evaporativa, con anomalías de precipitación y humedad del suelo modesta, este tipo es muy común en el Alto medio oeste (Osman, 2022).

Se concluyó que las sequías repentinas son un compuesto de distintas sequías de rápida intensificación y los análisis y pronósticos se beneficiarán de enfoques que reconozcan la existencia de múltiples factores fenomenológicos (Osman, 2022).

Drought And Vulnerability in Mexico's Forest Ecosystems. Investigación realizada en 2023, busca caracterizar la sequía en los últimos veinte años y evaluar la vulnerabilidad de los ecosistemas forestales en México con respecto a este fenómeno. Se realiza en áreas con grandes concentraciones de árboles en México como pueden ser bosques templados, tropicales o zonas semiáridas a niveles municipales (Agustín-Canales, 2023).

Para la obtención de los datos, esta investigación hizo uso del SPEI a 12 meses, el cálculo del índice de Vulnerabilidad (VI), el índice de vulnerabilidad de sequía (DVI) (Agustín-Canales, 2023).

Como técnicas de Minería de datos se utilizó el método K Means para generar cuatro grupos de sequía. El análisis de Moran Local bivariado que ayuda a evaluar la correlación espacial y la correlación entre la sequía y vulnerabilidad entre espacios de estudio vecinos. Y la Interpolación para mapear la presencia de sequía anualmente utilizando datos de 415 estaciones meteorológicas (Agustín-Canales, 2023).

Esta investigación arrojó que, dentro de las regiones estudiadas, se sufre sequía año con año por lo que es altamente vulnerable a este fenómeno. De los 21 años de estudio, la sequía abarcó más del 50% de la superficie nacional. Durante el año 2011 llegó a cubrir casi el 77% de la superficie del territorio nacional, siendo este, el año más grave. Las categorías más severas de sequía encontradas se localizaron en el norte, oeste y en la meseta central de México (Agustín-Canales, 2023).

De los municipios estudiados, cerca de un 20% se encuentra en una muy alta vulnerabilidad debido a la sequía. Mientras que, en las zonas forestales, la sequía ha afectado cerca del 90% de los bosques del país. En el 49% de las zonas forestales se experimentó sequía extrema y severa. Se identificó que la correlación alta entre sequía

y vulnerabilidad se dio en los estados de Durango, por lo que se identificó como atención urgente (Agustín-Canales, 2023).

Después de un análisis exhaustivo de estas investigaciones se obtuvo información importante para la investigación. La mayoría de estudios determinó un aumento en la sequía al pasar de los años, este aumento contrajo muchas afectaciones a la actividad humana como al medio ambiente, entre estas afectaciones se encontraron afectaciones en diversos cultivos alrededor del mundo, disminución en el volumen de agua de diversos ecosistemas, una menor cantidad de precipitaciones en zonas específicas, patrones de sequía anormales y prolongados etc.

Dentro de todas estas respuestas se puede rescatar aspectos técnicos para nuestra investigación, estos son:

La mayoría de investigaciones donde se busca agrupar los tipos de sequía se utiliza el algoritmo K means, seguido de las redes neuronales.

La mayoría de estas investigaciones requieren de una fuente confiable de datos como lo son las imágenes multiespectrales o los índices como el NNDI, SPEI, SPI. Estas fuentes de información han demostrado ser muy efectivas para poder analizar patrones y generar agrupaciones para la sequía.

Dentro de todos estos trabajos se pueden localizar áreas de oportunidad, ya que, muchas directamente realizan la ejecución de algoritmos sobre los índices, y aunque, algunas sí evalúan la precisión de las respuestas obtenidas, estas suelen ser sobre pocas herramientas, lo cual no permite obtener un panorama completo sobre algunas herramientas que potencialmente podrían tener mejores resultados sobre las ya establecidas.

Una vez enmarcado este aspecto clave, en este trabajo se desea analizar la versatilidad de los datos mientras son sometidos a distintos algoritmos de aprendizaje no supervisado. Estas pruebas permitirán saber cómo es que los datos de las imágenes multiespectrales se pueden comportar con diferentes algoritmos y al mismo tiempo evaluar la calidad de los resultados obtenidos. Siendo un pilar importante para futuros trabajos e investigaciones sobre el mismo tema, ya sea, reafirmando el dominio o recesión de los algoritmos mayormente utilizados en este tipo de investigaciones.

Pronostico y modelado predictivo de la sequía

Pronóstico De Sequías Usando Redes Neuronales Artificiales En La Cuenca Del Río Sonora, México. Este trabajo fue publicado en el año 2021, el objetivo de dicho trabajo es el de aplicar redes neuronales artificiales (RNA) para pronosticar las sequías meteorológicas en la parte media y alta de la cuenca del río Sonora, este trabajo utiliza datos de los índices SPI y SPEI a escalas 3, 6, 12 y 24 meses entre 1974 y 2013. Este estudio se centró en la cuenca del río Sonora, utilizando 19 estaciones climatológicas agrupadas en 4 regiones, R1, R2, R3 y R4 (Hernández-Vásquez, 2022).

Para este estudio se utilizarán los datos de índices como el SPI y SPEI, con distintas escalas para garantizar datos homogéneos. Para la identificación de la información oculta en los datos, este estudio utilizó análisis de componentes principales (PCA) para regionalizar la zona de estudio en regiones similares de precipitación. K Medoides para analizar los clusters y agrupar las estaciones. El uso de redes neuronales para pronosticar índices de sequía realistas de formas complejas y veraces. Algoritmo de Aprendizaje Resilient Propagation (RPROP+ y RPROP-) Utilizado para ajustar pesos sinápticos y de esa forma evitar convergencia. Validación cruzada para controlar la formación de grupos e identificación de datos mal agrupados o sobreajuste en los datos y así ajustar la capacidad de generalización (Hernández-Vásquez, 2022).

Este estudio concluyó que, hubo muchas tendencias al aumento de la intensidad de la sequía y las frecuencias de esta. Los eventos clave en aumento de la sequía ocurrieron en los años (1997, 1999, 2000 y de 2011 a 2013). El SPEI definió mejor los periodos, tendencia y aumento de la sequía que el SPI, demostrando que incluir la evapotranspiración es fundamental para la identificación de la sequía. La región R2 (Valles Inter montanos) resultó ser aquella con mayor ocurrencia en la sequía y la más vulnerable (Hernández-Vásquez, 2022).
El modelo de RNA resultó bastante efectivo puesto que sus resultados fueron satisfactorios. obtuvo un promedio de 0.76 en los resultados finales de validación, superando a los índices SPEI y SPI. Su eficiencia aumentó conforme la escala aumentaba, siendo los resultados de 24 meses mejor que 3 meses. Los resultados se asemejan al monitor de sequía en México (MSM) para enero marzo 2014 (Hernández-Vásquez, 2022).

CAPÍTULO 6 ÁREA DE APLICACIÓN

De la misma forma se abordarán temas frente al cumplo de aplicación de estos, como lo son el programa de visualización y análisis de imágenes multiespectrales, Landsat. Se analizará que es la sequía, los distintos niveles de sequía que maneja y organismos reguladores encargados del monitoreo de este fenómeno en México.

6.1 Sequía

El panel gubernamental sobre el cambio climático ha definido a la sequía como un periodo en el cual las condiciones de sequía se prolongan por tiempos considerados anormales hasta lograr un desequilibrio hidrológico grave en una zona geográfica. Para que un periodo sea considerado como anormal, es necesario conocer la zona geográfica a estudiar, Por ejemplo, mientras que en zonas lluviosas 5 días sin lluvias ya pueden ser un periodo anormal, en zonas áridas meses de sequía podrían no considerarse anormales. La sequía ya ha mostrado impactos a lo largo de los años en distintos ámbitos como son: La agricultura, el turismo, el estrés de la vegetación y la degradación del suelo (Agenda Hidalguense, 2025).

Para poder medir si existe o no sequía en una determinada zona, es necesario de valerse de herramientas como lo son los índices de sequía, estos pueden tener dos clasificaciones (Salas-Martínez, 2023).

La primera comprende aquellos que usan información meteorológica basada “in situ”; esta comprende desde el índice de precipitación y evapotranspiración estandarizado (SPEI), hasta, el porcentaje de precipitación normal (PNP), entre otros. Aunque este tipo de índices requiere de estaciones meteorológicas que obtengan datos completos y actualizados en tiempo real, por lo que en países como México esto puede limitar la obtención de los datos debido a la escasez de estas (Agenda Hidalguense, 2025).

Y utilizar la teledetección para recopilar datos de la zona, se tienen índices como el índice de vegetación de diferencia normalizada (NDVI), el índice de salud de la vegetación (VHI), el Índice de condición de la vegetación (VCI), etc. Estos se valen de otras herramientas como imágenes multiespectrales para obtener datos meteorológicos sobre la sequía, el inconveniente de estos es que las imágenes suelen estar disponibles cada 15 días.

Estas diferencias entre las clasificaciones e índices han permitido que surja un nuevo índice en unión con los anteriores, el NDDI (Índice de Sequía por Diferencia) el cual mejora aspectos como lo son el comportamiento, distribución, e intensidad de la

sequía, lo que lo ha favorecido para consolidarse como una herramienta necesaria para la medición de sequía a lo largo del mundo (CONAGUA, s.f.).

6.2 Servicio Meteorológico Nacional

El Servicio Meteorológico Nacional (SMN) es una dependencia gubernamental con el objetivo de proporcionar datos meteorológicos y de climatología, para realizar esta tarea el SMN se vale de herramientas como: estaciones automáticas, observatorios sinópticos, radares, estaciones de radio monitoreo y receptoras de imágenes de satélite, entre otras. Esta organización gubernamental desde 2014 proporciona datos meteorológicos basados en la metodología usada por USDM (U.S. Drought Monitor) y NADM (North American Drought Monitor) (CONAGUA, s.f.).

El Monitor de Sequía en México (MSM) se basa en la obtención de datos meteorológicos desde distintos índices o indicadores, los cuales con:

- Índice Estandarizado de Precipitación (SPI). Mide de forma numérica el exceso o déficit o exceso de lluvias y la anomalía de precipitación en porciento de lo normal. Ambos medidos en días, normalmente en (30,90,180,365).
- Índice Satelital de Salud de la Vegetación (VHI). Mide el grado de estrés de la flora a través de la radiación de esta.
- Modelo de Humedad del suelo Leaky Bucket CPC-NOAA. Estima la humedad de los suelos con la ayuda de un modelo hidrológico de una capa específica.
- Índice Normalizado de Diferencia de la Vegetación (NDVI). Esta métrica se basa en evaluar el tono y color de la vegetación
- Anomalía de la temperatura media. Se busca identificar la diferencia entre la temperatura observada y la temperatura de entre 10 y 30 años.
- Porcentaje de la Disponibilidad de Agua en las presas del país. Tomando mediciones sobre la cantidad de agua en los cuerpos de una zona geográfica.
- La aportación de expertos locales. La opinión de expertos sobre el tema aporta valor importante sobre la percepción del tema.

A través de todos estos índices se puede administrar la información por medio de capas, a través de un sistema de información geográfica (SIG) y mediante la opción de expertos en el tema se toman decisiones asignando un valor categórico a cada región, basados en la información obtenida, los valores probables son: *anormalmente seco (D0), sequía moderada (D1), sequía severa (D2), sequía extrema (D3) hasta sequía excepcional (D4).* Una vez identificados los valores, se trazan formas y se publican los resultados cada 15 días y cada mes se publican en el mapa regional o continental NADM (CONAGUA, s.f.).

6.3 Sequía En El Estado de Hidalgo

El Estado de Hidalgo está ubicado en la región central de los Estados Unidos Mexicanos. Esta región se caracteriza por su diversidad de biomas y regiones geo cultural que van desde regiones ricas en árboles, hasta zonas áridas como mesetas sierras, razón por la cual es un blanco ante el fenómeno de la sequía, no solo para las regiones que ya cuentan con una reducción en humedad y precipitaciones, si no, también zonas donde a través de los años, el volumen de agua en su vegetación ha ido disminuyendo, afectando a una gran parte de estos ecosistemas, y por supuesto, a la actividad humana (Martínez, 2024).

Para comprender la crisis hídrica que sufre este estado hay que comprender su distribución de recursos hídricos de los cuales se abastecen tanto los ecosistemas como la población. Los mayores cuerpos de agua renovable del estado se encuentran en la parte norte del estado, zonas donde el índice de pobreza socioeconómica es significativamente mayor. Siendo que el acceso a esta fuente de agua por parte de la población se ve muchas veces limitada por intereses externos a ellos y por la pobre infraestructura para una correcta administración de este recurso. Esta deficiencia lo vuelve altamente vulnerable a la sequía, puesto que una gestión deficiente del recurso requiere de una cantidad mayor de lluvia para compensarlo (Salas-Martínez, 2021).

El monitor de sequía en México (MSM), herramienta de la comisión nacional del agua (CONAGUA). Ha determinado al Estado de Hidalgo con una dinámica de “Aridez Volátil”, la cual se caracteriza por oscilaciones extremas de sequía generalizada y de una severidad mayor. Esta volatilidad sugiere una pérdida de resiliencia hidrológica en los sistemas del estado, los cuales dependen fuertemente de las precipitaciones para la distribución del agua en el Estado. Durante el periodo 2022-2023 el estado reportó cerca de 23 municipios en estado de sequía severa. Especialmente en la región conocida como Valle del Mezquital afectando cerca de 28,000 agricultores de la región. Siendo que el año 2024 cuando el MSM reportó que los 84 municipios tenían un grado importante de sequía. Lo cual enmarca al fenómeno como un problema a nivel del estado y no como casos aislados dentro de su territorio. Para mediados de este año se reportó una sequía de tipo D4 (Sequía excepcional) en cerca del 60% de la superficie estatal siendo este el nivel de sequía más grande durante la última década (El Sol de Hidalgo, 2025).

Para septiembre del 2025 el MSM registró que Hidalgo había alcanzado un nivel de cero sequías derivadas de intensas lluvias durante verano y otro del mismo año. Esto es un nivel de alerta por los cambios abruptos en las condiciones climáticas derivando en

una baja capacidad de amortización para el ecosistema. Los cuerpos de agua superficiales y subterráneos no logran retener agua a largo plazo no que únicamente da una recuperación temporal y superficial al problema, teniendo un presente desequilibrio hídrico (El Sol de Hidalgo, 2025).

Teniendo este contexto sobre la vulnerabilidad del estado de Hidalgo ante un inminente desequilibrio ecológico e hídrico. Es fundamental generar herramientas para el estudio, medición y prevención de la sequía. La falta de estudios sobre el tema crea un área de oportunidad para poder crear nuevas tecnologías capaces de generar conocimiento sobre este fenómeno (Gobierno del Estado de Hidalgo, 2013).

6.4 Imágenes Multiespectrales

La imagenología multiespectral (MSI) representa una de las tecnologías de teledetección más importantes y más utilizadas para el estudio de toda la superficie terrestre. Su utilidad radica en la capacidad de capturar información más allá de los límites de la percepción humana, esto se logra a través de un sensor multiespectral diseñado para aislar y medir la energía electromecánica en varios segmentos espectrales específicos y no contiguos. (Esri Support, n.d.)

Estos segmentos son conocidos como “Bandas espectrales” los cuales se encargan de captar información que es imperceptible para el ojo humano, abarcando regiones como el infrarrojo cercano, de onda corta y el térmico. A pesar de ser prácticamente invisibles para el ser humano, estos cuentan con una basta información para múltiples propósitos (U.S. Geological Survey, n.d.).

Su principio de basa en que conjuntos de satélites captan aquella cantidad de información que no es absorbida por los materiales, por ejemplo, las plantas normalmente obtienen su color de la clorofila, es un concepto que en el espectro visible absorbe el rojo y el azul, por ende, regresa el verde y es lo que observan los satélites, aquellas bandas que son rechazadas por los diversos elementos en el planeta. Esto da como resultado una imagen multicapa donde cada capa es una banda espectral específica. El conjunto de estas bandas puede decir el tipo de cobertura del suelo, la cantidad de humedad de este, la calidad del agua etc. (U.S. Geological Survey, n.d.).

Sus aplicaciones son muy variadas, estas pueden ir desde la gestión agrícola con el monitoreo de la salud de los cultivos, hasta la gestión de los cuerpos de agua midiendo la cantidad de información que reciben las bandas encargadas de gestionar parámetros relacionados al agua, sus aplicaciones permiten no solo monitorear si no,

de la misma forma generar análisis sobre periodos donde los fenómenos puedan o no ocurrir sobre un territorio o región, por ejemplo, tomando muestras de imágenes multiespectrales sobre el volumen de agua de las nubes, es posible hacer proximidades sobre desastres meteorológicos como huracanes o posibles frentes fríos (U.S. Geological Survey, n.d.).

6.4.1 Bandas Del Sensor OLI

El sensor Operational Land Imager (OLI) es el instrumento principal a bordo de satélites encargados de la observación de la tierra, como lo son el Landsat 8 y 9, su función es obtener imágenes de alta resolución de la superficie terrestre en diferentes partes del espectro electromagnético permitiendo un análisis detallado de la vegetación, agua, suelo y áreas urbanas (U.S. Geological Survey, n.d.).

Para las bandas multiespectrales tiene una precisión de un pixel representando cerca de 30 metros y las panorámicas una relación de 15 metros por pixel, aunque es en blanco y negro para mejorar la nitidez. Los datos obtenidos están en una taza de 12 bits lo que permite obtener muchos niveles de intensidad de la luz (U.S. Geological Survey, n.d.).

Las bandas espectrales capturadas por esta herramienta son 11 y cada una se define por su nombre, el rango multiespectral y un propósito:

- Banda 1: Aerosol Costero (0.43 - 0.45 µm). Esta banda captura la luz del espectro violeta profundo y es sensible a la dispersión de la luz por partículas finas. Este es capaz de captar contaminantes como humo o polvo y de observar aguas costeras poco profundas.
- Banda 2: Azul (0.45 - 0.51 µm). Este mide la luz azul visible y también es susceptible a la dispersión atmosférica. Se utiliza principalmente para identificar tipos de suelo y vegetación.
- Banda 3: Verde (0.53 - 0.59 µm). Captura la luz verde visible. Este es particularmente útil para monitorear la salud de los cultivos y zonas forestales protegidas.
- Banda 4: Rojo (0.64 - 0.67 µm). Identifica la luz roja visible. Su propósito es ayudar a diferenciar la vegetación de otro tipo de cobertura de suelo. También es útil en la cartografía identificando obras creadas por el hombre como carreteras o edificios.
- Banda 5: Infrarrojo Cercano o NIR (0.85 - 0.88 µm). Captura la energía en el infrarrojo cercano. Este es utilizado para realizar estudios de vegetación y

biomasa. Al ser reflejado con mucha fuerza por las plantas también permite observar el volumen de los cuerpos de agua.

- Banda 6: Infrarrojo de Onda Corta 1 o SWIR 1 (1.57 - 1.65 µm). Mide la primera parte del infrarrojo de onda corta. Este es útil para conocer la humedad de la vegetación como de las nubes y es capaz de identificar áreas vulnerables por la sequía.
- Banda 7: Infrarrojo de Onda Corta 2 o SWIR 2 (2.11 - 2.29 µm). Este obtiene la segunda parte del infrarrojo de onda corta. Identifica las zonas con alteraciones hidro ambientales que pueden indicar la presencia de minerales o mapear las áreas afectadas por incendios.
- Banda 8: Pancromática (0.50 - 0.68 µm). Esta banda de blanco y negro tiene una precisión de 15 metros por pulgadas de nitidez. Es utilizado para crear mapas de alta definición y observar la distribución urbana.
- Banda 9: Cirrus (1.36 - 1.38 µm). Esta banda únicamente detecta un tipo de nube alta y delgada llamada “Cirro”.
- Banda 10: Infrarrojo Térmico 1 (10.60 - 11.19 µm). Mide la radiación térmica por la superficie de la tierra. Este permite calcular la temperatura del suelo y agua identificando posibles zonas de calor.
- Banda 11: Infrarrojo Térmico 2 (11.50 - 12.51 µm). Mide la radiación térmica en un rango diferente al anterior. Mejora la precisión de la banda anterior, teniendo una mejora de precisión de las mediciones de temperatura. Al igual que el anterior, mide la temperatura terrestre.

6.5 Programa Landsat

El programa Landsat es una colaboración entre la USGS y la NASA. Mientras que la nasa se encarga de la parte de la logística, física, diseño, construcción y demás temas técnicos sobre los satélites. La USGS se encarga de la operación de estos mismos una vez ya estén en órbita, esto incluye la recopilación, procesamiento y distribución de la información obtenida de los múltiples sensores y sistemas a nivel mundial (NASA Scientific Visualization Studio, n.d.).

Landsat tuvo su primer lanzamiento en 1972 y desde entonces ha sido ícono en la imagenología multiespectral, proporcionando un registro riguroso y basto sobre los cambios en la superficie terrestre de nuestro planeta. La importancia de este programa radica en la creación de registro de datos sólido, confiable, coherente y equilibrado que ya cuenta con más de 50 años de registros constantes. Este registro constituye el mayor repositorio sobre la superficie terrestre jamás creado. El hecho de presentar un registro continuo e ininterrumpido, permite a los investigadores poder

percatarse de causas y consecuencias de múltiples fenómenos físicos, permitiendo a los mismos poder estudiar y prevenir o prever muchos desastres (NASA Scientific Visualization Studio, n.d.).

Aunque, la importancia de sus continuos y extensos registros son invaluables, la fidelidad de estos es un factor decisivo, puesto que, la consistencia y calibración de los instrumentos de medición es uno de los desafíos más grandes a los que se ha tenido que enfrentar este proyecto. Es por esto que este programa hace un importante énfasis en la calibración radiométrica y geométrica de sus instrumentos (U.S. Geological Survey, n.d.).

Como ya se mencionó Landsat es un instrumento con historia, es por ellos que ha habido distintas versiones de los satélites que han orbitado la tierra desde su inicio, el día de hoy, existen 9 versiones de programa Landsat cuyo único objetivo ha sido mejorar la información obtenida por los satélites, van mejoras desde aumentar la calidad de los instrumentos, agregar sensores nuevos y reemplazos por fallas catastróficas con el anterior (U.S. Geological Survey, n.d.).

CAPÍTULO 7 EXPERIMENTACIÓN Y RESULTADOS

Desarrollando paso a paso la metodología Knowledge Discovery in Databases (KDD) con la finalidad de descubrir información acerca del estudio. Esta metodología ampliamente utilizada para el descubrimiento de información ayudará a poder guiar el proceso desde la obtención de la información hasta el manejo de resultados. La finalidad de esta metodología será evaluar el desempeño teórico y práctico de los algoritmos de agrupamiento no supervisado para clasificar información de imágenes multiespectrales correspondientes a la región del estado de Hidalgo para la identificación de zonas de sequía y sus tipos.

En este sentido, se requiere aplicar una serie de pasos para guiar el conocimiento, los cuales serán detallados a continuación.

7.1 Selección y adquisición de datos

En esta fase se busca recopilar datos a partir de los cuales el conocimiento será moldeado, esta información debe ser recolectada de fuentes altamente confiables, consistentes, densas y con la menor cantidad de ruido y huecos posible. La calidad de la información es fundamental para resultados fidedignos y reales.

7.1.1 Delimitación Y Caracterización Del Área De Estudio

Antes de poder recopilar los datos se deben establecer límites sobre el estudio que se realizará. Se pueden marcar límites temporales, espaciales y en los mismos datos. Entre más delimitada sea la información, esta será más densa al igual que más precisa. Lo primero antes de recolectar datos, debe ser definir el universo mismo de los datos estableciendo así sus bases de la investigación.

7.1.1.1 Delimitación Del Espacio Físico Y Temporal

En el espacio físico se hará una obtención de imágenes multiespectrales para el territorio comprendido por el estado de Hidalgo, esto es debido a el conocimiento de la región y a el alto nivel de sequía vivido durante los últimos años. Este estado es ideal para guiar esta investigación debido a que no existen hechos de este tipo. Aunque hay muchos trabajos sobre el estudio de la sequía, no ha habido alguno que proponga clasificar la sequía mediante imágenes multiespectrales en el estado de Hidalgo. Esta brecha permite realizar una investigación novedosa y de alto impacto para esta región

vulnerable ante este fenómeno. Generando conocimiento que pueda generar nuevas investigaciones y formas de mitigar este fenómeno.

En el espacio temporal se seleccionará fechas durante las cuales, la región estudiada sufrió una de sus mayores crisis hídricas en la última década, esto de acuerdo a MSM, son las fechas de:

- 20 De febrero del 2024
- 05 de Julio del 2024
- 29 de Julio del 2024
- 15 De octubre del 2024

Al tener 4 puntos de referencia permite a la investigación tener un marco referencial amplio y capaz de detectar distintos tipos de patrones de sequía. La idea de utilizar fechas con un alto nivel de sequía es que al evaluar el desempeño de los distintos métodos de clustering, existen más datos agrupables para manejar y de los cuales los algoritmos pueden encontrar información útil y no solo generar ruido con estos mismos.

7.1.1.2 Criterios Para La Selección De Datos Satelitales Landsat 8

En el uso de las bandas de las imágenes multiespectrales, se realizó una delimitación a únicamente utilizar las bandas 4 (Rojo), 5 (Infrarrojo Cercano) y 6 (Infrarrojo de onda corta), estas bandas son comúnmente utilizadas para este tipo de estudio por su relación con los fenómenos que muestran altos niveles de sequía.

El espectro rojo es esencial, ya que es comúnmente utilizado para identificar la luz roja visible e identificar la vegetación y la cobertura del suelo, la banda 5 captura la energía del infrarrojo cercano y comúnmente estudia la vegetación y la biomasa de la misma, también ha sido utilizado para medir los cuerpos de agua, por último la banda 6 mide el infrarrojo de onda corta, lo que permite poder ser utilizado para conocer la humedad de la vegetación así como las nubes y áreas vulnerables a la sequía.

Los satélites Landsat permiten obtener imágenes a través de niveles, esto ayuda fuertemente a eliminar el ruido dependiendo el nivel de solidez que se requiera en los datos, para esta investigación se utilizara el nivel 2, este goza de tener correcciones atmosféricas, permitiendo tener una imagen más clara sin ruido efectos atmosféricos. De esta forma se obtienen resultados que respetan las condiciones reales de la superficie.

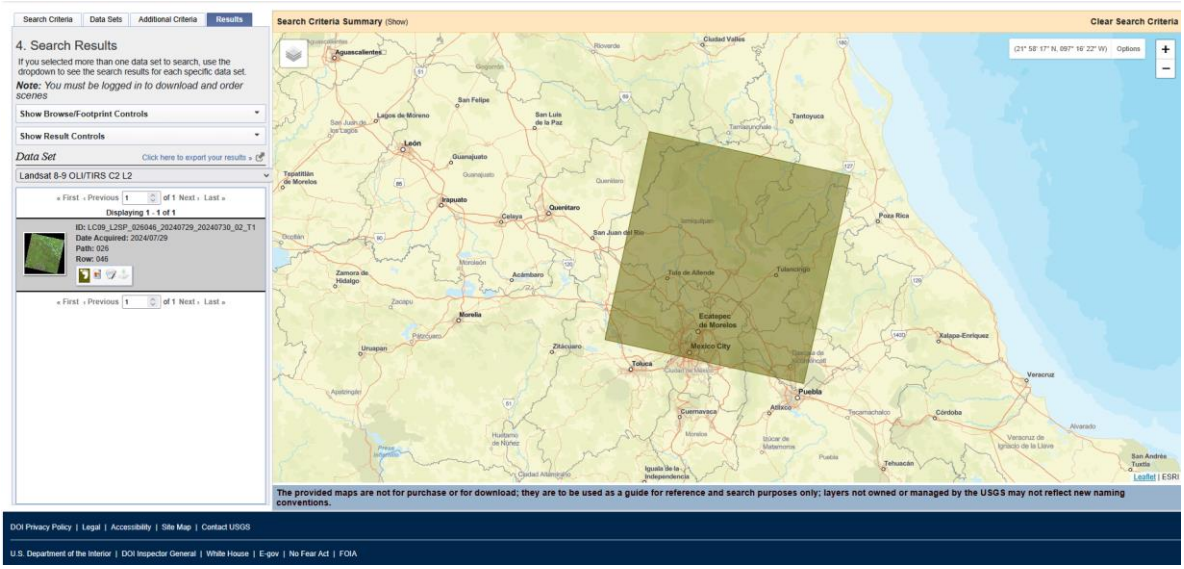

Figura 4 Sistema USGS para la obtención de Imágenes Multiespectrales

7.1.2 Proceso De Adquisición De Datos

Este proceso se realizará a través de la USGS, agencia encargada de proveer imágenes multiespectrales y mapas cartográficos de la superficie terrestre. A través de esta agencia se obtendrán imágenes de los puntos en el tiempo marcados con anterioridad y de la misma forma se realizará la categorización correcta. Interfaz mostrada en la figura 4 sobre la configuración del sistema para la obtención de las imágenes.

Los parámetros para poder obtener las imágenes dentro de esta investigación son los siguientes:

- Nivel de procesamiento: L2SP
  Este parámetro dicta el nivel de procesamiento que las imágenes obtendrán al ser exportadas, se ve puesto en dos para poder eliminar el ruido atmosférico de las imágenes, esto incluye la reflectancia superficial y la temperatura de la superficie. Este parámetro es lo que se conoce como “Listo para el análisis”.
- Ubicación Geográfica: 026046
  Esta ubicación es la correspondiente al estado de Hidalgo en mayor parte de su territorio. Dentro del formato de Landsat, corresponde al sistema de referencia mundial 2 (WSR-2) con path 026 y fila 046
- Versión de colección: 02

Esta colección refiere a las mejoras más recientes en procesamiento, lo que incluye una mejora en la precisión geométrica y radiométrica. La métrica seleccionada corresponde a la colección de datos 02 de Landsat.

- Nivel de calidad: T1
  Este parámetro dicta la designación de calidad, garantiza que las imágenes tienen una excelente precisión geométrica con errores en un margen de 12 metros y adecuadas para análisis científicos y series temporales. Esta categoría enmarca la calidad como de Tier 1 o de alta calidad.
- Satélite: LC08 y LC09
  Enmarca la versión del proyecto y satélite del que se obtendrán los datos, para la resolución del proyecto utilizara Landsat 8 y 9, ambos son altamente compatibles con la diferencia de que el satélite Landsat 9 posee sensores de una resolución radiométrica superior lo que permite obtener imágenes con una mayor atención a las variaciones.
- Fechas de adquisición y su imagen correspondiente:
  - 20 de febrero del 2024, Figura 5
  - 05 de julio del 2024, Figura 6
  - 29 de julio del 2024, Figura 7
  - 09 de octubre del 2024, Figura 8

Estas fechas aseguran puntos clave de sequía de acuerdo a la MSM donde la sequía alcanzó puntos críticos en el estado de Hidalgo, asegurando un proceso de clustering eficiente y con un margen amplio de datos a tomar.

Como se puede observar en las imágenes obtenidas mediante el programa Landsat, se observa una disminución considerable de vegetación durante las dos fechas escogidas para el estudio en las figuras 5 y 6 estas imágenes correspondientes a febrero y la primer semana de julio, demuestran también, un nivel bajo en los cuerpos de agua, mientas que, por su parte, las figuras 7 y 8, correspondientes a la segunda semana de julio y a octubre, muestran un aumento considerable de la vegetación y aumento en los cuerpos de agua considerable.

Estas figuras muestran una realidad en el país y en el estado de Hidalgo, donde la sequía aparece de forma significativa y severa durante largos periodos de tiempo. Y a partir de la segunda semana de julio, esta repentinamente desaparece sin dejar rastro de aquellas zonas verdaderamente áridas, aumentando considerablemente los cuerpos de agua y la vegetación visible por las imágenes.

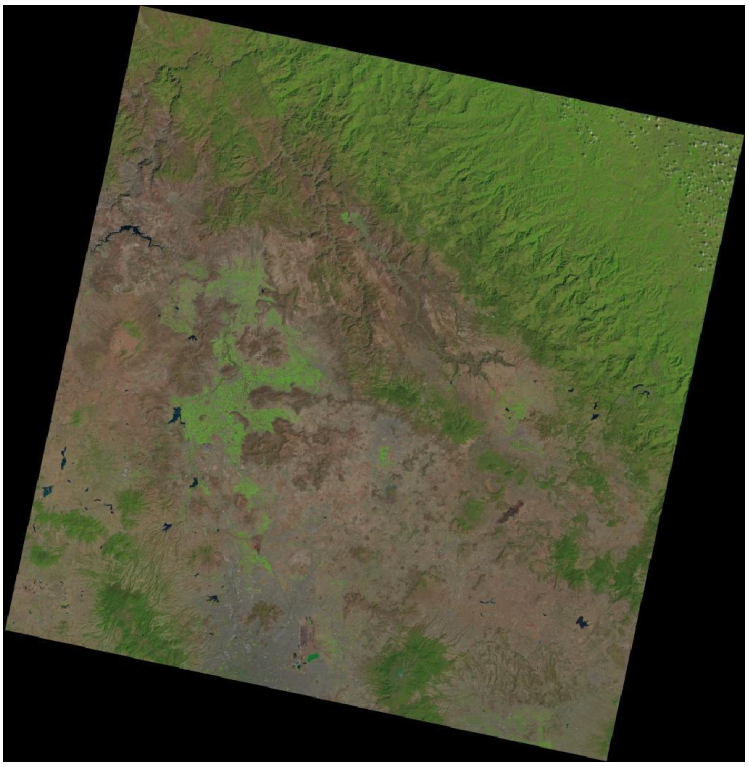

Figura 5 Imagen Obtenida del USGS Correspondiente a febrero

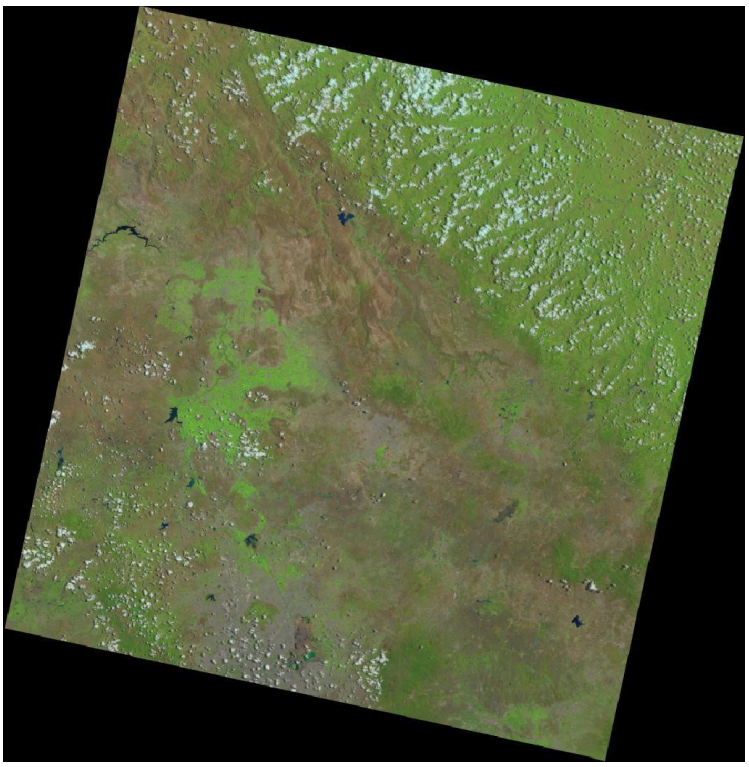

Figura 6 Imagen Obtenida del USGS Correspondiente a Julio

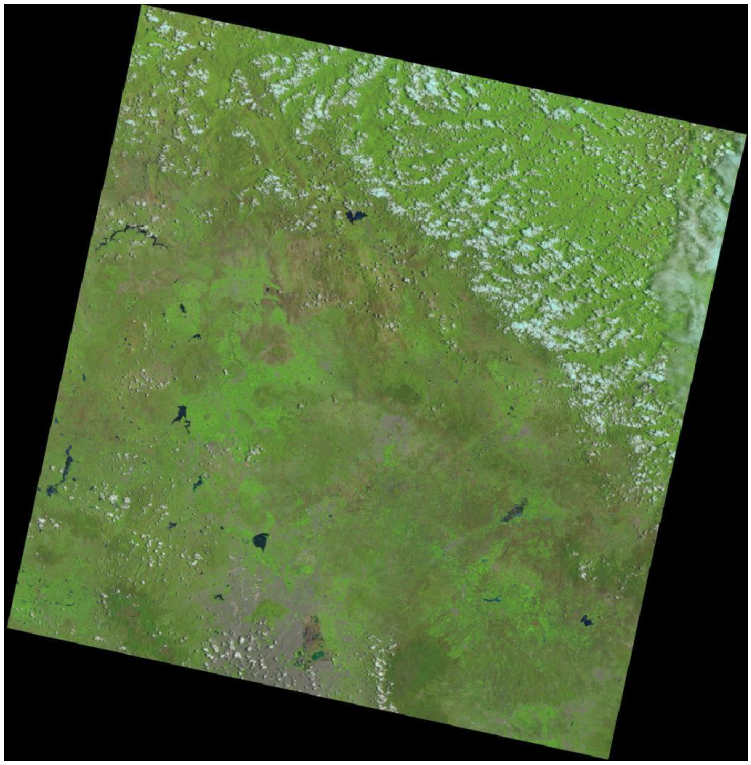

Figura 7 Imagen Obtenida del USGS Correspondiente a julio

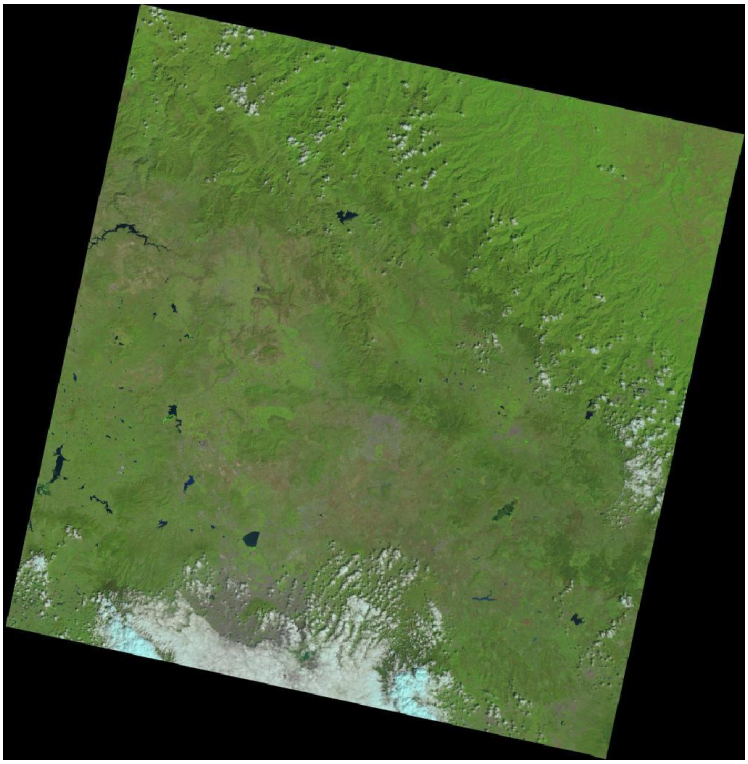

Figura 8 Imagen Obtenida del USGS Correspondiente a octubre

7.2 Pre Procesamiento Y Acondicionamiento

Una vez obtenidos y descargados datos de las imágenes multiespectrales en formatos TTF, ahora se escogen las bandas necesarias para poder aplicar los algoritmos de forma que así lo dictan las limitaciones.

7.2.1 Verificación De Calidad Y Extracción De Bandas Multiespectrales

Cada una de las imágenes espectrales ejecutadas arrojó un número delimitado de imágenes TTF las cuales contienen una sola banda multiespectral por imagen y un archivo de metadatos. Como primer paso de la verificación de integridad, se hace una inspección visual sobre los metadatos corroborando que los parámetros establecidos son los solicitados por la investigación. Dentro de esta inspección de meta data, se debe evaluar la fecha de adquisición, de procesamiento, el nivel de procesamiento, etc.

Una vez hecha la inspección y corroboración de los datos, las imágenes TTF se extraerán los datos en forma de texto como se muestra en la figura 9, una vez hecho, se separarán las bandas dejando únicamente aquellas que son necesarias para la investigación, para este estudio, se dejarán las bandas 4 (Rojo), 5 (Infrarrojo), 6 (Infrarrojo de onda corta). Dando por entendido que las demás bandas no se incluyen en el agrupamiento. Usar estas tres bandas juntas permite observar la vegetación y si estado de salud basados en como absorben o rechazan los espectros de la luz gracias a las Bandas 4 y 5, también se requiere observar los cuerpos de agua y el volumen que estos contengan con ayuda de la banda 6. Esta combinación es crucial para la identificación de la sequia en base a bandas multiespectrales.

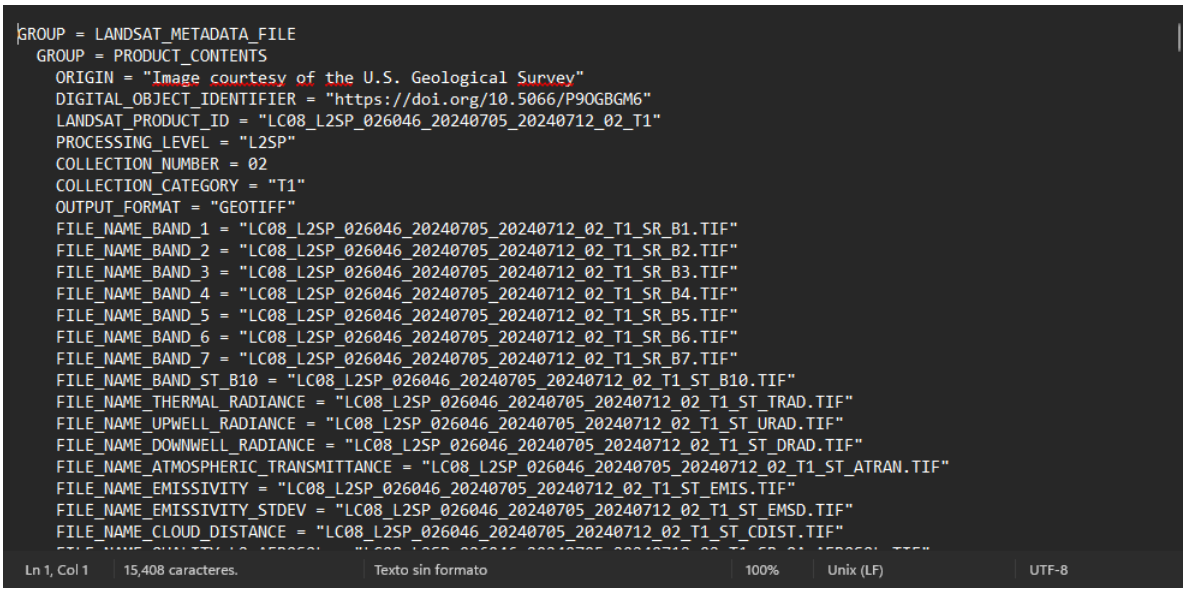

Figura 9 Representación en texto de una imagen del programa Landsat

7.3 Transformación Y Extracción De Características

Esta fase de la metodológica es la encargada de armar la base de datos en datos que los algoritmos puedan leer y a los que se puedan ser aplicados. Estas bases de datos deben ser transformados en un formato estructurado y enriquecido, construyendo un espacio con características multidimensionales y que sean los más separables posible.

7.3.1 Construcción Del Vector De Características Por Pixel

Para poder construir la base de datos utilizable, se creará un vector de características, donde cada píxel de la imagen se convertirá en un arreglo de datos correspondientes a las características necesarias para la investigación. Dicho de otra forma, cada pixel de la imagen, contiene todas las bandas multiespectrales obtenidas por el programa Landsat, las cuales deben ser separadas para solo contener las estudiadas, una vez que todos los pixeles han sido convertidos en arreglos numéricos, se agregan de forma consecutiva en un archivo en el formato adecuado (para csv cada banda se divide por comas y cada objeto se separa con un salto de línea) y el conjunto de todos los objetos es el vector de características. Para este caso de estudio, cada pixel se dividirá en un arreglo de tres métricas, estas corresponden a las bandas multiespectrales ya delimitadas en esta investigación, la cuales son:

- Banda 4 (Rojo): Es este bando, aportará a los datos la cantidad de radiación de este espectro que es absorbida por la clorofila por medio de la fotosíntesis, lo que significa,

que un aumento en esta banda, simboliza una baja función fotosintética por parte de la vegetación.

- Banda 5 (Infrarrojo Cercano): Esta banda demuestra la estructura celular interna de las hojas de una vegetación sana, lo que refleja fuertemente un infrarrojo cercano, Una disminución en esta banda puede representar estrés vegetal, deshidratados o la reducción de biomasa.

- Banda 6 (Infrarrojo de Onda Corta 1): Esta banda demuestra el contenido de los cuerpos de agua, gracias a que este espectro es altamente sensible para detectar este fenómeno, puesto que el agua absorbe la radiación en esta longitud de onda. Siendo que un aumento en esta banda simboliza una disminución en los cuerpos de agua.

7.3.2 Normalización De Los Datos

Una vez comprendidos y vectorizados todos los datos es necesario normalizarlos a un modo donde los algoritmos de agrupamiento no puedan ser influenciados por las medidas anormales de las distancias, manteniendo un orden y una proporción equidistante con estos datos. Para solucionar este problema, cada una de las bandas requieren ser aplicadas técnicas de re escalado de min-max, que re escalan cada característica a un rango común de los datos proceso mostrado en la figura 10.

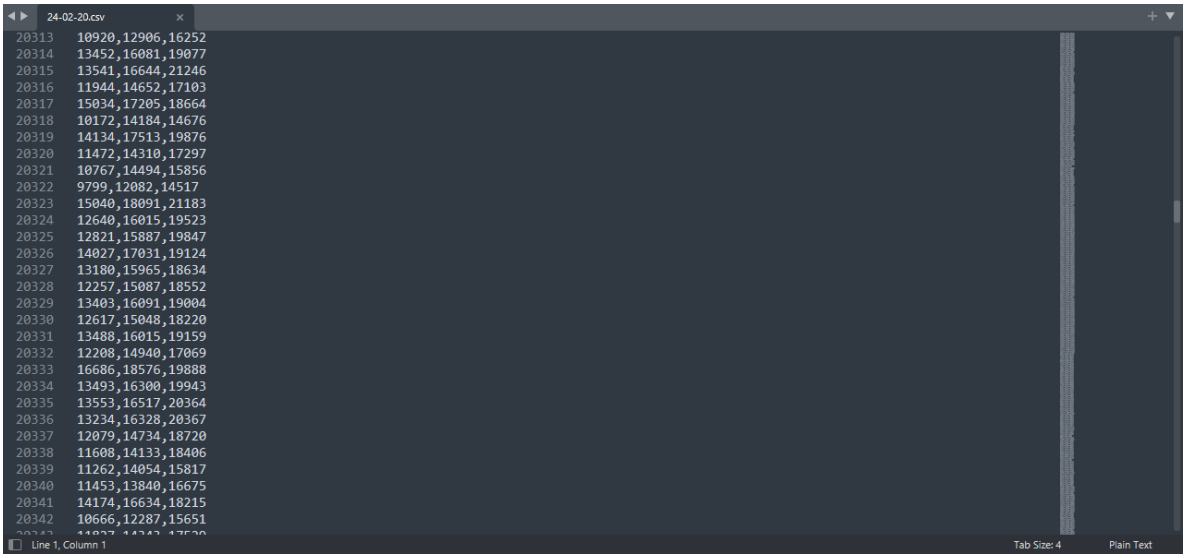

Figura 10 Representación Numérica de los Datos ya Vectorizados

7.4 Minería de Datos. Aplicación de algoritmos

Esta es la fase más importante de toda esta metodología, es donde se ejecutarán diversos algoritmos de minería de datos mediante los que se espera obtener información importante u oculta de los datos. Para esto se hará uso de una herramienta de minería de datos, donde puedan ser cargados los conjuntos y devueltas las agrupaciones de una forma simple y sencilla. Esta herramienta será WebMinerX, una herramienta que usa JavaScript y Python con scikit-learn para poder ejecutar los algoritmos correctamente.

7.4.1 Fundamentos Para La Segmentación De Imágenes Satelitales

Como se revisó en el estado del arte, los algoritmos de agrupamiento están fuertemente ligados para identificar fenómenos relacionados con la meteorología y con las imágenes multiespectrales, muchos de estos estudios se basaron en diferentes algoritmos siendo K Means el usado por excelencia, aunque, también se estudiaron algoritmos como el método jerárquico. Por lo que existe la posibilidad de encontrar un algoritmo que mejore la eficiencia partición de las imágenes multiespectrales y así facilitar el agrupamiento no dependiendo completamente de expertos para poder generar este conocimiento.

Al haber tratado los datos como vectores, es posible generar agrupaciones donde grupos con características similares pueden ser generados con relativa simplicidad, ahora es dependiente del algoritmo encontrar esas distancias y generar esos grupos. También es importante mencionar que la naturaleza de cada algoritmo mencionado afectará el resultado y la agrupación. Ya que se escogieron 3 algoritmos que representan un tipo de agrupamiento no supervisado distinto, los resultados esperados entre ellos, no deberían tener similitudes entre sí.

7.4.2 Implementación Del Algoritmo K Means

K Means es el algoritmo con un mayor número de usos en materia de clustering para la identificación de fenómenos meteorológicos con el uso de índices y de imágenes multiespectrales. Su uso es sencillo y altamente eficiente lo que lo hace adecuado altamente escalable con grandes conjuntos de datos, ideal para esta investigación (Jovic, 2014).

7.4.2.1 Parámetros Utilizados

- Distancia: Euclidiana
  Se utilizará la distancia euclidiana, para medir la disimilitud entre los vectores de características de los píxeles.
- Número de Clusters: 6
  Es el número requerido de acuerdo al NNDI Y a la MSM
- Inicialización de clusters: Objetos iniciales
  Para poder hacerla repetible y replicable, no se utilizó ningún tipo de semilla o inicialización aleatoria de los centroides.

7.4.2.2 Ejecución Del Algoritmo

La ejecución de este algoritmo es bastante sencilla, la herramienta representada en la figura 11, una vez seleccionada la carga inicial, donde una vez cargado el archivo con los vectores, este sistema permite hacer alguna transformación de los datos además de mostrar una previsualización de los mismos además de permitir un básico formateo de eliminación de caracteres o columnas según sea el caso.

Una vez cargado el archivo y estando de acuerdo con la configuración de los datos, se procede con la configuración de parámetros. La herramienta ya tiene parámetros predefinidos como el seed y la distancia. Aunque los parámetros de configuración permiten elegir la inicialización y número de centroides, únicamente se indica el número de centroides en 6 e inicialización no aleatoria.

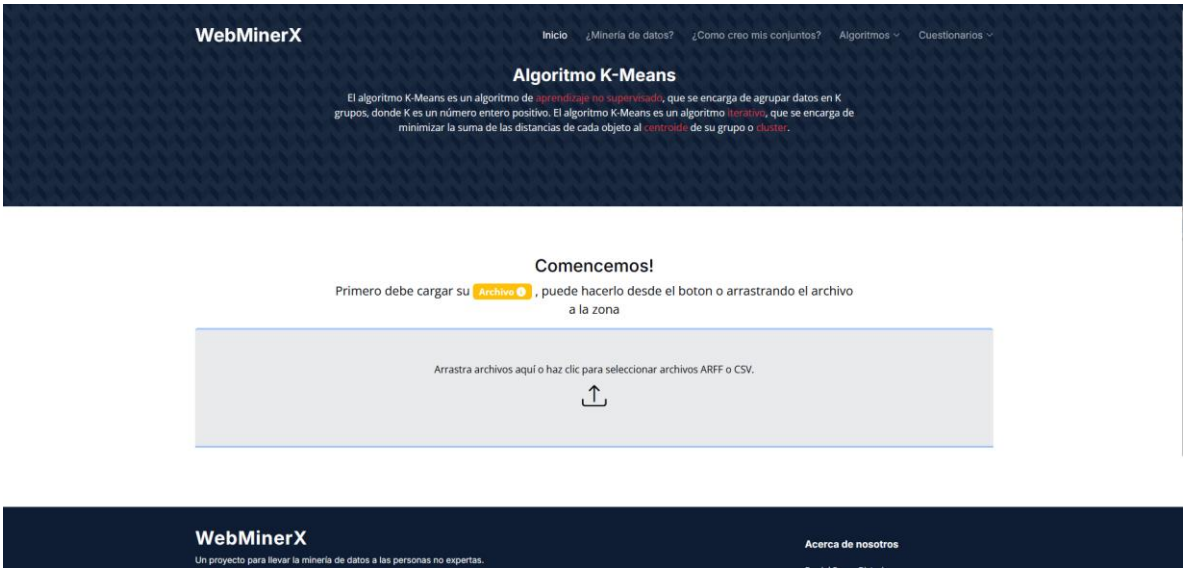

Figura 11 Sistema WebMinerX, Ejecución del Algoritmo K Means

Cuando la Iteración ha terminado, entonces el sistema dará una visualización de los datos y preguntará al usuario sobre si guardar los datos. Se guardan y solo una vez guardada la agrupación y etiquetada con la fecha correspondiente, se visualizará la siguiente fecha y así sucesivamente hasta obtener los 4 grupos de k means.

7.4.3 Implementación Del Algoritmo Jerárquico

El algoritmo jerárquico es uno de los métodos más comunes en métodos de agrupamiento, este algoritmo también tiene usos dentro de la rama del clustering de imágenes multi espectrales. Su característica principal, es, que, aunque se defina previamente. Este realizará todas las uniones en forma de árbol invertido hasta lograr la unidad de clustering. Aunque de una forma similar a K means, su salida será únicamente de los clustering requeridos (Zwass, 2025).

El objetivo de este algoritmo es poder generar uniones desde lo más general hasta la unidad, lo que permite crear una serie de uniones coherente y visibles en un gráfico final llamado dendrograma. Una de sus desventajas es el gran coste computacional que este algoritmo tiene, lo que muchas veces lo hace muy difícilmente escalable.

7.4.3.1 Parámetros Utilizados

- Distancia: Euclidiana
  Se utilizará la distancia euclidiana, para medir la disimilitud entre los vectores de características de los píxeles.
- Número de Clusters: 6
  Es el número requerido de acuerdo al NNDI Y a la MSM
- Método de enlace: Ward
  Este método fusiona el par de clústeres que resulta en el menor aumento de la varianza total intra clúster. Tiende a producir clusters compactos y de tamaño similar, lo que a menudo es deseable en la segmentación de imágenes de cobertura terrestre.

7.4.3.2 Ejecución Del Algoritmo

El procedimiento es sumamente similar al anterior, una vez cargados los archivos correspondientes a la primera fecha, se elige únicamente el número de clusters. El sistema por sí mismo está cargado con los demás parámetros para una ejecución repetible y eficiente como se muestra en la figura 12. Normalmente el sistema emite

un dendrograma, debido al enorme trabajo computacional, se saltó esta función dentro del sistema.

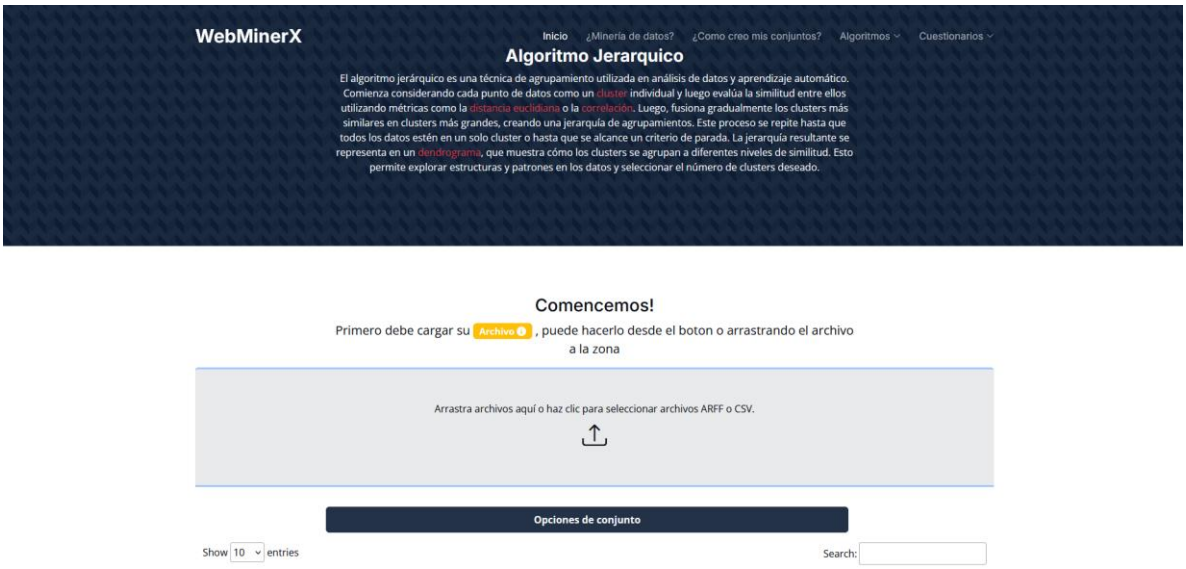

Figura 12 Sistema WebMinerX, Ejecución del Algoritmo Jerárquico

7.4.4 Implementación Del Algoritmo DBScan

Destiny Based Spatial Clústering of Applications with Noise (DBScan) ofrece un enfoque distinto a los anteriores, este se basa en generar grupos basados en poblaciones de datos. Las zonas densamente pobladas serán aquellas que formen un grupo. Este no requiere de un número de clusters, el algoritmo se basa en los parámetros necesarios para determinar por sí mismo el número óptimo de grupos

De ser necesario generará una agrupación adicional de datos anormales o atípicos, ideales para investigaciones donde se requiere analizar anomalías (Zwass, 2025).

7.4.4.1 Parámetros Utilizados

- Distancia: Euclidiana
  Se utilizará la distancia euclidiana, para medir la disimilitud entre los vectores de características de los píxeles.
- Mínimos de puntos: 500
  Este número de puntos se establece a que, debido al tamaño de los datos, no se permiten agrupaciones pequeñas o que no aporten información
- Épsilon: 12500
  Esta variable controla la distancia máxima que pueden tener dos puntos para

pertenecer a un mismo grupo, permitiendo así poder cuestionar o alejar puntos de una agrupación. Este número se obtuvo gracias a repetidas iteraciones para conseguir el número requerido de puntos necesarios.

7.4.4.2 Ejecución Del Algoritmo

El procedimiento es similar al anterior, para la carga de archivos, la diferencia radica en la selección de las variables, donde, a diferencia de algoritmos de k means o DBScan, en este algoritmo se selecciona el mínimo de puntos para formar un grupo y la épsilon, el cual es la distancia máxima para pertenecer a un grupo como se muestra en la figura 13.

Algo que hay que mencionar de este algoritmo, es que, al no tener datos o registros de ejecuciones pasadas, este algoritmo tuvo que ejecutarse n cantidad de veces hasta lograr parámetros que satisfagan las necesidades de la investigación.

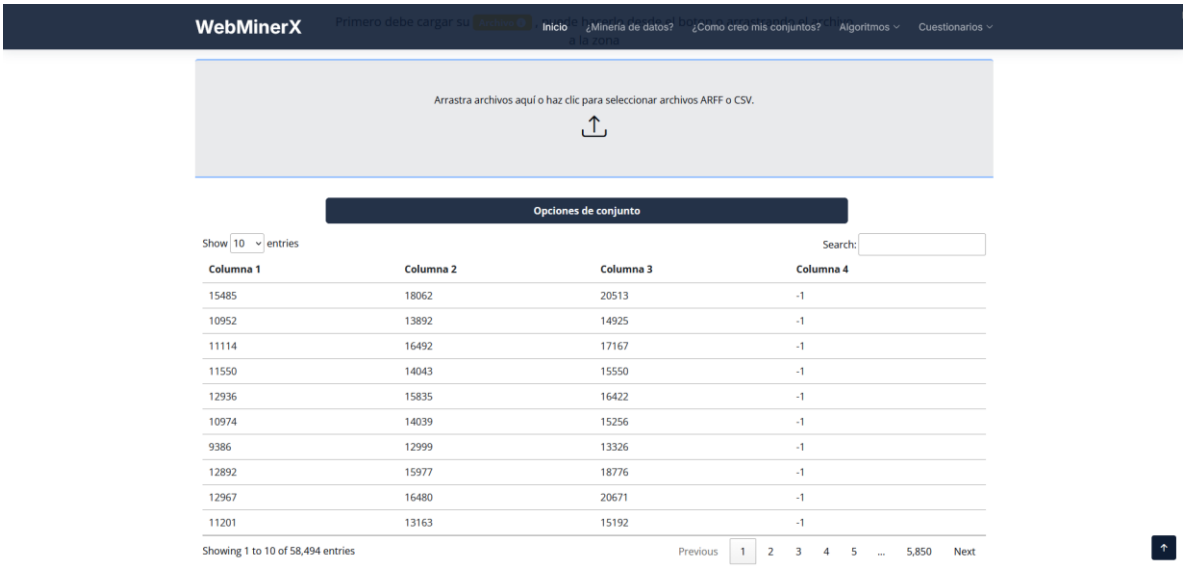

Figura 13 Sistema WebMinerX, Ejecución del Algoritmo DBScan

7.5 Evaluación, interpretación y visualización de resultados

Dado que el agrupamiento es una forma de aprendizaje no supervisado, no se dispone de una muestra de la cual comparar directamente los resultados. Por lo tanto, es necesario utilizar métricas de validación interna que evalúen la calidad de la partición de los datos basándose únicamente en la separación, o unión de los datos. Ya sea la

separación de los datos con otros clusters o la unión de los datos dentro del mismo clúster.

A continuación, se harán pruebas con distintos índices de validación para evaluar la calidad de cada una de las agrupaciones generadas para cada uno de las fechas marcadas, generando así tres validaciones por cada fecha por cada algoritmo de los datos.

7.5.1 Análisis Del Coeficiente De Calinski Harabasz

Este índice de validación también es conocido como razón de varianza, basa su evaluación de calidad en que tan densos y separados entre sí están los clusters, este índice define la razón de la dispersión inter clusters e intra clusters. Su evaluación resulta en una métrica donde un mayor valor representa una menor dispersión de los datos y viceversa (Fayyad, 1996).

La aplicación propuesta para realizar los agrupamientos, permite realizar estas evaluaciones con base a múltiples índices, incluyendo Calinski Harabasz entre su repositorio. A lo cual, de la misma forma que se vino trabajando, se usará para la ejecución del índice. Su proceso es similar, como se muestra en la figura 14, requiere una carga inicial de datos previamente etiquetados, estos datos deben ser identificados con un clúster identificado como número entero en la última columna de los datos. Una vez cargados los datos, se podrán hacer arreglos básicos de formato en la tabla de vista previa, en caso de requerirse. El último paso es únicamente ejecutar el algoritmo y copiar la métrica deseada.

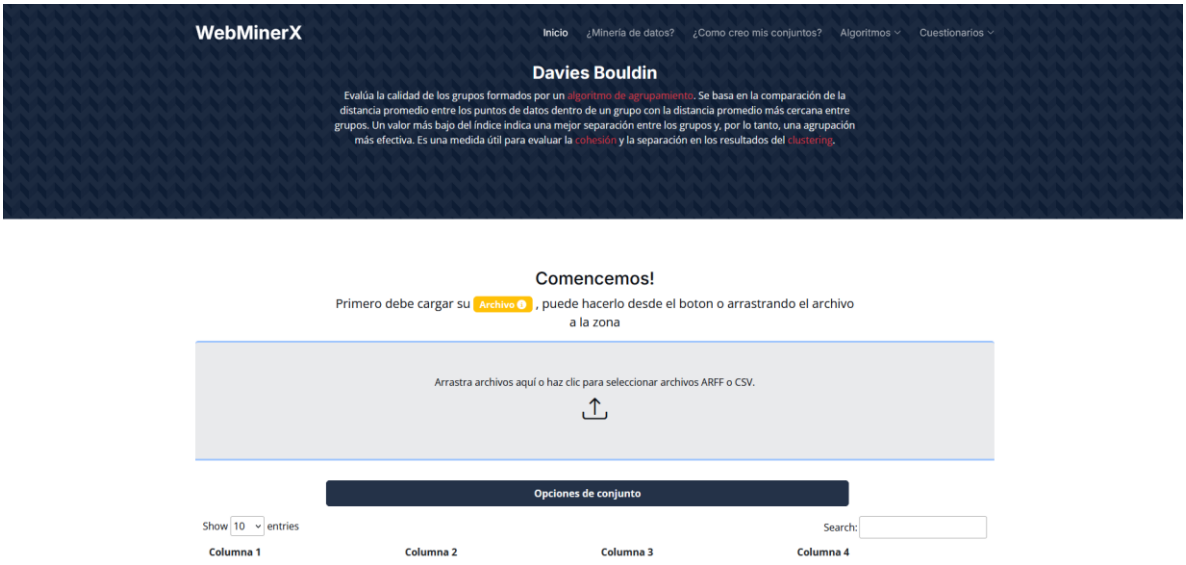

Figura 14 Sistema WebMinerX, Ejecución del Índice Davies Bouldin

7.5.2 Análisis Del Coeficiente De Davies-Bouldin (DBI)

El índice de Davies-Bouldin evalúa la calidad del agrupamiento en base a la relación entre la dispersión dentro de los clusters y la separación entre ellos. Esta métrica se define a sí misma como el promedio de todos los clusters de la similitud con su clúster más similar. Esta métrica se basa en la calidad donde un número más bajo indica un mejor agrupamiento, lo que indica que son más compactos (Fayyad, 1996).

La ejecución de este algoritmo es sencilla, se realiza la carga inicial de los datos y de la misma forma que el trabajo se ha realizado. La aplicación definirá métricas estándar para la ejecución del algoritmo y arrojará la métrica deseada representado con la figura 15.

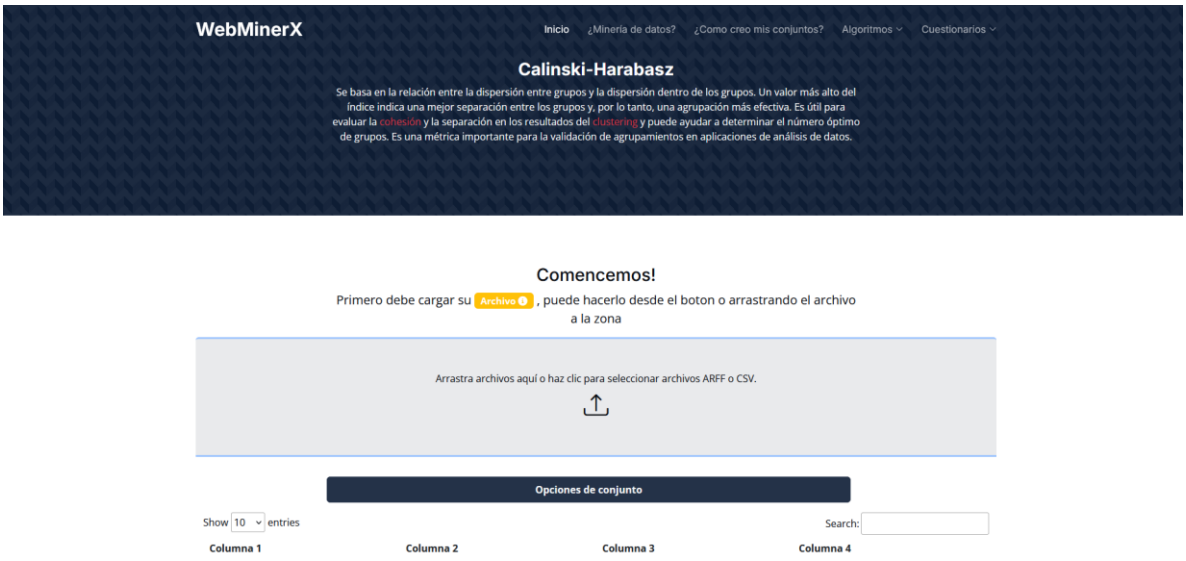

Figura 15 Sistema WebMinerX, Ejecución del Índice Calinski Harabasz

7.5.3 Análisis Del Coeficiente De Silhouette

El coeficiente de Silhouette es un índice de calidad que evalúa la cohesión intra clúster así como la separación Inter clúster, sus resultados se dan en un rango entre -1 y 1, donde un índice menor a cero indica una cohesión fallida y deficiente, mientras que un número mayor a 1 da como resultado un agrupamiento útil y con mayor precisión mientras el número aumente (Piatetsky-Shapiro, 2012).

Nuevamente se cargan los clusters dentro de la aplicación y se ejecuta el algoritmo, de la misma forma que muchos índices, estos no requieren de parámetros especiales,

además de los embebidos, así como los algoritmos anteriores fueron ejecutados y mostrado en la figura 16.

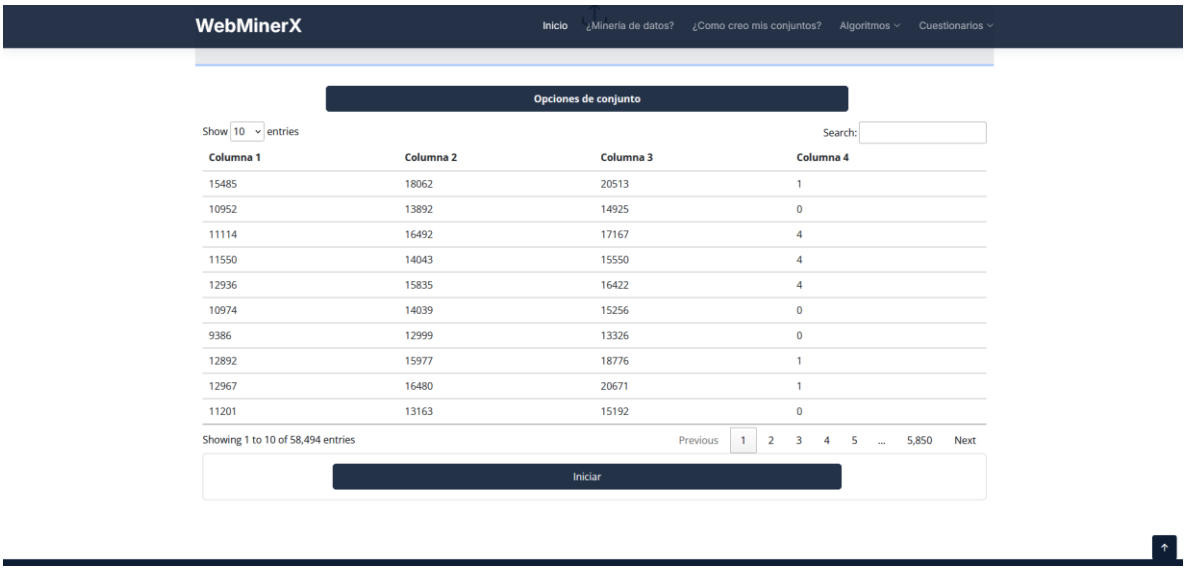

Figura 16 Sistema WebMinerX, Ejecución del Índice Silhouette

7.5.4 Interpretación De Los índices

Una vez realizados los índices de validación, estos se almacenaron en una tabla comparativa, donde cada uno de los algoritmos con cada una de las fechas generan un índice y a su vez, estos deben ser evaluados, de acuerdo a las reglas de cada índice.

Esto resume en 36 evaluaciones correspondientes 12 evaluaciones de cada índice distinto al cual pertenecen 3 validaciones por cada fecha. De esta forma se obtiene que de acuerdo a Davies Bouldin la mejor agrupación unánimemente es DBScan (aunque hay cosas a tomar en cuenta). Para Calinski Harabasz su mejor resultado fue K Means en 3 de las cuatro fechas, la cuarta fue mejor valorada para el algoritmo Jerárquico. Por último, para el índice Silhouette, existen opiniones divididas, donde para la fecha 1 y 3 DBScan lleva la mejor métrica y las fechas 2 y 4 son mejor valoradas por el algoritmo Jerárquico. Como se observa en la tabla 2.

Obteniendo de esta forma que DBScan lleva el 50% de conjuntos denotados como mejor valorados, mientras que el algoritmo jerárquico lleva el 25% y K Means obtiene el 25% restante.

Tabla 2 Resultados de la Validación de los Índices Sobre las Agrupaciones

Algoritmo	Índice	24-02-20	24-07-05	24-07-29	24-10-15
K Means	Davies	0.8751171	0.8347362	0.8317217	0.7893732
Jerárquico		1.0008542	0.8739486	0.8571458	0.8268167
DBSCAN		0.2178277	0.3121389	0.1511411	0.5181221
K Means	Calinski-Harabasz	111116.0150303	80037.2785778	68876.6968198	47047.2466717
Jerárquico		39036.1934555	57491.2082037	68137.9684877	95684.8491332
DBSCAN		14154.5002342	3950.4359654	1020.3554882	1427.9242355
K Means	Silhouette	0.3478245	0.3469972	0.3671854	0.3808788
Jerárquico		0.3063048	0.2926747	0.3382776	0.3103607
DBSCAN		0.2178277	0.3065766	0.3147501	0.4475070

Algo que hay que tomar en cuenta para el algoritmo de DBScan es que, aunque los grupos fueron satisfechos, el ruido fue un factor decisivo, el cual no se incluyó en los índices, por lo que, para concluir si estas agrupaciones son realmente útiles, se requiere más investigación antes de usar este algoritmo como un algoritmo confiable. A lo cual, se requiere hacer ajustes en la tabla de resultados de los coeficientes, eliminando al algoritmo DBScan de la tabla, para dejar así algoritmos con resultados útiles y completos. Dando como resultado la tabla 3.

Tabla 3 Resultados de los Índices Excluyendo DBScan

Algoritmo	Índice	24-02-20	24-07-05	24-07-29	24-10-15
K Means	Davies	0.8751171	0.8347362	0.8317217	0.7893732
Jerárquico		1.0008542	0.8739486	0.8571458	0.8268167
K Means	Calinski-Harabasz	111116.0150303	80037.2785778	68876.6968198	47047.2466717
Jerárquico		39036.1934555	57491.2082037	68137.9684877	95684.8491332
K Means	Silhouette	0.3478245	0.3469972	0.3671854	0.3808788
Jerárquico		0.3063048	0.2926747	0.3382776	0.3103607

En esta nueva evaluación de los grupos se destaca K Means con un 58% de agrupaciones mejor valoradas, mientras que jerárquico obtiene el 42% de agrupaciones mejor valoradas, siendo que el índice de Silhouette es el único en dar una valoración unánime.

Una vez agrupados los datos y obtenidos los conjuntos etiquetados con su grupo perteneciente, estos se procesaron de nuevo para regenerar la imagen original, pero generando un espectro sobre de las agrupaciones realizadas con el mejor algoritmo, el algoritmo k means.

Las imágenes generadas de este último proceso son las figuras 17,18,19,20. Estas figuras se encuentras agrupadas con su imagen original del programa Landsat para generar el contraste original. Como se puede apreciar, algunas deben valores diferentes y la agrupación no fue significativamente igual para todas.

La figura 17 muestra la agrupación más acertada identificando con valores bajos las áreas como más vegetación visible por el ojo, mientras que los grupos superiores a 5 indican una alta falta de vegetación en la región. A diferencia de la figura 19, donde se observa mayor ruido en la imagen, debido a la naturaleza de la imagen donde la sequía es menos visible, el algoritmo realizo distintos grupos de forma invertida a la figura 17, marcando con grupos bajos a zonas densamente pobladas por vegetación y grupos altos a aquellas zonas desérticas.

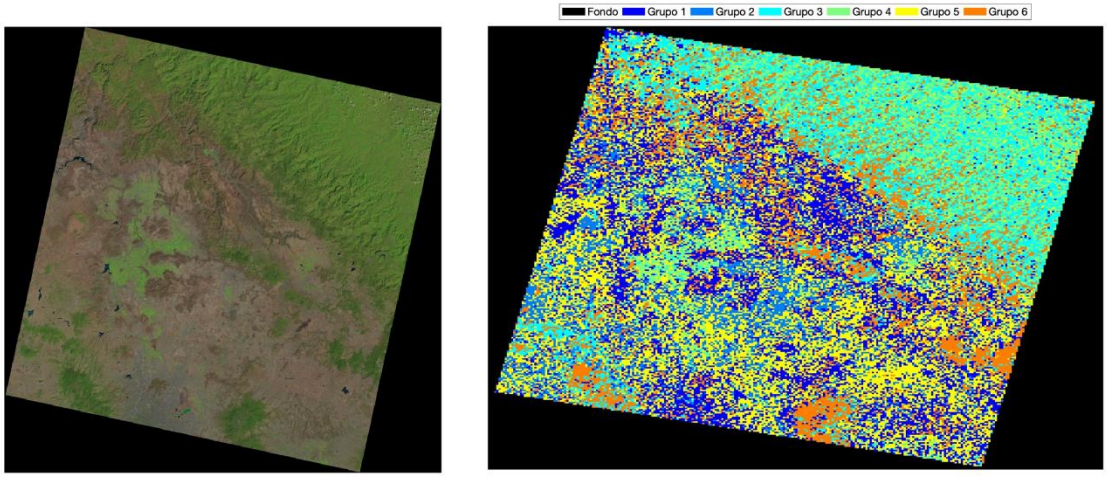

Figura 17 Imágenes reconstruidas basadas en K Means de febrero

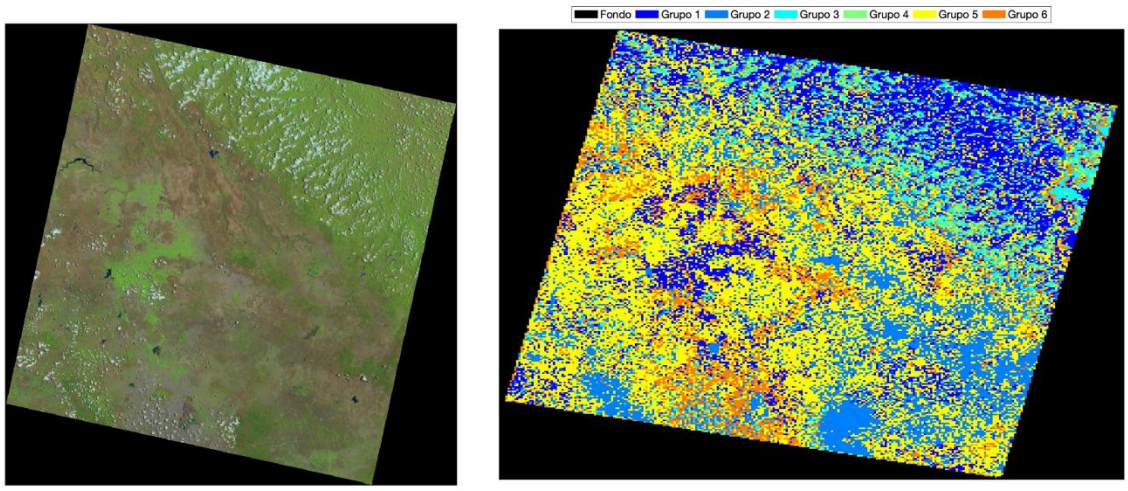

Figura 18 Imágenes reconstruidas basadas en K Means de 5 de julio

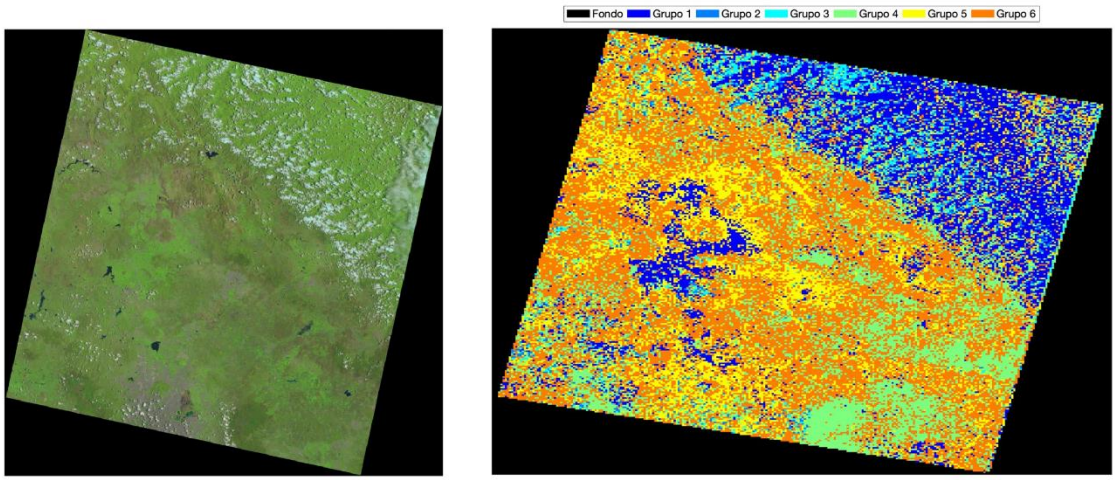

Figura 19 Imágenes reconstruidas basadas en K Means de 29 de julio

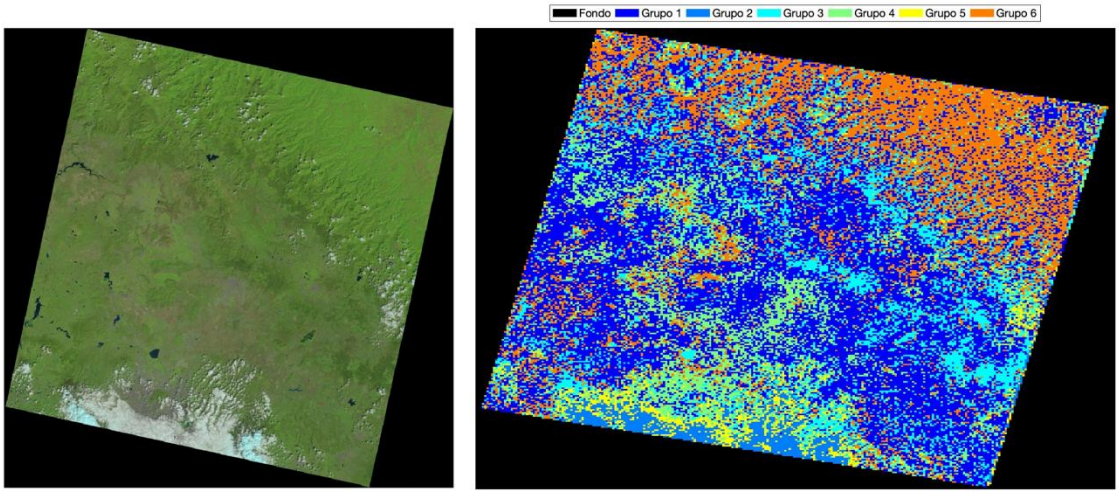

Figura 20 Imágenes reconstruidas basadas en K Means de octubre

7.6 Uso del conocimiento

Es por esto que se puede identificar que, al igual que las investigaciones pasadas lo han expuesto. K means es un algoritmo altamente eficaz, en el campo teórico, por lo menos en este caso de estudio, para sus delimitaciones y variables, en comparación a los otros dos algoritmos, seguido del algoritmo jerárquico y DBScan, algoritmo el cual se concluye, requiere más investigación antes de obtener datos y tomar sus agrupaciones como válidas y funcionales.

Con la información dada se puede inferir que la identificación de patrones de sequía en el estado se logró con distintos márgenes de error y de calidad. Esta investigación aún no tiene resultados lo suficientemente concluyentes para definir si el algoritmo k means es lo suficiente preciso para ser utilizado como herramienta sin tener grupos de muestra alimentante estudiados y sin tener una base fiable y útil.

Como muchos artículos lo indicaron, K Means es el algoritmo más utilizado para la medición de fenómenos relacionados con la sequía, lo cual fue demostrado por los índices de validación, especialmente Silhouette, el cual, indico al algoritmo k means con mejor cohesión en el mes de julio, muestra donde las imágenes reconstruidas mostraron menos diferencias y una mayor exactitud por sobre las imágenes satélites.

El uso de este algoritmo resulta factible debido a su capacidad para identificar regiones afectadas por los distintos niveles de sequía que, a su vez, afectan las bandas multiespectrales, permitiéndole al algoritmo generar una agrupación útil y de calidad.

CAPÍTULO 8 CONCLUSIONES

Durante la realización de este trabajo se buscó categorizar los distintos niveles de sequía ubicados en el estado de Hidalgo basados en distintos algoritmos de minería de datos, buscando crear una correcta categorización además de evaluar estos modelos buscando cual se desempeña mejor en esta área de estudio. Esto debido a los pocos trabajos realizados sobre este tema, principalmente para las regiones de México e Hidalgo. Este trabajo también busca ser un punto de apoyo para los expertos en materia de la sequía, sugiriendo herramientas que posiblemente puedan no ser consideradas para su trabajo y determinación, buscando que, de esta forma, no se sustituya la intervención humana. pero si se puedan tomar decisiones informadas y con nuevas herramientas que faciliten esta tarea.

Se realizó un enfoque basado en el modelo metodológico KDD mediante la aplicación de diferentes métodos de agrupamiento basados en métodos no supervisados, los cuales fueron K Means, Agrupamiento Jerárquico y DBScan ejecutados a imágenes multiespectrales del programa Landsat de distintos periodos de tiempo y evaluando cada una de las agrupaciones con índices de validación para evaluar su desempeño lógico, descubriendo así la eficiencia de cada uno para este caso de estudio.

Los resultados obtenidos, basados en los índices, demuestran una eficiencia por parte del algoritmo K Means, obteniendo una mejora en los índices de validación utilizados excepto el índice de Silhouette, puesto que este demostró un mejor desempeño para el algoritmo jerárquico. Este estudio también demostró una clara deficiencia en el algoritmo DBScan puesto que encontrar una configuración óptima de los parámetros para obtener las agrupaciones necesarias requirió un exceso de tiempo y coste computacional y, aun así, encontrando una cantidad de ruido que supera las agrupaciones hechas por el algoritmo, por lo cual, se separó de los resultados.

Ahora bien, basándose en las imágenes reconstruidas, se puede decir que el algoritmo K Means realizó un agrupamiento correcto (tomando en cuenta algunas consideraciones), puesto que las separaciones corresponden con diferencias apreciables visualmente dentro de las imágenes, lo que hay que tomar en cuenta es que, muchas veces el algoritmo destino diferentes números de clúster a regiones semejantes, es decir, en algunas imágenes asignaba números bajos del clúster para identificar regiones afectadas por la sequía, mientras que en otras, estas regiones áridas eran representadas mediante clusters altos, lo cual no debería ser un problema si esta búsqueda se hicieran en grandes volúmenes de datos, con variaciones altas entre sí. Algo que hay que añadir, es que el hecho de obtener nubes en el radar, afecta

considerablemente a la agrupación, permitiendo al algoritmo detectar grupos diferentes a los ya establecidos.

El trabajo realizado presento interrogantes sobre el uso de algoritmos de minería no supervisados como herramienta para la identificación de sequía a través de imágenes multiespectrales obtenidas a través del programa Landsat, el trabajo realizado demostró que, de forma teórica, como ya no han hecho algunos investigadores, el algoritmos K Means demostró un resultado superior a los algoritmos comparados como lo fueron el algoritmo Jerárquico y DBScan, todo esto basándose únicamente en la calidad de los resultados como distribución de los grupos y densidad de estos mismos y en el uso de múltiples índices de validación que así lo demostraron.

En el entorno practico, también demostró una clara ventaja sobre la identificación de zonas que enfrentan distintos niveles de sequía y de falta de vegetación. Aunque los resultados no fueron congruentes como la asignación de grupos para distintas fechas evaluadas. Lo recomendable seria poder cargar grandes volúmenes de datos y muestrearlos o identificar cada fecha como una misma muestra independiente de las demás para obtener así mejores resultados sobre la aplicación del algoritmo K Means para la identificación de niveles de sequía por medio de imágenes multiespectrales.

REFERENCIAS

(Salas-Martínez, 2021) Salas-Martínez, F., Valdés-Rodríguez, O. A., Palacios-Wassenaar, O. M., & Márquez-Grajales, A. (2021, October). Analysis of the evolution of drought through spi and its relationship with the agricultural sector in the central zone of the state of veracruz, mexico. Agronomy, 11(11), 2099. URL: https://doi.org/10.3390/agronomy11112099, doi:10.3390/agronomy11112099

(Salas-Martínez, 2023) Salas-Martínez, F., Valdés-Rodríguez, O. A., Palacios-Wassenaar, O. M., Márquez-Grajales, A., & Rodríguez-Hernández, L. D. (2023, May). Methodological estimation to quantify drought intensity based on the nddi index with landsat 8 multispectral images in the central zone of the gulf of mexico. Frontiers in Earth Science, 11. URL: https://doi.org/10.3389/feart.2023.1027483, doi:10.3389/feart.2023.1027483

(Thiem, 2024) Thiem, H. (2024, September). Multi-year drought and heat waves across Mexico in 2024.

(Martínez, 2024) Martínez, R. (2024, May). Consecuencias de la sequía en México - UNAM Global.

(Han, 2012) Han, J., Kamber, M., & Pei, J. (2012, February). Data mining: concepts and techniques. Choice Reviews Online, 49(06), 49–3305. URL: https://doi.org/10.5860/choice.49-3305, doi:10.5860/choice.49-3305

(Rubiños, 2024) Rubiños, M., Díaz-Longueira, A., Timiraos, M., Michelena, Á., García-Ordás, M. T., & Alaiz-Moretón, H. (2024, January). A Comparative Analysis of Algorithms and Metrics to Perform Clustering.

(Wegmann, 2021) Wegmann, M., Zipperling, D., Hillenbrand, J., & Fleischer, J. (2021, January). A review of systematic selection of clustering algorithms and their evaluation.

arXiv (Cornell University). URL: https://arxiv.org/abs/2106.12792, doi:10.48550/arxiv.2106.12792

(CONAGUA, s.f.) CONAGUA. (n.d.). Monitor de Sequía en México.

(Edenhofer, 2013) Edenhofer, O. & Seyboth, K. (2013, 1). Intergovernmental Panel on Climate Change (IPCC).

(Jovic, 2014) Jovic, A., Brkic, K., & Bogunovic, N. (2014). An overview of free software tools for general data mining. 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1112-1117). doi:10.1109/MIPRO.2014.6859735

(Zwass, 2025) Zwass & Vladimir. (2025, 8). Information system | Definition, Examples, & Facts.

(Fayyad, 1996) Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 83–90). AAAI Press / MIT Press.

(Piatetsky-Shapiro, 2012) Piatetsky-Shapiro, G. & Fayyad, U. M. (2012). An introduction to sigkdd and a reflection on the term 'data mining'. SIGKDD Explorations, 13(2), 103–104.

(Comisión Nacional del Agua (CONAGUA), n.d.) Comisión Nacional del Agua (CONAGUA). (n.d.). Evaluación de la vulnerabilidad a la sequía. CONAGUA.

(Comisión Nacional del Agua (CONAGUA), 2025) Comisión Nacional del Agua (CONAGUA). (2025, jul). Programa Nacional Contra la Sequía Monitoreo de la Sequía.

(Agenda Hidalguense, 2025) Agenda Hidalguense. (2025, apr). Condiciones de sequía en 19 municipios de Hidalgo.

(González, 2024) González, A. (2024, feb). Los 84 municipios de hidalgo presentan sequía: conagua. El Sol de Hidalgo.

(Gobierno del Estado de Hidalgo, 2013) Gobierno del Estado de Hidalgo. (2013). Programa Estatal de Acción ante el Cambio Climático de Hidalgo. Editorial de la Universidad Autónoma del Estado de Hidalgo - Secretaría de Medio Ambiente y Recursos Naturales del Estado de Hidalgo.

(El Sol de Hidalgo, 2025) El Sol de Hidalgo. (2025, apr). Hidalgo sobrevive a sequías intensas durante una década. El Sol de Hidalgo.

(U.S. Geological Survey, n.d.) U.S. Geological Survey. (n.d.). Multispectral sensors collect information across the entire electromagnetic spectrum | U.S. Geological Survey.

(U.S. Geological Survey, n.d.) U.S. Geological Survey. (n.d.). The Land Analysis System (LAS) for multispectral image processing | U.S. Geological Survey.

(Esri Support, n.d.) Esri Support. (n.d.). Multispectral Image Definition | GIS Dictionary.

(U.S. Geological Survey, n.d.) U.S. Geological Survey. (n.d.). What is the Landsat satellite program and why is it important?

(NASA Scientific Visualization Studio, n.d.) NASA Scientific Visualization Studio. (n.d.). The Landsat Program.

(U.S. Geological Survey, n.d.) U.S. Geological Survey. (n.d.). Landsat 8.

(Xulu, 2019) Xulu, S., Peerbhay, K., Gebreslasie, M., & Ismail, R. (2019). Unsupervised clustering of forest response to drought stress in zululand region, south africa. Forests, 10(7). URL: https://www.mdpi.com/1999-4907/10/7/531, doi:10.3390/f10070531

(Wang, 2024) Wang, R., Zhang, X., Guo, E., Cong, L., & Wang, Y. (2024). Characteristics of the spatial and temporal distribution of drought in northeast china, 1961–2020. Water, 16(2). URL: https://www.mdpi.com/2073-4441/16/2/234, doi:10.3390/w16020234

(Shim, 2021) Shim, I., Kim, H., Hong, B., An, J., & Hwang, T. (2021). Drought vulnerability assessment and cluster analysis of island areas taking korean island areas at eup (town) and myeon (subcounty) levels as study targets. Water, 13(24). URL: https://www.mdpi.com/2073-4441/13/24/3657, doi:10.3390/w13243657

(Salvador, 2018) Salvador, M. Á. C., Pérez, M. L. M., Rodríguez, F. V., Luna, M. Y., Gasca, A. M., Manzano, A., Serrano, S. M. V., Beguería, S., & González-Hidalgo, J. C. (2018). Análisis espacio-temporal de la sequía en España peninsular. Influencia de los principales patrones de teleconexión.

(Olivares, 2018) Olivares, B. & Zingaretti, M. (2018, 02). Análisis de la sequía meteorológica en cuatro localidades agrícolas de venezuela mediante la combinación de métodos multivariados. UNED Research Journal, 10,. doi:10.22458/urj.v10i1.2026

(Hernández-Vásquez, 2022) Hernández-Vásquez, C., Ibáñez-Castillo, L., Ramón, A.-R., Monterroso Rivas, A. I., & Cervantes, R. (2022, 05). Pronóstico de sequías meteorológicas usando redes neuronales artificiales en la cuenca del río sonora, méxico. Tecnología y ciencias del agua, 13, 242-292. doi:10.24850/j-tyca-2022-03-06

(Osman, 2022) Osman, M., Zaitchik, B. F., Badr, H. S., Otkin, J., Zhong, Y., Lorenz, D., Anderson, M., Keenan, T. F., Miller, D. L., Hain, C., & Holmes, T. (2022). Diagnostic classification of flash drought events reveals distinct classes of forcings and impacts. Journal of Hydrometeorology, 23(2), 275 - 289. URL: https://journals.ametsoc.org/view/journals/hydr/23/2/JHM-D-21-0134.1.xml, doi:10.1175/JHM-D-21-0134.1

(Agustín-Canales, 2023) Agustín-Canales, N. S., Cruz-Sánchez, Y., Borja-de la Rosa, Ma. A., González-Tepale, Ma. R., & Monterroso-Rivas, A. I. (2023). Drought and vulnerability in mexico's forest ecosystems. Forests, 14(9). URL: https://www.mdpi.com/1999-4907/14/9/1813, doi:10.3390/f14091813

(Mariscal, 2010) Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137–166. Survey académico que revisa modelos incluyendo CRISP-DM y SEMMA. Consultado de [25, 33, 42, 43]. doi:10.1017/S0269888910000032

(Saltz, 2021) Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. Analiza las limitaciones de CRISP-DM para la ciencia de datos moderna y la coordinación de equipos.

(Inc., 2024) Inc., S. I. (2024). Introduction to SEMMA. Fuente primaria que define SEMMA como una organización lógica del conjunto de herramientas.

(Azevedo, 2008) Azevedo, A. & Santos, M. (2008). KDD, semma and crisp-dm: A parallel overview. Fuente comparativa académica

(Pedregosa, 2011) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. Fuente de [9, 10, 41, 42].

(Hall, 2009) Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. SIGKDD Explorations, 11(1), 10–18. Fuente de [29, 32, 43, 44].

(Guzmán, 2024) Guzmán Vera, O. A. (2024). Aplicación Web para la Ejecución de Algoritmos Descriptivos de Minería de Datos para Usuarios No Expertos. Universidad Autónoma del Estado de Hidalgo. Fuente. Tesis de Licenciatura. Asesora: Dra. Anilú Franco Arcega.

(Charrad, 2014) Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. Journal of Statistical Software, 61(6), 1–36. Fuente de [25, 26, 45, 46, 47]. doi:10.18637/jss.v061.i06