UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

EVALUACIÓN DE SOLUCIONES AGRIVOLTAICAS MEDIANTE EL USO DE UNA HERRAMIENTA BASADA EN UN SISTEMA DE INFORMACIÓN GEOGRÁFICO (SIG)

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

GIOVANNI ALESSANDRO BENEDETTO RODRÍGUEZ

PROFESOR GUÍA:
RODRIGO PALMA BEHNKE

PROFESORA CO-GUÍA:
MARCIA MONTEDONICO GODOY

MIEMBROS DE LA COMISIÓN:
MARCELO IBARRA LEIVA
ÁLVARO SILVA MADRID

Este trabajo ha sido parcialmente financiado por FONDECYT /ANID N°1241556

SANTIAGO DE CHILE
2024

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: GIOVANNI ALESSANDRO BENEDETTO RODRÍGUEZ
FECHA: 2024
PROF. GUÍA: RODRIGO PALMA BEHNKE

EVALUACIÓN DE SOLUCIONES AGRIVOLTAICAS MEDIANTE EL USO DE UNA HERRAMIENTA BASADA EN UN SISTEMA DE INFORMACIÓN GEOGRÁFICO (SIG)

Los efectos de la crisis climática han hecho necesario modernizar la matriz eléctrica, promoviendo una mayor participación de fuentes renovables, con la energía fotovoltaica desempeñando un papel crucial. Sin embargo, esto ha incrementado el uso del suelo para la generación eléctrica, lo que podría generar conflictos con el sector agropecuario. En este contexto, surge el concepto de sistemas agrivoltaicos, que combinan la generación de energía y la producción agrícola en un mismo terreno.

El presente trabajo tiene como objetivo caracterizar diversas soluciones agrivoltaicas a nivel global y, a partir de los datos recopilados, construir un modelo que permita identificar y caracterizar zonas donde aún no se han desarrollado proyectos para la implementación de soluciones agrivoltaicas.

Para alcanzar el objetivo, se plantea la siguiente propuesta metodológica. Primero, se buscan proyectos agrivoltaicos georreferenciados utilizando diversas fuentes de información y, a partir de esa información, se desarrolla una base de datos. Luego, se realiza un análisis exploratorio de datos con el fin de preprocesarlos para posteriormente construir un modelo que se ajuste a los datos.

Los principales resultados de este proyecto son la elaboración de un sistema de información georreferenciado, un estudio estadístico de las características más relevantes de los proyectos seleccionados, y un modelo basado en el algoritmo de agrupación de datos mixto K-prototypes, que identifica tres grupos de proyectos agrivoltaicos.

El primer grupo está conformado por soluciones agrivoltaicas de pequeña escala, con una capacidad promedio de 0,45 [MW] y una superficie promedio de 1 [ha], empleando paneles monofaciales de arreglo fijo a más de 3 metros de altura. En estos proyectos se cultivan principalmente hortalizas y frutas, y se encuentran ubicados principalmente en Europa, India, Chile y Estados Unidos.

El segundo grupo incluye soluciones agrivoltaicas de gran tamaño, con una capacidad promedio de 55 [MW] y una superficie promedio de 172 [ha]. Estos proyectos utilizan paneles monofaciales con seguimiento en un eje, donde el cultivo principal son praderas, están ubicados en Estados Unidos, específicamente en el estado de California.

El último grupo abarca proyectos de tamaño intermedio, con una capacidad promedio de 10 [MW] y una superficie promedio de 28 [ha]. Estos proyectos emplean paneles monofaciales que pueden ser de arreglo fijo o contar con seguimiento en un eje a una altura de 0,7 metros, y están destinados principalmente al cultivo de praderas. Se ubican principalmente en el centro de Estados Unidos.

"When you walk through a storm
Hold your head up high
And don't be afraid of the dark
At the end of a storm
There's a golden sky
And the sweet silver song of a lark
Walk on through the wind
Walk on through the rain
For your dreams be tossed and blown
Walk on, walk on
With hope in your heart
And you'll never walk alone
You'll never walk alone
".

"You'll Never Walk Alone" - Rodgers y Hammerstein

Agradecimientos

En primer lugar, doy gracias a Dios por la vida que me ha dado y permitirme disfrutar de la compañía de mis seres queridos.

Quiero agradecer a mis padres por todas sus enseñanzas, amor y apoyo incondicional a lo largo de estos años. Su paciencia y comprensión han sido pilares fundamentales en mi desarrollo personal. Gracias también a mi hermano Giuseppe, quien siempre ha cuidado de mí y me ha apoyado en todos mis sueños.

También quiero expresar mi agradecimiento a mis tíos y tías, Magdalena, Luis, Lucrecia y Alejandro, quienes siempre me han hecho sentir respaldado y querido. Agradezco a mis primos y primas por su apoyo, en especial a Valeria, quien ha sido como una hermana mayor para mí, y a Vicente, quien siempre me ha acompañado en cada una de mis locuras. Agradezco también a mi mami Lastenia por todo el cariño y amor que me ha entregado. Además, no puedo olvidar a mis pequeños sobrinos, cuyas travesuras me han sacado tantas canas como carcajadas, aportando alegría a mi día a día.

Quiero agradecer a mis amigos del liceo, en especial a José, Alonso, Gonzalo, Cristóbal y Nicolás, por las risas interminables y las aventuras compartidas, pero más importante aún, les agradezco de corazón por haberme acompañado durante todos estos años. También quiero extender mi agradecimiento a mis amigos de la universidad, en especial a Felipe y Gabriel, por la fraternidad y apoyo mutuo, así como las trasnochadas que compartimos para salvar los ramos.

De igual manera, quiero agradecer a mis profesores guía, Rodrigo Palma y Marcia Montedonico, por su paciencia y enseñanzas. Gracias por guiarme y darme la oportunidad de realizar este trabajo.

Agradezco a todas las personas que me han acompañado en este camino. Aunque no los haya nombrado, siempre estaré agradecido de haberlos conocido.

Finalmente, se agradece el financiamiento parcial del proyecto FONDECYT /ANID N°1241556.

Tabla de contenido

Índice de tablas

Índice de ilustraciones

Capítulo 1

Introducción

1.1. Motivación

El planeta Tierra enfrenta una situación extremadamente compleja debido al cambio climático y sus consecuencias, las cuales amenazan la estabilidad del ecosistema y la calidad de vida de la humanidad. Los efectos del calentamiento global, originados principalmente por la acumulación de gases de efecto invernadero en la atmósfera, se manifiestan en fenómenos climáticos extremos, proliferación de incendios forestales, olas de calor, lluvias fuera de temporada, sequías, aumento del nivel del mar y alteraciones en los ecosistemas. A medida que el planeta se calienta, las consecuencias sociales, económicas y ambientales se vuelven cada vez más graves y difíciles de manejar.

El incremento de la población mundial, junto con las crecientes demandas de energía eléctrica y térmica, ha exacerbado las emisiones de dióxido de carbono (CO_2) hacia la atmósfera, como se ilustra en el gráfico de la Figura 1.1. En el caso de Chile, los principales causantes de las emisiones de gases de efecto invernadero (GEI) son los sectores de energía, procesos industriales y uso de productos (IPPU), agricultura y residuos, siendo la mayor parte de estas emisiones de GEI originada en el sector energético [4].

El sector energético representó el 75 % de las emisiones totales de GEI en el año 2020, alcanzando un total de 79724 de kilotoneladas de CO_2 equivalente (kt CO_2 eq), lo que supone un aumento del 139 % desde 1990 y una disminución del 5 % desde 2018 [5], según se muestra en la Figura 1.2. Cabe destacar que el sector energético incluye todos los consumos cuyo uso principal implica el empleo de combustibles fósiles en el país y sus emisiones fugitivas asociadas. Por ende, no se limita únicamente a la generación de electricidad. No obstante, la producción de electricidad tiene la mayor participación en las emisiones de GEI dentro de este sector. En el año 2020, sus emisiones totalizaron 29842 [kt CO_2 eq] [6].

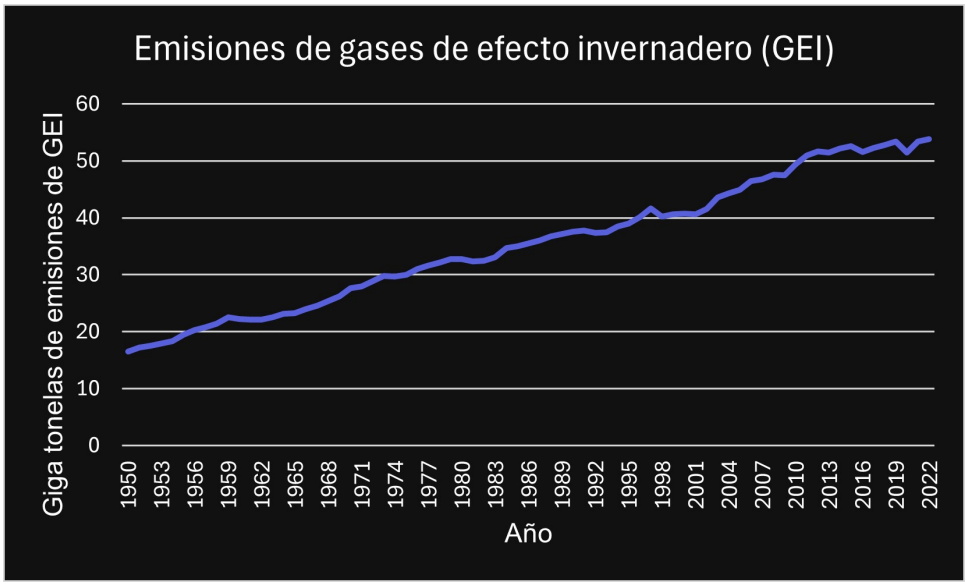

Figura 1.1: Emisiones de gases de efecto invernadero a nivel mundial entre los años 1950 y 2022. Fuente: Jones et al. (2024) [1].

Esto significa que el sector energético presenta las mayores oportunidades para cumplir con los compromisos de Chile de alcanzar la neutralidad de carbono para el año 2050, compromisos que fueron suscritos en virtud del Acuerdo de París. Este tratado internacional sobre el cambio climático es jurídicamente vinculante y tiene como objetivo limitar el calentamiento global a niveles muy inferiores a 2 $[°C]$, preferiblemente a $1,5$ $[°C]$, en comparación con los niveles preindustriales. [7]

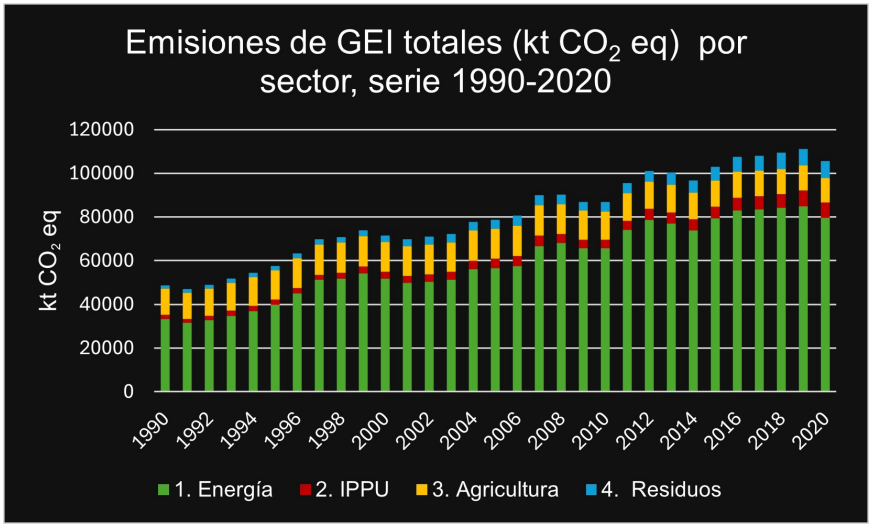

Figura 1.2: Emisiones totales de GEI por sector. Fuente: Equipo Técnico Coordinador del MMA.

En 2018, el Grupo Intergubernamental de Expertos sobre el Cambio Climático (IPCC, por sus siglas en inglés) publicó un informe especial sobre los impactos del calentamiento global a $1,5$ $[°C]$ [8]. Este informe resalta una serie de daños causados por el cambio climático que podrían evitarse si se establece un límite de calentamiento global de $1,5$ $[°C]$ en lugar de 2

[$^{\circ}C$] o más. Por ejemplo, para el año 2100, el aumento del nivel del mar a nivel global sería 10 [cm] menor con un calentamiento global de $1,5$ [$^{\circ}C$]. El informe también señala que limitar el calentamiento global a $1,5$ [$^{\circ}C$] requeriría cambios "rápidos y significativos" en la tierra, la energía, la industria, los edificios, el transporte y las ciudades.[9]

Para abordar esta problemática, es esencial reconsiderar la matriz energética y adoptar fuentes de energía renovable, como la generación solar. Sin embargo, la generación solar requiere grandes extensiones de terreno, y la creciente demanda de alimentos debido al aumento de la población compite por el uso de estas tierras. Esto da lugar a un conflicto en el uso del suelo entre la generación solar y la producción de alimentos.

A medida que la humanidad enfrenta el desafío del cambio climático, es crucial encontrar soluciones sostenibles que equilibren la necesidad de energía con la de alimentos sin comprometer aún más la salud de nuestro planeta. De este modo, nace el concepto de sistema agrivoltaico [10], proponiendo tanto el uso del terreno tanto para la producción de energía mediante tecnología fotovoltaica, en combinación armónica y optimizada con la producción agrícola y ganadera. Un ejemplo de estos sistemas se presenta en la figura 1.3. Esta solución no solo resuelve el conflicto de uso del suelo, sino que también tiene un impacto en el microclima que se forma debajo de los paneles, modificando las condiciones climáticas en el terreno. Otra ventaja es la reducción de la evaporación resultante de los rayos del sol, lo cual disminuye la necesidad de riego. Cuando los paneles solares tienen una altura adecuada y una distancia apropiada entre ellos, este sistema puede proteger los cultivos agrícolas y reducir la necesidad de agua [11].

Figura 1.3: Ejemplo de un sistema agrivoltaico: Jack's Solar Garden (EE. UU.). Fuente: Jack's Solar Garden [2].

Para llevar a cabo la implementación eficiente de sistemas agrivoltaicos, resulta esencial realizar una detallada caracterización del territorio, donde la evaluación del clima, la identificación de los parámetros estructurales del arreglo de paneles y el uso del suelo desempeñan un papel crucial. La comprensión de las variables climáticas, junto con los parámetros estructurales, puede asegurar una producción de energía sostenible. Además, evaluar la calidad del suelo y su uso actual posibilita la integración armónica de la agricultura con la generación de energía solar.

Por otra parte, la utilización de herramientas analíticas robustas se presenta como un recurso fundamental para detectar patrones recurrentes en áreas geográficas específicas. En este contexto, la disponibilidad de un sistema de información geográfica (SIG) para el análisis detallado de datos geoespaciales y la caracterización previamente mencionada puede brindar una perspectiva precisa sobre qué solución agrivoltaica se está implementando en estas zonas. Para realizar esta evaluación de manera eficaz, es esencial emplear una variedad de técnicas de agrupación de datos. Estas herramientas posibilitan la identificación de agrupaciones naturales o patrones subyacentes en los datos, lo que simplifica la segmentación de áreas con atributos similares y permite identificar las ubicaciones más adecuadas para la implementación de sistemas agrivoltaicos.

1.2. Objetivos

Objetivo General

A continuación se presenta el objetivo general que este trabajo busca lograr.

Desarrollar una herramienta destinada a caracterizar territorios con implementación de sistemas agrivoltaicos. A partir del análisis y estudio de los datos recopilados, se pretende construir un modelo que permita caracterizar las soluciones agrivoltaicas para zonas específicas.

Objetivo Específico

A su vez, los objetivos específicos del trabajo se listan a continuación.

1. Elaborar un mapa georreferenciado de proyectos agrivoltaicos que incluya los datos más relevantes que caracterizan a estos sistemas.
2. Identificar y estudiar las características más relevantes de los sistemas agrivoltaicos seleccionados, en relación a territorios específicos donde ya se encuentran proyectos desarrollados.
3. Construcción de un modelo a partir de los datos recopilados que permita caracterizar zonas específicas donde no se hayan desarrollado proyectos para la implementación de soluciones agrivoltaicas, en relación a territorios específicos donde ya se encuentran proyectos desarrollados.
4. Validación de modelo a partir de casos de estudio.

1.3. Estructura del documento

La estructura del presente trabajo se organiza en los siguientes siete capítulos.

En el primer capítulo, Introducción, se establece el contexto general de la investigación, resaltando la importancia de las soluciones agrivoltaicas en el marco de la sostenibilidad energética y agrícola. Se presentan también el objetivo general y los objetivos específicos del trabajo.

El segundo capítulo, Antecedentes y estado del arte, ofrece una revisión de la literatura existente sobre sistemas agrivoltaicos, bases de datos y algoritmos de agrupamiento de datos. Se presta especial atención bases de datos SIG ya existentes para sistemas agrivoltaicos.

El tercer capítulo, Metodología propuesta, describe en detalle los métodos y procedimientos utilizados en el estudio. Se explican los criterios de selección de datos, las herramientas tecnológicas empleadas y los enfoques de análisis de datos. Este capítulo también justifica la elección de la metodología y su adecuación para alcanzar los objetivos planteados.

El cuarto capítulo, Desarrollo de base agrivoltaica, documenta el proceso de creación de la base de datos que constituye el núcleo de la herramienta SIG utilizada en la investigación. Se describen las fuentes de datos, la selección de características más relevantes y la elaboración del modelo relacional de la base de datos.

El quinto capítulo, Análisis exploratorio de datos, presenta un examen inicial de los datos recolectados, utilizando técnicas de análisis exploratorio para identificar patrones, tendencias y posibles correlaciones dentro de la base de datos agrivoltaica.

En el sexto capítulo, Análisis de identificación de agrupamiento, se aborda el proceso de identificación de agrupamientos en los datos, empleando algoritmos de aprendizaje automático no supervisado. Se presentan los resultados obtenidos, junto con su interpretación en el contexto de la evaluación de soluciones agrivoltaicas.

El último capítulo, Conclusiones, sintetiza los hallazgos principales de la investigación, evalúa el cumplimiento de los objetivos propuestos y sugiere posibles líneas de investigación futura.

1.4. Alcances del trabajo

Los alcances de este trabajo son los siguientes:

- Las fuentes de información empleadas son en su mayoría secundarias, excepto en el caso de los proyectos chilenos, para los cuales se consultan fuentes primarias.
- Los proyectos agrivoltaicos que no puedan ser georreferenciados con la información proporcionada por las fuentes no son considerados en la construcción de la base de datos ni en la elaboración del SIG agrivoltaico.
- Los proyectos agrivoltaicos que estén fuera de servicio son considerados para la construcción de la base de datos y SIG agrivoltaico.

Capítulo 2

Antecedentes y estado del arte

2.1. Sistemas agrivoltaicos

Los sistemas agrivoltaicos (ver figura 2.1), también llamados AgroPV o AgriPV, se refieren a la combinación de una misma área de tierra para la agricultura y la generación de electricidad a través de paneles fotovoltaicos. El objetivo principal de estos sistemas es solucionar el conflicto entre el sector eléctrico y el sector agropecuario, ya que en muchos casos los lugares más adecuados para las plantas de energía solar, coinciden con tierras agrícolas clasificadas en las clases de capacidad de uso de suelos I (suelos de alta calidad), II (suelos con ligeras limitaciones) y III (suelos con limitaciones moderadas) [12]. Esto presenta un grave problema, porque es muy probable que las tierras que respaldan una agricultura viable y diversa tengan más valor como tierra para infraestructura energética. Esta competición por el uso del terreno podría ser particularmente seria en regiones densamente pobladas, áreas montañosas y pequeñas islas habitadas [13].

Figura 2.1: Sistema agrivoltaico de 1 [MWp] en el estado de Guyarat, India. Fuente: NSEFI.

2.1.1. Beneficios

La integración de la tecnología fotovoltaica, comúnmente abreviada como PV (siglas del inglés "Photovoltaic"), con la agricultura presenta un beneficio clave: la capacidad de disminuir la demanda eléctrica de los agricultores sin comprometer el rendimiento de los cultivos.

Esto no solo implica una ventaja financiera sustancial al brindar a los agricultores dos fuentes de ingresos, sino que también representa una oportunidad de inversión significativa en los sectores de PV y agricultura. En primer lugar, los agricultores pueden generar su propia electricidad mediante sistemas de PV, cubriendo sus necesidades energéticas internas y, además, comercializando el excedente de energía. En segundo lugar, continúan generando ingresos a través de la producción de cultivos. Esta sinergia no solo mejora la situación financiera de los agricultores, sino que también abre puertas a oportunidades de inversión, aprovechando los parques fotovoltaicos ya existentes [11].

Una ventaja notable de los sistemas agrivoltaicos, como ya fue mencionado, es su potencial para abordar el conflicto entre la tierra agrícola y la producción de energía. Estos sistemas transforman de manera efectiva la competencia por la tierra en una mezcla armoniosa de agricultura y producción de energía solar. Esta relación simbiótica es altamente ventajosa, ya que combina la generación de energía con una agricultura sostenible.

Esta doble utilización de la tierra no solo mantiene algunos de los servicios ecosistémicos que provee el suelo y optimiza el aprovechamiento económico del suelo, sino que también puede generar efectos beneficiosos entre la producción agrícola y el sistema agrivoltaico. Dependiendo del diseño del sistema, la construcción puede asumir importantes funciones de protección (por ejemplo, protección contra el granizo u otros eventos climáticos extremos), así como contribuir a la recogida de agua de lluvia con los dispositivos adecuados. Especialmente en los años más cálidos y en las regiones secas, cabe esperar una menor evaporación del agua del suelo en la instalación debido a la sombra [14].

La instalación de paneles en un campo agrícola puede generar sombra sobre los cultivos, lo que afectará directamente las necesidades de luz de las plantas. La luz es esencial para la fotosíntesis y un crecimiento óptimo, y se distinguen tres aspectos: calidad, cantidad y duración, todos los cuales influyen significativamente en los cultivos. Cada tipo de cultivo tiene un punto de saturación lumínica, donde la radiación fotosintéticamente activa por encima de ese punto no mejora la fotosíntesis ni el crecimiento. Identificar este punto de saturación lumínica para un cultivo específico es crucial para comprender su tolerancia a la sombra [11].

Algunos estudios sugieren que las bayas y frutas pueden obtener beneficios iniciales de hasta un 40 % de sombra. En contraste, los forrajes, las verduras de hojas y los tubérculos no muestran una variación evidente en su rendimiento para el mismo porcentaje de sombra. Por otro lado, el maíz y los cultivos de leguminosas son altamente susceptibles a la sombra, lo que lleva a la clasificación de los cultivos en tres grupos: intolerantes a la sombra, tolerantes a la sombra y amantes de la sombra [15]. Sin embargo, otros estudios sugieren que la pertenencia de un cultivo a alguno de estos grupos puede variar y que no es aplicable de manera universal a todas las ubicaciones alrededor del mundo. Por lo tanto, se requieren experimentos de campo para verificar la pertenencia de los cultivos a los grupos [16].

2.2. Sistemas de información geográfico (SIG)

Un Sistema de Información Geográfica, abreviado como SIG, engloba un conjunto de herramientas, tecnologías y procesos destinados a la recopilación, almacenamiento, análisis,

interpretación y visualización de datos geoespaciales. Según el libro *Sistemas de información geográfica* [17], es fundamental que un SIG posibilite la ejecución de las siguientes operaciones:

- Lectura, edición, almacenamiento y, en términos generales, gestión de datos espaciales.
- Análisis de dichos datos. Esto puede incluir desde consultas sencillas a la elaboración de complejos modelos, y puede llevarse a cabo tanto sobre la componente espacial de los datos (la localización de cada valor o elemento) como sobre la componente temática (el valor o el elemento en sí).
- Generación de resultados tales como mapas, informes, gráficos, etc.

De manera similar, un Sistema de Información Geográfica (SIG) puede concebirse como un "mapa de orden superior", entendiendo que representa una forma más poderosa y avanzada de llevar a cabo todas las tareas que, antes de la aparición de los SIG, se realizaban mediante el uso de mapas y cartografía en el sentido clásico. En otras palabras, los SIG constituyen un avance más allá de la función tradicional de los mapas. Sin embargo, esta definición resulta demasiado simplista, ya que mapas y SIG no son conceptos equivalentes en el contexto actual de estos últimos [17].

Un mapa es simplemente una representación de un conjunto de datos espaciales, y aunque esta representación es de gran importancia, en el ámbito de un SIG, es solo un componente más dentro de una serie de elementos. Además, un SIG no solo incluye datos y representación, sino también las operaciones que pueden hacerse sobre el mapa, parte integral del sistema conformado por el SIG.

Para una definición más precisa, se puede afirmar que un SIG es un sistema que integra tecnología informática, personas e información geográfica, y cuya función principal es capturar, analizar, almacenar, editar y representar datos georreferenciados [18].

Los SIG, según el libro *Sistemas de información geográfica* [17], se componen de elementos básicos, los cuales son:

- Datos: son indispensables para otorgar sentido al resto de los componentes de un SIG y permitirles desempeñar su función en el sistema. La información geográfica, que constituye la esencia misma de los SIG, tiene su base en los datos. Por lo tanto, comprender a fondo los datos y su naturaleza es crucial para una adecuada comprensión de los SIG.
- Análisis: las funcionalidades fundamentales de los SIG siempre incluyen, en mayor o menor medida, una serie de procesos que posibilitan la obtención de resultados y el análisis de los datos espaciales. Estas formulaciones representan procesos que pueden ser desde muy simples hasta enormemente complejos, aplicables en diversos campos o de manera general. Su origen puede ser muy diverso y no necesariamente proviene del ámbito puro de la geografía, ya que van desde simples consultas o mediciones hasta la creación de modelos elaborados que utilizan datos de variables numerosas, generando resultados complejos.
- Visualización: cualquier tipo de información puede representarse de manera gráfica, lo cual suele facilitar la interpretación de dicha información o de una parte de ella. Muchas características de la información, como la presencia de patrones sistemáticos, son más fáciles de estudiar cuando se respaldan con algún elemento visual, ya que este aporta

un nuevo enfoque.

- Tecnología: incluye tanto el hardware en el que se ejecutan las aplicaciones SIG como las propias aplicaciones, es decir, el software SIG.
- Factor organizativo: el sistema SIG requiere una organización y una coordinación adecuada entre sus distintos elementos. El factor organizativo ha ganado progresivamente importancia en el entorno de los SIG a medida que han evolucionado, generando sistemas más complejos y un mayor número de interrelaciones entre los diversos componentes que lo conforman.

En resumen, un SIG es una herramienta esencial para la gestión de información geográfica, indispensable para trabajar con cualquier tipo de datos georreferenciados en la actualidad.

2.3. Bases de datos

Una base de datos es un sistema compuesto por un conjunto de datos almacenados que abordan diferentes temáticas y se organizan de diversas formas, pero que tienen algún tipo de conexión o contexto común que facilita su acceso directo y su interrelación. Estos datos están guardados sobre un soporte físico [19].

Según lo expuesto en el libro *Bases de datos* [20], "una base de datos es un conjunto de datos almacenados entre los que existen relaciones lógicas y ha sido diseñada para satisfacer los requerimientos de información de una empresa u organización".

Una definición más elaborada es proporcionada en el libro *Diseño de base de datos* [21], donde el término bases de datos se entiende como "la representación a nivel integrado de una colección estructurada de datos que contiene físicamente el diseño lógico de un conjunto de entidades, instancias de las diferentes entidades del sistema de información que se está modelando en una organización y las interrelaciones de las entidades; representación que necesita de una gestión de datos a fin de ser utilizada de manera compartida por todos los usuarios de una organización en la resolución de sus necesidades de información".

El desarrollo de una base de datos es un proceso complicado que surge de la necesidad de almacenar información del mundo real para facilitar un acceso rápido y eficiente. Un diseño inadecuado de la base de datos puede dar lugar a problemas como redundancias, inconsistencia de los datos e incoherencias.

Con estas definiciones en mente, en esta sección se abordarán en detalle las definiciones y conceptos básicos de las bases de datos, el modelado de datos, el diseño y creación de bases de datos, la gestión y mantenimiento de bases de datos. Esta comprensión integral proporcionará el fundamento necesario para explorar cómo las bases de datos pueden satisfacer de manera efectiva las necesidades de información de las organizaciones.

2.3.1. Modelo entidad-relación

Manejar datos es sencillo cuando son pocos, pero a medida que su volumen crece, se hace necesario utilizar distintos modelos para facilitar su diseño y gestión.

Un modelo proporciona mecanismos de abstracción para representar una parte del mundo cuyos datos son de interés. Esta representación, realizada en términos de un modelo específico, se denomina esquema, y el conjunto de datos que representa constituye la base de datos.

El modelo entidad-relación es una representación conceptual de los datos, propuesta por Peter Chen en 1976. Presenta un elevado nivel de abstracción y se basa en comprender la realidad como un conjunto de entidades y las relaciones entre ellas [20].

Una entidad es un objeto que existe y puede ser distinguido de otros, como por ejemplo una persona, que se diferencia de cualquier otra persona.

Las entidades se pueden relacionar entre ellas, estas asociaciones son conocidas como relaciones. Los conjuntos de relaciones pueden tener restricciones. La más frecuente es la cardinalidad de asignación, que limita el número de entidades relacionadas con una entidad de otro conjunto de entidades . Las posibles restricciones de cardinalidad según el libro [19], son:

- Uno a Uno (1:1): un registro de una entidad A se relaciona con solo un registro en una entidad B.
- Uno a Varios (1:M): un registro en una entidad A se relaciona con varios registros en una entidad B. Pero los registros de B se relacionan con uno solo en la entidad A.
- Varios a Uno (M:1): un registro en una entidad A se relacionan con solo un registro en una entidad B. Pero la entidad en B se puede relacionar con varios registro de la entidad A.
- Varios a Varios (M:M): una entidad en A se puede relacionar con una o con muchas entidades en B y viceversa.

Una entidad puede contar con varias características de las cuales algunas sean de interés conservar, estas son conocidas como atributos. Según el libro [19], se tiene tres tipos de atributos:

- Atributos de una entidad.
- Atributos que sirven para identificar entidades, conocidos como claves. Estos son los que poseen la propiedad identificatoria.
- Atributos de una relación, conocidos como atributos descriptivos.

También hay restricciones de integridad. Se trata de condiciones de obligado cumplimiento para los datos de las bases de datos. Hay de dos tipos: inherentes y explícitas.

Las restricciones inherentes son aquellas que no son determinadas por los usuarios, sino que son definidas por el hecho de que la base de datos sea relacional. Según el libro [19] estas son:

- No puede haber dos entidades iguales.
- El orden de las entidades no importa.
- El orden de los atributos dentro de una entidad no importa.

El modelo relacional permite a los usuarios incorporar restricciones personales o explícitas

a los datos. Las más usadas son las siguientes:

- Clave primaria: hace que los atributos marcados como clave primaria no puedan repetir valores.
- Unicidad: impide que los valores de los atributos marcados de esta forma puedan repetirse.
- Obligatoriedad: prohíbe que el atributo marcado de esta forma no tenga ningún valor.
- Integridad referencial: prohíbe colocar valores en una clave externa que no estén reflejados en la tabla donde ese atributo sea clave primaria.
- Regla de validación: condición que debe cumplir un dato concreto para que sea actualizado.

2.3.2. Bases de datos relacionales

Es una representación lógica de los datos basada en la teoría de conjuntos y la lógica de predicados. Organiza los datos en tablas (relaciones), que constan de filas (tuplas) y columnas (atributos). Edgar Frank Codd definió las bases del modelo de datos relacional a finales de los años 60. Con el tiempo, este modelo se fue adoptando cada vez más, hasta convertirse en el modelo de bases de datos más popular.

Antes de adentrarse más en las bases de datos relacionales, es necesario convertir el diseño conceptual del modelo entidad-relación al modelo relacional. Para eso, se deben seguir los siguientes pasos descritos en el libro [19].

1. En el caso de una entidad fuerte, es decir, una entidad cuya existencia es independiente en la base de datos, el paso a tablas de entidades fuertes es:
   - Nombre de la tabla = Nombre de la entidad
   - Campos de la tabla = Atributos de la entidad
   - Llave primaria = Identificadores primarios de la entidad
2. Por otro lado, las entidades débiles no pueden ser identificadas de manera única solo con sus propios atributos. Su existencia depende de una entidad fuerte. Paso a tablas de entidades débiles:
   - Nombre de la tabla = Nombre de la entidad
   - Campos de la tabla = Atributos de la entidad + clave primaria de la entidad de la que depende
   - Llave primaria = Discriminante (identificador débil de la entidad débil) + identificador primario de la entidad de la que depende
3. Paso a tablas de relaciones:
   - Nombre de la tabla = Nombre de la relación
   - Campos de la tabla = Identificadores primarios de las entidades relacionadas + posibles atributos descriptivos
   - Llave primaria = llave compuesta, como mínimo, de los identificadores principales de las entidades relacionadas
   - En el caso de que la cardinalidad de la relación sea M:M, se debe crear una tabla independiente. De lo contrario, se debe modificar levemente al menos una de las

entidades relacionadas. Si la relación es de cardinalidad 1:M o M:1, el identificador primario de la entidad "1" se incluye como clave foránea en la nueva tabla correspondiente a la entidad "M". Para las relaciones en las que ambas entidades tienen una cardinalidad de "1", la alternativa más común consiste en colocar como llave foránea el identificador primario de una de las entidades en la nueva tabla correspondiente a la otra entidad (es indiferente qué identificador se escoja).

Las bases de datos relacionales están construidas sobre tablas, también llamadas relaciones. Las tablas se representan gráficamente como una estructura rectangular formada por filas y columnas. Cada columna, también llamada atributo, guarda una característica o propiedad determinada de la tabla, mientras que cada fila, también llamada tupla, contiene una ocurrencia o ejemplar de la instancia representada en la tabla [19].

Las tablas deben cumplir con las siguientes propiedades:

- Unicidad de nombre: cada tabla debe tener un nombre único.
- Atributos con dominios atómicos: cada intersección entre una fila y una columna debe contener un solo valor.
- Columnas con nombres distintos: cada columna debe tener un nombre distinto. También es aconsejable que estos nombres no se repitan entre tablas.
- Independencia del orden de las columnas: el orden de las columnas en una tabla no importa.
- Independencia del orden de las filas: el orden de las filas en una tabla no importa.
- Unicidad de clave primaria: dos filas de una tabla no pueden tener el mismo valor de clave primaria.
- Unicidad de atributos: no puede haber dos filas con todos sus atributos iguales.

En el modelo relacional podemos distinguir los siguientes tipos de claves:

- Llave candidata: corresponde al conjunto de atributos que pueden identificar las tuplas de forma unívoca y mínima en una tabla. Una llave candidata no contiene atributos redundantes y cada valor dentro de la llave candidata es único.
- Llave primaria: es aquella llave candidata que es escogida por el desarrollador para identificar de manera única las tuplas en una tabla. La llave primaria no puede contener valores nulos y debe ser única para cada registro de la tabla.
- Llave foránea: es un atributo o un conjunto de atributos en una tabla que se refiere a la clave primaria de otra tabla. La llave foránea establece una relación entre las dos tablas, asegurando la integridad referencial, lo que significa que los valores de la llave foránea deben corresponder a valores existentes en la clave primaria de la tabla relacionada.

Por último, es necesario abordar el valor "Null". Este constituye un valor especial que representa la ausencia de datos en una columna. Cuando se desconoce algún valor de un atributo para una fila determinada, se le asigna el valor "Null".

2.3.3. Normalización

La normalización en una base de datos es un proceso destinado a minimizar la redundancia, el mantenimiento y los problemas de actualización de los datos en dicha base [19].

Este proceso se divide en niveles denominados formas normales, que van desde la primera forma normal (1FN) hasta la quinta forma normal (5FN). En esta memoria solo se estudia hasta la tercera forma normal (3FN), ya que las primeras tres formas son generalmente suficientes para la mayoría de las bases de datos.

En primer lugar, es importante definir el concepto de dependencia funcional. Una dependencia funcional ocurre cuando dos columnas, A y B, de una tabla R, están relacionadas de tal manera que B es funcionalmente dependiente de A. Esto significa que para cada valor de A, existe un único valor de B asociado con él. Esta relación se denota como $A \rightarrow B$.

Por ejemplo, consideremos una tabla de empleados con las columnas $ID_{Empleado}$, $Nombre$, $Departamento$ y $Salario$. Podemos tener una dependencia funcional entre $ID_{Empleado}$ y $Nombre$, lo que significa que para cada $ID_{Empleado}$ específico, existe un único "Nombre" correspondiente. En este caso, la dependencia funcional se expresa como $ID_{Empleado} \rightarrow Nombre$.

Además, cuando un atributo B en una relación R depende funcionalmente de otro atributo A en la misma relación R, sin depender funcionalmente de ningún subconjunto de A, se dice que la dependencia funcional es completa. Si esto no ocurre, entonces se dice que existe una dependencia funcional transitiva. En este caso, dados los atributos o conjuntos de atributos A, B y C, C depende funcionalmente de manera transitiva de A si sucede que $A \rightarrow B$ y luego $B \rightarrow C$.

Una relación se considera estar en la primera forma normal (1FN) si ninguno de sus atributos tiene dominios que sean conjuntos. En otras palabras, para que una relación esté en 1FN, todos sus atributos deben contener un único valor, es decir, deben ser atómicos.

Una relación se encuentra en la Segunda Forma Normal (2FN) si cumple con los requisitos de la Primera Forma Normal (1FN) y, además, cada atributo que no forma parte de la clave primaria depende completamente de dicha clave primaria de manera funcional.

Una relación se dice que está en la Tercera Forma Normal (3FN) si cumple con los criterios de la Segunda Forma Normal (2FN) y, además, no presenta dependencias funcionales transitivas entre la clave primaria y sus atributos que no pertenecen a ninguna clave alternativa. En otras palabras, en 3FN no ocurre que un atributo dependa funcionalmente de otro u otros atributos que no sean parte de la clave primaria.

2.4. Algoritmos de agrupamiento de datos

El agrupamiento (en inglés "clustering") es una técnica de análisis de datos que consiste en agrupar un conjunto de objetos de tal manera que los objetos en el mismo grupo (en inglés "cluster") sean más similares entre sí que a los objetos de otros grupos.

Hay varias técnicas de agrupamiento disponibles. En la figura 2.2, se muestra el rendimiento de diferentes algoritmos proporcionados por la librería "scikit-learn" para varias distribuciones de datos. A continuación, se analizarán algunas de estas técnicas.

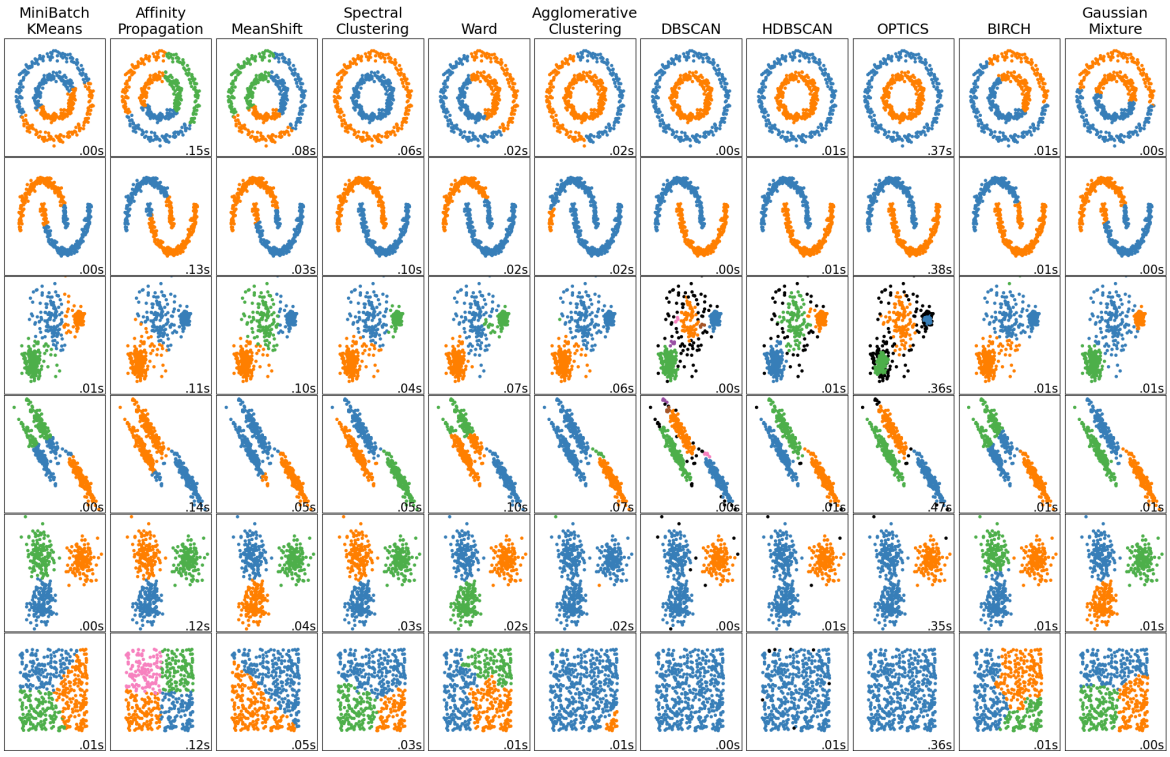

Figura 2.2: Una comparación de los algoritmos de agrupación de la librería de python "scikit-learn". Fuente: Clustering en sckit-learn [3].

2.4.1. Tipos de datos

Antes de analizar cualquier algoritmo de agrupación, es importante distinguir entre los tipos de datos existentes, ya que esto determina qué técnicas y métodos pueden ser utilizados para analizarlos. Los datos se pueden clasificar en tres tipos: numéricos, categóricos y ordinales.

Los datos numéricos, también conocidos como cuantitativos, representan cantidades y pueden ser medidos y expresados en números. Estos datos pueden ser continuos o discretos. Por ejemplo, la altura de una persona o el número de hijos en una familia.

Los datos categóricos, también conocidos como cualitativos, describen características o cualidades que no tienen una cantidad inherente. Estos datos representan categorías sin un orden específico y no se pueden ordenar de manera que tenga sentido cuantitativo. Un ejemplo

de datos categóricos es el color del cabello en un grupo de personas (rubio, castaño, negro, etc.).

Por último, los datos ordinales describen características o cualidades de manera similar a los datos categóricos, pero tienen una secuencia lógica. Aunque existe un orden, la distancia entre las categorías no es necesariamente igual. Por ejemplo, un dato ordinal puede ser los niveles de acuerdo en una encuesta, como: muy de acuerdo, de acuerdo, neutral, en desacuerdo, muy en desacuerdo. El orden es claro, pero la diferencia entre cada nivel puede no ser equidistante.

2.4.2. Métricas de distancias

Al igual que con los tipos de datos, es necesario estudiar distintas métricas de distancia que se utilizan para asignar los datos a los grupos. Esto se debe a que la distancia euclidiana, la métrica más utilizada por defecto en la mayoría de los algoritmos de agrupación, pierde su significado cuando se trata de atributos no numéricos.

Para los datos numéricos, es común utilizar la distancia euclidiana, como se menciona anteriormente. La distancia euclidiana entre los puntos $P = (p_1, p_2, ..., p_n)$ y $Q = (q_1, q_2, ..., q_n)$ esta definida por la expresión 2.1 [22].

$$\mathrm{d}(P, Q) = \sqrt{\sum_{\mathrm{i}=1}^{n} (p_{\mathrm{i}} - q_{\mathrm{i}})^2} \tag{2.1}$$

En el caso de datos categóricos, no se utiliza una medida de distancia como tal, sino medidas de similitud entre dos muestras, como el índice de Jaccard o el coeficiente de Sorensen-Dice.

Sin embargo, cuando se trata de un conjunto de datos mixtos, es decir, datos numéricos y categóricos, Gower en 1971 propuso una forma de medir cuán diferentes son dos conjuntos de datos.

Sean dos individuos, i y j con v caracteres, que pueden ser comparados en un carácter k. A estos se les puede asignar un puntaje $S_{\mathrm{i}jk}$, que es cero cuando i y j se consideran diferentes, y una fracción positiva cuando tienen algún grado de acuerdo o similitud. A veces, no es posible hacer comparaciones porque falta información o, en el caso de variables dicotómicas, un carácter no existe en ambos, i y j. La posibilidad de hacer comparaciones se puede representar por una cantidad $\delta_{\mathrm{i}jk}$ igual a 1 cuando el carácter k se puede comparar entre i y j, y 0 en caso contrario. Cuando $\delta_{\mathrm{i}jk}$, $S_{\mathrm{i}jk}$ es desconocido pero convencionalmente se establece en cero. La similitud entre i y j se define como el puntaje promedio ponderado de las similitudes calculadas para todos sus caracteres [23]:

$$S_{\mathrm{i}j} = \frac{\sum_{k=1}^{v} S_{\mathrm{i}jk} \cdot \delta_{\mathrm{i}jk}}{\sum_{k=1}^{v} \delta_{\mathrm{i}jk}} \tag{2.2}$$

Cuando todas las comparaciones son posibles ($\sum_{k=1}^{v} \delta_{ijk} = v$), v representa el número total de caracteres. En caso contrario, v es el número de caracteres sobre los cuales se realiza la comparación.

Así, la distancia de Gower (expresada en la ecuación 2.2), varía entre 0 y 1. Un valor de 1 significa que los dos individuos no difieren en ningún carácter, mientras que 0 significa que difieren al máximo en todos sus caracteres.

2.4.3. K-means

El algoritmo K-means agrupa los datos tratando de separar muestras en n grupos de varianza igual, minimizando un criterio conocido como la inercia o la suma de los cuadrados dentro del grupo. Este algoritmo requiere que se indique el número de grupos. Este algoritmo se adapta bien a grandes cantidades de muestras y se ha aplicado en una amplia variedad de campos y áreas de estudio [3].

K-means lo que hace es dividir un conjunto de N muestras X en K grupos disjuntos C, donde cada uno de los grupos queda descrito por la media μ_j de las muestras del grupos. Estas medias son conocidas como "centroides". Cabe señalar que, en general, no son puntos de X, aunque se encuentran en el mismo espacio.

El objetivo del algoritmo k-means es seleccionar centroides que minimicen la inercia, o la suma de los cuadrados dentro del grupo, los cuales sen encuentran definidos en la ecuación 2.3. En otras palabras, k-means encuentra centros de grupos que minimizan la suma de distancias entre los datos y el centro de cada grupo [3].

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(\|x_i - \mu_j\|^2) \tag{2.3}$$

La inercia puede reconocerse como una medida de cuán coherentes son internamente los grupos, pero sufre de algunos inconvenientes:

- La inercia asume que los grupos son convexos e isotrópicos, lo cual no siempre es el caso. Por lo tanto k-means, no maneja de manera efectiva grupos alargados o variedades con formas irregulares.
- La inercia no está normalizada como métrica, se sabe que valores más bajos son preferibles, y cero es el óptimo. Sin embargo, en espacios de alta dimensionalidad, las distancias euclidianas tienden a inflarse, un fenómeno conocido como la "maldición de la dimensionalidad". Aplicar un algoritmo de reducción de dimensionalidad antes de utilizar KMeans puede mitigar este problema y mejorar la eficiencia computacional.

En términos básicos, el algoritmo K-means consta de tres pasos fundamentales. En el primer paso, se seleccionan los centroides iniciales, siendo el método más básico elegir k muestras del conjunto de datos X. Después de esta inicialización, K-means se compone de un ciclo que alterna entre los siguientes dos pasos. El primer paso asigna cada muestra a su centroide más cercano. Luego, en el segundo paso, se calculan nuevos centroides tomando el valor medio de todas las muestras asignadas a cada centroide anterior. Se calcula la diferencia

entre los centroides antiguos y los nuevos, y el algoritmo repite estos dos últimos pasos hasta que esta diferencia sea menor que un umbral predefinido. En otras palabras, se repite hasta que los centroides no se muevan significativamente [3].

2.4.4. K-modes y K-prototypes

K-means es un algoritmo diseñado para trabajar exclusivamente con datos numéricos, utilizando medidas de distancia como la euclidiana para evaluar la similitud entre ellos. Sin embargo, estas limitaciones pueden superarse mediante modificaciones específicas que permiten adaptar el algoritmo para datos categóricos:

- Utilizar una medida de disimilitud de coincidencia simple para objetos categóricos.
- En lugar de calcular medias, se utilizan modas para representar los centroides de los grupos.
- Emplear un método basado en frecuencias para encontrar las modas.

Estas adaptaciones dan origen al algoritmo conocido como K-modes. Sin embargo, cuando se trabaja con conjuntos de datos mixtos que contienen tanto variables numéricas como categóricas, se recomienda utilizar el algoritmo K-prototypes. Este método combina tanto K-means como K-modes, el algoritmo ajusta simultáneamente centroides numéricos y modas categóricas para minimizar la distancia entre los puntos de datos y sus centroides asignados, permitiendo así agrupar datos mixtos de manera efectiva [24]. Este enfoque tiene como objetivo minimizar una función de costo combinada que evalúa tanto la distancia o diferencia entre los datos [25].

2.4.5. Agrupamiento jerárquico

La agrupación jerárquica es un método de agrupación que busca construir una jerarquía de grupos. Las estrategias para la agrupación jerárquica generalmente se dividen en dos categorías:

- Aglomerativo: Este es un enfoque “de abajo hacia arriba”. Se empieza con cada punto como grupo individual, luego en cada paso se junta el par de grupos más cercano hasta que quede sólo un grupo (o k grupos).
- Divisivo: Este es un enfoque “de arriba hacia abajo”. Se empieza con un grupo que contenga todos los puntos, luego en cada paso se divide un grupo en dos hasta que todo grupo contenga un solo punto (o haya k grupos).

Para determinar qué grupos fusionar (en el caso de métodos aglomerativos) o dónde dividir un grupo (en el caso de métodos divisivos) en agrupamientos jerárquicos, se requiere una medida de disimilitud entre conjuntos de observaciones. Esta medida se logra utilizando una definición de distancia d, como la euclidiana o la distancia Gower. Además, se emplea un criterio de vinculación que especifica cómo se calcula la disimilitud entre conjuntos, tomando en cuenta las distancias entre pares de observaciones en esos conjuntos. La elección de la métrica de distancia y del criterio de vinculación puede tener un impacto significativo en el resultado del agrupamiento. La métrica determina qué objetos son más similares, mientras que el criterio de vinculación influye en la forma de los grupos resultantes [25].

Sean A y B dos conjuntos de observaciones, los criterios de vinculación más utilizados son definidos como:

- Máx - Enlace Completo: minimiza la distancia máxima entre las observaciones de pares de grupos. Este enlace es poco susceptible a valores atípicos, tiende a quebrar grupos grandes y tiende a formar grupos esféricos.

$$max\{(\mathrm{d}(x, y) : x \in A, y \in B\} \tag{2.4}$$

- Mín - Enlace Simple: minimiza la distancia entre las observaciones más cercanas de pares de grupos. Este enlace puede manejar formas no elípticas, tiende a romper grupos y es sensible a valores atípicos y ruido.

$$min\{(\mathrm{d}(x, y) : x \in A, y \in B\} \tag{2.5}$$

- Promedio: minimiza el promedio de las distancias entre todas las observaciones de pares de grupos. Este enlace es un compromiso entre los enlaces min y max, tiende a formar grupos esféricos y es menos susceptible a valores atípicos y ruido.

$$\frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} \mathrm{d}(x, y) \tag{2.6}$$

- Ward: minimiza la suma de las diferencias al cuadrado dentro de todos los grupos. Es un enfoque de minimización de la varianza y en este sentido es similar a la función objetivo de k-means, pero abordada con un enfoque jerárquico aglomerativo.

Estos algoritmos ofrecen la ventaja de permitir la selección del número de grupos cortando el dendrograma, pero también requieren un tiempo de ejecución mayor [25].

2.4.6. Índices de evaluación de la calidad de los grupos

Estos índices son útiles para evaluar la coherencia y separación de los grupos obtenidos mediante algoritmos de agrupación, ayudando a seleccionar el número óptimo de grupos o a comparar diferentes resultados de agrupamiento.

Análisis de silueta

El análisis de silueta se emplea para evaluar la separación entre los grupos obtenidos. El gráfico de silueta ilustra cuán cerca está cada punto de un grupo respecto a los puntos de los grupos vecinos, lo que ofrece una manera visual de evaluar parámetros como el número de grupos [26].

El coeficiente de silueta para cada elemento se calculan como:

$$s = \frac{b - a}{max(a, b)} \tag{2.7}$$

Donde:

- a es el promedio de las distancias entre el elemento y todos los otros elementos del grupo al cual el elemento analizado fue etiquetado.
- b es la distancia entre el elemento y el centro del grupo más cercano.

Después de calcular los coeficientes individuales, el coeficiente de silueta se obtiene promediando todos estos valores.

Este análisis produce coeficientes de silueta, con un rango de $[-1, 1]$. Los coeficientes de silueta cercanos a $+1$ indican que la muestra está claramente separada de los grupos vecinos. Un valor de 0 sugiere que la muestra está en o cerca del límite de decisión entre dos grupos vecinos, mientras que valores negativos indican que esas muestras podrían haber sido asignadas incorrectamente al grupo [26].

Índice de Davies-Bouldin

Fue propuesto por David L. Davies y Donald W. Bouldin en 1979. Este índice busca medir qué tan bien separados y compactos están los grupos en una partición de datos. Cuantifica la relación entre la distancia media dentro de los grupos (dispersión) y la distancia media entre los grupos. Se busca minimizar este índice, donde valores más bajos indican una mejor separación entre los grupos y una mayor compactación dentro de los grupos, lo que sugiere una mejor partición de grupos [27].

$$DB = \frac{1}{N} \sum_{\mathrm{i}=1}^{N} R_{\mathrm{i}} \tag{2.8}$$

Donde N es el número de grupos y R_{i} es el valor máximo de $R_{\mathrm{i}j}$ con $\mathrm{i} \neq j$, el cual se define como:

$$R_{\mathrm{i}j} = \frac{S_{\mathrm{i}} + S_j}{M_{\mathrm{i}j}} \tag{2.9}$$

Donde S_{i} y S_j son la dispersión de los grupos i y j respectivamente, y $M_{\mathrm{i}j}$ es la distancia entre los centroides.

2.5. Manejo de valores faltantes

Manejar los valores faltantes en un conjunto de datos es crucial para el éxito de un modelo de aprendizaje automático debido a que los datos incompletos pueden introducir sesgos. Los valores faltantes pueden llevar a una interpretación errónea de la información y afectar la capacidad del modelo para aprender patrones y relaciones relevantes. Si no se abordan adecuadamente, estos datos incompletos pueden resultar en una disminución de la precisión del modelo y en una mayor variabilidad en los resultados [28]. Los valores faltantes pueden clasificarse en tres tipos principales:

- Completamente al azar (MCAR): se caracterizan por la ausencia de un patrón que pueda ser explicado por otras variables en el conjunto de datos. En este caso, la falta de datos ocurre de manera aleatoria, y cualquier observación tiene la misma probabilidad de faltar sin que esto dependa de otras variables. En este caso, es útil considerar que cada observación tiene una probabilidad de faltar y que las que no están presentes se han perdido de manera aleatoria. Una solución común para manejar este tipo de datos es eliminar las filas o columnas que no contengan los datos o imputar valores.
- Al azar (MAR): se refiere a la situación en la que la pérdida de datos está relacionada con alguna otra variable del conjunto de datos, pero no con la variable en sí misma. Aunque la falta de información está influenciada por otras variables, sigue siendo aleatoria en su naturaleza. En este caso, la falta de información no es completamente aleatoria. La solución más común para manejar este tipo de datos faltantes es la imputación de valores, utilizando información de las variables relacionadas para estimar los valores faltantes.
- No al azar (MNAR): se refiere a la situación en la que los datos faltantes están directamente relacionados con la propia variable en cuestión. Este es el peor tipo de datos faltantes, ya que puede conducir a estadísticas descriptivas sesgadas y afectar gravemente la precisión de las inferencias y predicciones. Para abordar este problema, es recomendable intentar recuperar los datos faltantes de la fuente original del conjunto de datos o cruzar la información con otros conjuntos de datos que puedan proporcionar los datos necesarios.

La decisión sobre qué solución tomar también puede verse afectada por el porcentaje de datos faltantes, la importancia de la columna y el impacto en el análisis o modelo.

2.6. Bases de datos SIG para sistemas agrivoltaicos

A continuación, se llevó a cabo un análisis de diversas herramientas SIG utilizadas en proyectos agrivoltaicos a nivel global. También, se indicaran los datos presentados por estas herramientas y se señalarán los posibles aspectos de mejoras de estas.

2.6.1. inSPIRE Agrivoltaics Map

OpenEI es una página web que sirve como una base de datos especializada en energías renovables y eficiencia energética. Gran parte de la información que se encuentra en OpenEI es generada por la comunidad, lo que significa que es una colección de contenido y datos proporcionados por colaboradores que incluyen expertos y entusiastas. Los usuarios de OpenEI pueden acceder, editar, añadir y descargar datos de forma gratuita. Actualmente, OpenEI cuenta con más de 230034 páginas que abarcan una amplia variedad de temas relacionados con la energía, desde energía renovable y eficiencia energética hasta políticas y regulaciones, así como aprendizaje automático y análisis de datos [29].

OpenEI es gestionada y desarrollada por el "National Renewable Energy Laboratory" (NREL), un laboratorio estadounidense especializado en el desarrollo e investigación de energías renovables, eficiencia energética, integración de sistemas de energía y transporte sostenible. El NREL recibe financiamiento y respaldo del "U.S Department of Energy" (DOE).

Dentro de OpenEI, se encuentra un mapa dinámico, llamado "inSPIRE Agrivoltaics Map", que representa un censo de las distintas instalaciones ubicadas en todo Estados Unidos [30]. A continuación, se describen los datos relevantes proporcionados por esta base de datos:

- Actividad agrivoltaica, se diferencia según si el sistema agrivoltaico será utilizado como hábitat para agentes polinizadores, para el pastoreo de ganado, el cultivo de vegetales o como techo solar de un invernadero.
- Tamaño del sistema, medido en megawatt.
- Tamaño del sitio, medido en acres.
- Tipo de tecnología fotovoltaica utilizada, por ejemplo, si los paneles son monofaciales o bifaciales.
- Tipo de arreglo, si los paneles son fijo o cuentan con seguimiento en alguno de sus ejes.
- Tipo de plantación o ganado.
- Servicio al ecosistema, por ejemplo, si el proyecto es amigable con los distintos agentes polinizadores o si la vegetación utilizada es nativa de la zona.
- Orientación del panel, esto solo si es de arreglo fijo.
- Altura del panel con respecto del suelo.

Por otro lado, esta herramienta es exclusiva para proyectos ubicados en suelo norteamericano, lo que significa que no incluye proyectos agrivoltaicos que se encuentren en Europa, Asia o otras partes del mundo. Otra desventaja de esta herramienta se relaciona con su carácter comunitario, lo que conlleva que muchos proyectos no estén completamente caracterizados, ya que la información proviene de la comunidad. Por último, es importante señalar que esta herramienta no proporciona información meteorológica sobre la ubicación de los proyectos, como la temperatura, la radiación solar o el tipo de clima.

2.6.2. Mapa agrivoltaico de India

Este SIG fue desarrollado y es mantenido por la "National solar energy federation of India" (NSEFI), un organismo defensor de las políticas solares en India y una organización paraguas que representa a empresas de energía solar a lo largo de toda la cadena de valor solar. La NSEFI colabora de manera complementaria con los gobiernos central y estatal para lograr los objetivos nacionales de India de alcanzar 100 [GW] para 2022 y un objetivo de energías renovables de 450 [GW] para 2030 [31].

En lo que respecta al ámbito agrivoltaico, la NSEFI ha producido varios informes abordando la situación legal, los desafíos futuros y desarrollos de proyectos agrivoltaicos en India. En este último aspecto, destaca la creación de un sistema SIG que presenta los proyectos agrivoltaicos censados hasta la fecha.

El SIG realiza una distinción entre proyectos AgroPV completamente elevados y aquellos que integran la agricultura como complemento de la planta de energía. Además, proporciona una descripción detallada para cada proyecto, que no se limita únicamente a texto, sino que también puede incluir contenido multimedia como imágenes y vídeos [32]. Sin embargo, es

importante señalar que la información en la descripción no siempre abarca todos los detalles que podrían estar disponibles para otros proyectos. En muchos casos, estas descripciones omiten información sobre la capacidad de la instalación o el tipo de cultivo utilizado, simplemente redirigiendo al usuario a un enlace externo al SIG. Este planteamiento presenta un problema significativo, ya que muchos de estos enlaces están rotos, lo que dificulta el acceso a los datos.

Los enlaces que permanecen activos dirigen al usuario a un resumen de proyectos en funcionamiento en India, proporcionando datos relevantes como la capacidad de las plantas de energía, los cultivos usados, los principios operativos y aspectos tecnoeconómicos. Sin embargo, la principal desventaja de esta herramienta radica en la falta de estandarización en la visualización de datos para todos los proyectos, obligando al usuario a buscar información fuera del SIG, como se mencionó anteriormente. Otro aspecto desfavorable es que ni el SIG ni los enlaces externos incorporan datos climatológicos de la zona donde se ubican los proyectos.

2.6.3. Mapa de instalaciones agrivoltaicas en Alemania

Este mapa interactivo muestra las instalaciones agrivoltaicas en Alemania, creado por el Fraunhofer ISE, un instituto de investigación solar de origen alemán.

El mapa proporciona información detallada sobre cada instalación, incluyendo el nombre del proyecto, el tamaño del sistema medido en [kWp], la superficie ocupada en hectáreas, el tipo de cultivo utilizado y la categoría del proyecto. La categoría indica si el proyecto está clasificado como "entre hileras" (paneles con altura libre vertical menor a 2.1 metros) o "sobre cultivo" (paneles con altura libre vertical superior a 2.1 metros). Además, la herramienta permite completar una encuesta para añadir nuevos proyectos al mapa. No obstante, al igual que otras herramientas similares, no incluye datos meteorológicos o climatológicos de las zonas donde se encuentran los proyectos registrados [33].

Capítulo 3

Metodología propuesta

3.1. Preámbulo

En esta sección, se detalla la propuesta metodológica que es utilizada en el desarrollo de este trabajo, la cual se muestra en la figura 3.1. Además, se especifica que herramienta y bibliotecas fueron utilizadas para la implementación.

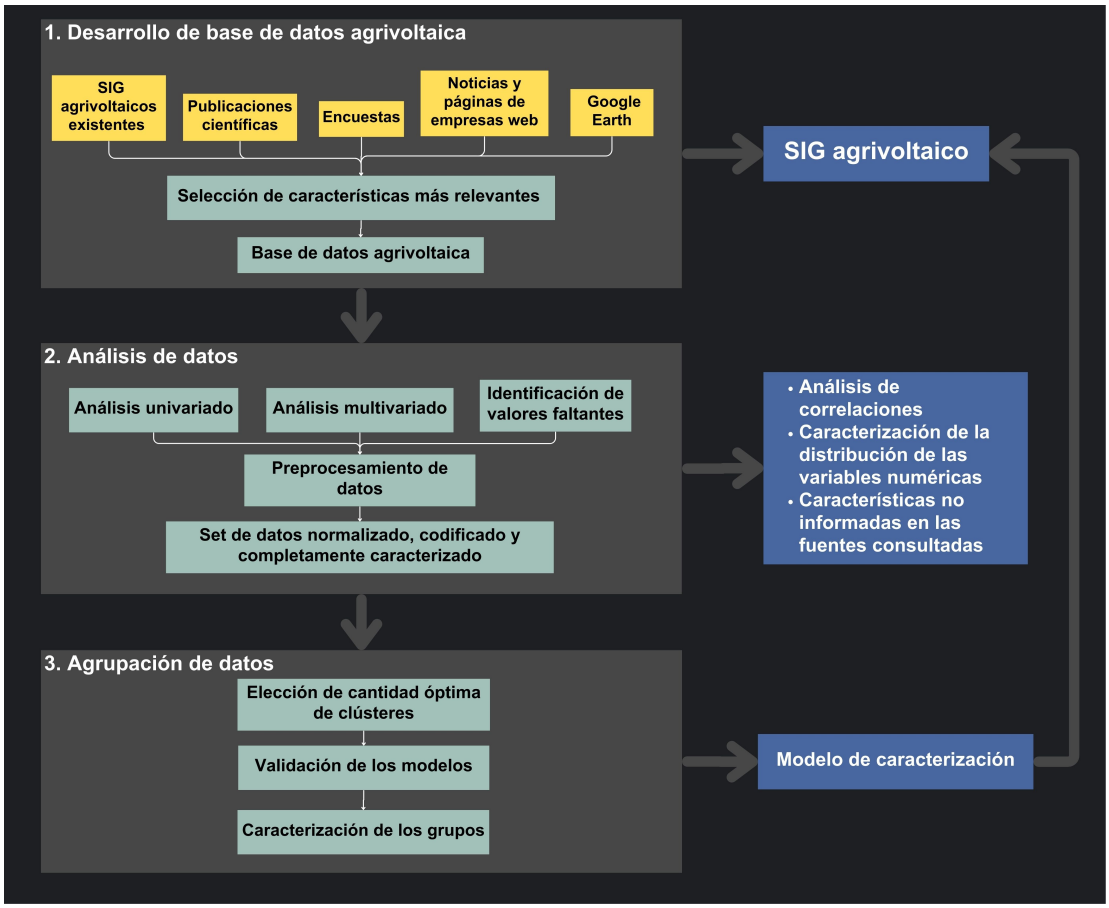

Figura 3.1: Diagrama de propuesta metodológica.

La propuesta metodológica se divide en tres partes, cada una representada en color gris, correspondiendo a los objetivos específicos definidos en la sección 1.2. El primer recuadro gris corresponde al proceso necesario para cumplir con el objetivo específico número 1, el segundo recuadro gris al objetivo específico número 2, y el tercer recuadro gris abarca los objetivos específicos 3 y 4. Los recuadros amarillos representan los datos iniciales del proyecto, los recuadros celestes son los subprocesos obtenidos a medida que avanza el trabajo, y los recuadros azules indican los resultados esperados de cada proceso. Es importante mencionar que el modelo de caracterización forma parte del SIG agrivoltaico como una característica, lo que genera una retroalimentación entre los recuadros azules.

A continuación, se detalla cada paso necesario para la elaboración de la memoria, según lo mencionado anteriormente.

3.2. Paso 1: Desarrollo de base de datos agrivoltaica

El primer paso consiste en buscar proyectos agrivoltaicos georeferenciados utilizando varias fuentes, como sistemas de información geográfica (SIG) especializados en agrivoltaica, publicaciones científicas, encuestas a proyectos existentes, noticias sobre iniciativas en curso y sitios web de empresas dedicadas al sector. Durante esta búsqueda de información, puede ser necesario complementar utilizando herramientas adicionales como Google Earth.

No toda la información obtenida es relevante, por lo que es necesario realizar una selección de características. Para esto se debe analizar la información de fichas de proyectos tanto fotovoltaicos como agrivoltaicos.

Una vez seleccionadas estas características más relevantes, toda la información debe ser incorporada en una base de datos del tipo relacional. Para la gestión de esta base de datos se utiliza MariaDB.

Con la base de datos ya desarrollada, el siguiente paso es crear una herramienta que contenga un mapa con marcadores, los cuales muestran las características relevantes de todos los proyectos registrados. Esta herramienta es una página web, desarrollada completamente en lenguaje HTML. Los mapas para la visualización de los datos son elaborados con Leaflet, una biblioteca de JavaScript de código abierto utilizada para crear mapas interactivos.

3.3. Paso 2: Análisis de datos

Con el resultado del proceso anterior, se procede a realizar un análisis exploratorio de los datos con el fin de descubrir patrones, relaciones y tendencias, así como identificar cualquier anomalía o error presente. Para ello, primero se realiza un análisis univariado, un análisis multivariado y una identificación de valores faltantes. Sin embargo, antes es necesario utilizar la biblioteca SQLAlchemy para trabajar con la base de datos en Python, junto con la biblioteca Pandas para manipular los datos de forma eficiente.

El análisis univariado se enfoca en examinar cada variable de forma individual. Para realizar esto, se calculan medidas de tendencia central, como la media, mediana y moda, que ayudan a resumir la ubicación central de los datos. Además, se calculan medidas de

dispersión para las variables numéricas, como la desviación estándar, varianza y rango, que miden cuánto se desvían los datos de la tendencia central. También se utilizan gráficos de histogramas para observar la frecuencia de cada valor en una variable.

El análisis multivariado implica el estudio simultáneo de dos o más variables en un conjunto de datos para determinar las relaciones entre ellas. Para las variables numéricas, se utiliza el análisis de correlaciones, una técnica que permite identificar patrones y relaciones entre las distintas variables de un conjunto de datos. La correlación se puede medir utilizando diferentes métodos, como el coeficiente de correlación de Pearson. Para las variables categóricas, se emplean tablas de contingencia, las cuales permiten calcular el número de ocurrencias de una variable para cada una de sus categorías en comparación con los valores de otra variable.

La identificación de valores faltantes es un subproceso que consiste en buscar y detectar la cantidad y ubicación de estos valores. El objetivo de este subproceso es determinar la mejor manera de manejar los datos faltantes, evaluando si es más adecuado eliminar las columnas con datos faltantes o imputar los valores ausentes.

A partir de estos análisis, se obtiene una caracterización de la distribución de las variables, se determina qué variables no aportan información relevante al modelo y se decide la mejor manera de tratar los datos faltantes. Con esta información, se procede al preprocesamiento de los datos con la biblioteca de scikit-learn, que incluye escalar las variables numéricas si se asemejan a una distribución normal, de lo contrario, se busca aproximar la distribución arbitraria a una gaussiana. Además, se codifican las variables categóricas y se imputan los datos faltantes.

Una vez realizado el preprocesamiento, se obtiene un conjunto de datos completamente caracterizados, el cual servirá como entrada para el siguiente proceso: el agrupamiento de datos.

3.4. Paso 3: Agrupación de datos

Ahora, con los datos completamente listos para el agrupamiento de datos, es necesario determinar la cantidad óptima de grupos a formar para los algoritmos de k-prototypes y jerárquico. Para esto, se utilizan índices de evaluación como el puntaje de silueta y el índice de Davies-Bouldin. Para trabajar con los algoritmos mencionados, se hace uso de las librerías de Python kmodes y SciPy.

Una vez obtenidos el número óptimo de grupos para cada uno de los algoritmos, se procede a realizar la agrupación de los datos. Posteriormente, se validan y comparan los modelos utilizando los índices mencionados, y se escoge el algoritmo que ofrezca un mejor desempeño para la construcción del modelo.

Una vez obtenido el modelo, se analizan los grupos formados y se identifican las características que comparten estos grupos. Este resultado también forma parte del SIG agrivoltaico, por lo que debe ser incluido como una característica adicional.

3.5. Herramientas y biblioteca utilizadas

A continuación, se detallan las herramientas y bibliotecas más relevantes utilizadas para la realización de este trabajo.

3.5.1. Google Earth

Google Earth es un programa y servicio desarrollado por Google que permite visualizar imágenes satelitales, mapas y datos geográficos de la Tierra en 3D. Esta herramienta facilita la obtención de información geográfica, como la altitud sobre el nivel del mar de proyectos específicos o el tipo de clima presente en una región. Sin embargo, la identificación detallada del tipo de clima en un territorio específico no está disponible de forma predeterminada en Google Earth. Para obtener esta información, es necesario descargar archivos KML, un formato basado en XML utilizado para representar datos geográficos en aplicaciones de mapeo como Google Earth, Google Maps y otras plataformas de información geográfica.

En este caso, se ha seleccionado un mapa mundial de clasificación climática de Köppen-Geiger de alta resolución (5 minutos de arco) [1], que es representativo para el periodo de 1986-2010, que fue realizado a partir del trabajo realizado por Rubel et al. (2017) [34]. Este mapa proporciona una clasificación detallada de los tipos climáticos en diferentes áreas geográficas, siendo fundamental para estudios climáticos, agrícolas y análisis geoespaciales. A continuación, se muestra en la figura 3.2 cómo se vería el mapa mundial de clasificación climática dentro de la aplicación Google Earth.

Figura 3.2: Mapa mundial de clasificación climática de alta resolución dentro de la aplicación Google Earth.

3.5.2. MariaDB

Para la gestión de la base de datos, se utilizó MariaDB, un sistema de gestión de bases de datos relacional de código abierto derivado de MySQL. El uso de un gestor de bases de

[1] Link de descarga del archivo KML

datos es fundamental para la creación y administración eficiente de una base de datos. Un gestor de bases de datos proporciona una estructura organizada para almacenar y recuperar datos, asegurando que la información esté correctamente indexada y sea fácilmente accesible. Esto mejora la eficiencia y rapidez de las consultas, lo que es esencial para aplicaciones que requieren acceso rápido a grandes volúmenes de datos. La elección de MariaDB sobre otros gestores se debe a la experiencia previa con este sistema. Además, al ser un derivado de MySQL, es relativamente fácil migrar a otro gestor compatible con MySQL.

3.5.3. Bibliotecas de python

Entre las bibliotecas más relevantes utilizadas en este trabajo se destacan: SQLAlchemy, Pandas, scikit-learn, kmodes, gower y SciPy.

- **SQLAlchemy**: es una biblioteca para trabajar con bases de datos en Python. Simplifica la interacción con las bases de datos relacionales al permitir a los desarrolladores trabajar con objetos Python en lugar de escribir consultas SQL manuales.
- **Pandas**: es una biblioteca de análisis de datos en Python que proporciona estructuras de datos flexibles y herramientas eficientes para la manipulación y análisis de datos numéricos y de series temporales. Es ampliamente utilizada para tareas de limpieza, transformación y análisis de datos, facilitando el manejo de grandes conjuntos de datos.
- **scikit-learn**: es una biblioteca de aprendizaje automático para Python que ofrece una amplia gama de algoritmos de aprendizaje automático, herramientas para modelado predictivo y análisis de datos. En este trabajo, se utilizaron específicamente sus herramientas de preprocesamiento para escalar, transformar e imputar variables, así como para aplicar métricas en la evaluación de la agrupación de datos.
- **kmodes**: corresponde a la biblioteca utilizada para la implementación de los algoritmos de agrupamiento kmodes y kprotoypes.
- **gower**: es una biblioteca en Python para calcular la distancia de Gower, una métrica que se utiliza para calcular la similitud entre observaciones en conjuntos de datos mixtos. Es útil en tareas de agrupamiento.
- **SciPy**: es una biblioteca de Python que ofrece rutinas y algoritmos eficientes para el cálculo científico y técnico. En este trabajo, se utiliza específicamente para implementar el algoritmo de agrupamiento jerárquico.

Capítulo 4

Desarrollo de base de datos agrivoltaica

En este capítulo se aborda la recopilación de información sobre los proyectos agrivoltaicos, incluyendo las fuentes de información consultadas. Además, se explica la construcción y modelación de la base de datos relacional, detallando los pasos seguidos para diseñar la estructura de la base de datos, la normalización de las tablas, y la implementación de las relaciones y dependencias funcionales entre los diferentes conjuntos de datos. También se presenta el SIG desarrollado a partir de la base de datos.

4.1. Fuentes de información

El primer paso en la construcción de la base de datos consiste en buscar información sobre proyectos agrivoltaicos. Para obtener esta información, se consultan los Sistemas de Información Geográfica (SIG) agrivoltaicos mencionados en la sección 2.6. Además, se revisan sitios de noticias relacionados con el ámbito fotovoltaico y agrivoltaico, páginas web de empresas dedicadas a la realización de proyectos agrivoltaicos, y artículos científicos pertinentes. A través de los SIG agrivoltaicos se obtiene información de proyectos en India, EE. UU., Canadá, Puerto Rico y Alemania, mediante artículos científicos se obtiene información de los Países Bajos, y a través de sitios web de noticias y empresas se recaba información de proyectos en China, España, Israel, Suiza, Francia e Italia. En el caso de Chile, la información se obtiene mediante un censo de los distintos proyectos. En el anexo A se puede encontrar un mapa de los proyectos registrados en Chile, elaborado a partir de la información obtenida.

Hasta el momento, se han registrado 586 proyectos, la gran mayoría pertenecientes al SIG de OpenEI, que es la base de datos más extensa consultada hasta la fecha. Es importante volver a mencionar que existen más iniciativas en otras partes del globo, pero como no es posible georreferenciarlas, no se consideraron para la construcción de la base de datos.

4.2. Características relevantes

Para la selección de las características relevantes, se estudia la estructura de las fichas informativas de proyectos fotovoltaicos y agrivoltaicos, las cuales incluyen una serie de características importantes de estos proyectos. Dado que los proyectos agrivoltaicos son esencial-

mente proyectos fotovoltaicos con la adición de cultivos en la superficie donde se encuentran los paneles solares, se opta por considerar varias de las características relevantes mostradas en estas fichas. Además, se agregan características específicas para describir los cultivos y el clima. Las características relevantes son:

- ID: corresponde a un número único e identificador para cada proyecto.
- Nombre del proyecto.
- Dueño o propietario del proyecto.
- Empresa encarga del desarrollo del proyecto.
- Latitud en la que se encuentra el proyecto, medida en grados decimales.
- Longitud en la que se encuentra el proyecto, medida en grados decimales.
- Capacidad instalada, medida en megawatt [MW].
- Superficie: área que ocupa el proyecto, medida en hectáreas.
- Periodo de inicio: año en el que se inició la operación del proyecto.
- Altura sobre el nivel del suelo en la que se encuentran los módulos fotovoltaicos, medido en metros. En caso de tener seguimiento, se considera la distancia entre el suelo y el eje.
- Altitud: metros sobre el nivel del mar en los que se encuentra el proyecto.
- País en el cual se encuentra el proyecto.
- Región: subdivisión administrativa en la que se encuentra el proyecto.
- Clima: tipo de clima de la zona del proyecto según la clasificación de Köppen-Geiger. La clasificación de Köppen-Geiger es un sistema de categorización climática desarrollado por el climatólogo alemán Wladimir Köppen. Este sistema clasifica los climas del mundo en varias categorías basadas en la temperatura y la precipitación, y está diseñado para reflejar los patrones de vegetación que son típicos de cada clima.
- Tipo de cultivo: especifica el tipo de cultivo presente en el proyecto. Los tipos de cultivo son: frutales, cereales, hortalizas, praderas, berries, leguminosas, vides o cultivos industriales.
- Tipo de diseño: corresponde al tipo de disposición de los paneles. Se distingue entre las siguientes cuatro categorías:
    1. Entre hilera: los cultivos están entre las hileras de paneles solares.
    2. Sobre cultivo: los paneles están sobre una estructura encima de los cultivos.
    3. Invernadero: los paneles forman parte del techo de un invernadero.
    4. Vertical: los paneles están ubicados verticalmente con respecto al suelo.
- Tipo de panel: se distingue entre monofacial, bifacial o semitransparente.
- Tipo de seguimiento: se distingue entre arreglo fijo, seguimiento de un eje o seguimiento de dos ejes.
- Estado del proyecto: el proyecto se encuentra en operación, en construcción (esta categoría incluye proyectos aprobados o proyectados) o fuera de servicio.

4.3. Modelo entidad-relación de la base de datos

El primer paso para la construcción del modelo es definir las entidades, sus atributos y cómo se relacionan entre sí. La idea es que este modelo abarque todas las características relevantes.

La primera y más importante entidad es el "proyecto agrivoltaico" en sí mismo. Un proyecto agrivoltaico tiene los siguientes atributos: número identificatorio (ID), nombre, dueño, desarrollador, latitud, longitud, capacidad [MW], superficie [ha], año de inicio de operación, altitud respecto al nivel del mar y la altura entre los módulos fotovoltaicos y el suelo. Aunque se podría pensar que la altura respecto al suelo es un atributo de otra entidad llamada "panel", la realidad es que un tipo de panel puede estar a diferentes alturas. Esto complicaría el modelo, por lo que es mejor considerar que la altura es una característica propia de cada proyecto agrivoltaico.

Otra entidad es la ya mencionada "panel". Los paneles tienen como atributos un ID y una clase o tipo. Entre sus tipos se distinguen tres: monofacial, bifacial y semitransparente. Los paneles solo tienen una relación, y es con la entidad "proyecto agrivoltaico", dado que un proyecto agrivoltaico incluye paneles fotovoltaicos para la generación de energía. Esta relación puede definirse como M:M , ya que un proyecto agrivoltaico puede tener muchos tipos de paneles y muchos tipos de paneles pueden estar presentes en un proyecto agrivoltaico.

De igual manera los proyectos agrivoltaicos tiene un "diseño", entidad que hace referencia a la disposición de los paneles. Esta entidad tiene como atributos un ID y un tipo de disposición, pudiendo ser estas: entre hilera, sobre cultivo, invernadero o vertical. Esta relación es M:M, lo que significa que un proyecto puede tener muchos diseños (disposiciones) y que estos diseños pueden estar en varios proyectos.

Además, los proyectos agrivoltaicos pueden tener o no un sistema de seguimiento en los paneles, lo que introduce una relación con la entidad "seguimiento". Esta entidad tiene como atributos un ID y el tipo de seguimiento, que puede ser fijo, de un eje, o de dos ejes. La relación es M:M, ya que un proyecto puede incluir varios tipos de seguimiento y un tipo de seguimiento puede estar presente en varios proyectos.

Cada proyecto, al tener una ubicación geográfica específica, está asociado a un tipo de clima correspondiente a esa ubicación. Por lo tanto, se introduce una nueva entidad llamada "clima". Esta entidad tiene como atributos un ID y una clase, donde las clases corresponden a la clasificación de Köppen-Geiger. Esta relación es 1:M, ya que un proyecto tiene un único clima asociado, pero un clima puede estar presente en varios proyectos.

En un proyecto agrivoltaico, no solo se genera energía, sino que también es importante la producción agrícola. Por ello, se introduce la entidad "cultivo". Esta entidad tiene como atributos un ID y una categoría, que puede ser una de las siguientes: frutales, cereales, hortalizas, praderas, berries, leguminosas, vides y cultivos industriales. Esta relación es M:M, ya que no hay restricciones sobre la cantidad de cultivos que puede haber en un proyecto agrivoltaico, y un cultivo puede estar presente en varias ubicaciones al mismo tiempo.

También, se tiene la entidad "país", que tiene como atributos un ID y un nombre. A su vez

esta entidad se relaciona con otra entidad llamada “región”. La relación entre estas entidades, es que un país contiene varias regiones, pero una región solo pertenece a un país, por lo que la relación es de 1:M.

Por otro lado, una “región” puede contener varios “proyectos agrivoltaicos”, estableciendo así una relación entre estas entidades. La entidad “región” tiene como atributos un ID y un nombre. Sin embargo, aunque una región puede tener múltiples proyectos, cada proyecto solo puede pertenecer a una región. Por lo tanto, la relación entre “región” y “proyectos agrivoltaicos” es de 1:M.

El estado de un proyecto es único para cada proyecto, por lo que se establece una relación 1:M entre las entidades “estado” y “proyectos agrivoltaicos”. La entidad “estado” cuenta con los atributos ID y tipo de estado.

Cabe mencionar que muchas de las relaciones M:M expuestas anteriormente, al observar los datos, son más bien del tipo 1:M. Por ejemplo, los proyectos agrivoltaicos catastrados tienden a tener un solo tipo de seguimiento. Sin embargo, esto no significa que en el futuro no pueda aparecer algún tipo de proyecto con más de un tipo de seguimiento. La idea de este modelado entidad-relación y su posterior modelo relacional no es solo almacenar datos de la manera más óptima posible, sino también garantizar que sea fácil de mantener a lo largo del tiempo.

Como resultado se obtiene el siguiente diagrama de modelo entidad-relación (figura 4.1). Este muestra las entidades, atributos y las relaciones descritas anteriormente.

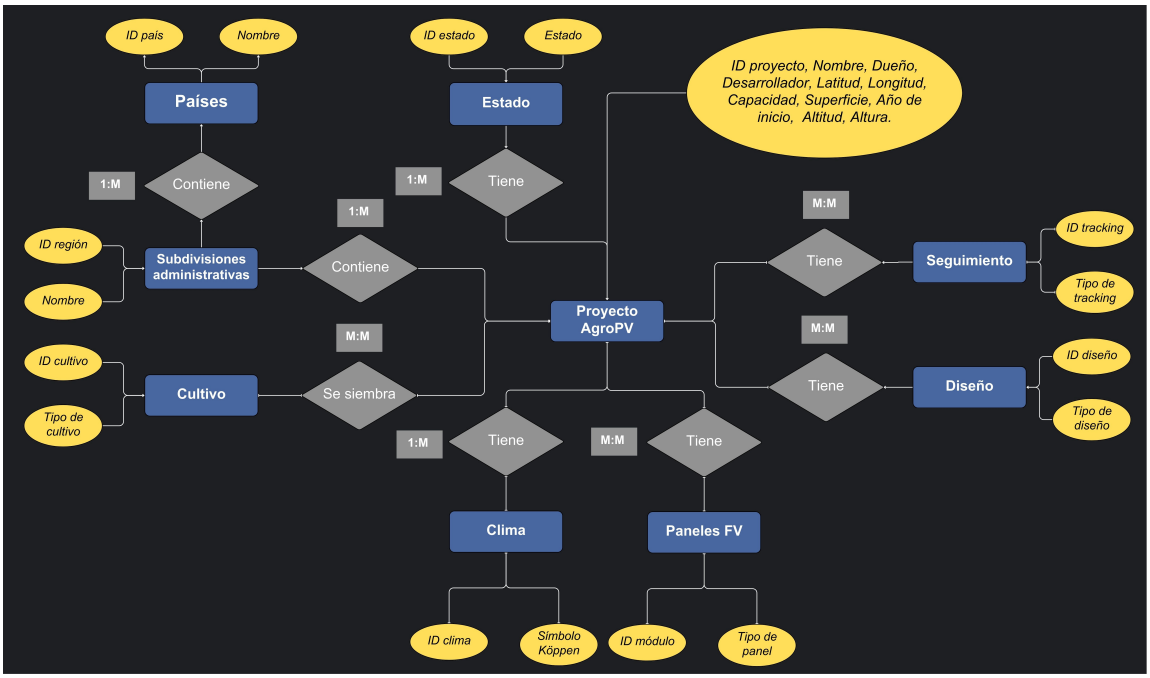

Figura 4.1: Diagrama entidad-relación para el caso de estudio.

4.4. Modelo relacional de la base de datos

El modelo entidad-relación debe ser transformado en un modelo relacional, el cual se desarrolló en DataGrip, un entorno de desarrollo integrado (IDE) para bases de datos creado por JetBrains. Este IDE es compatible con el gestor de bases de datos MariaDB.

Para llevar a cabo este proceso, se deben seguir los pasos explicados en la sección 2.3.2 para pasar de un modelo entidad-relación a un modelo relacional. En primer lugar, se deben convertir las entidades fuertes en tablas. En este caso de estudio, las entidades fuertes son: "panel", "diseño", "seguimiento", "clima", "cultivo", "país" y "estado". El nombre de estas tablas es el mismo que el de las entidades, y los campos o columnas de la tabla serán los atributos de las entidades. La clave primaria de cada tabla es el número identificatorio (ID) de cada entidad.

Entonces, por ejemplo, en el caso de la entidad "país", se tiene una tabla con el nombre "*Pais*", donde sus atributos serán "id_*pais*", que actuará como la clave primaria, y "Nombre". Es importante mencionar que al trabajar con sistemas de gestión de bases de datos, se recomienda evitar el uso de caracteres especiales, aunque en la actualidad la gran mayoría de los sistemas de gestión pueden manejar correctamente estos caracteres. Por lo tanto, la palabra "país" se escribirá sin tilde en el nombre de la tabla.

En el caso de las entidades débiles, como la entidad "región" y "proyectos agrivoltaicos", el nombre de las tablas será el mismo que el de las entidades. Los campos de la tabla incluirán los atributos de la entidad junto con la clave primaria de la entidad a la que depende, por ejemplo en el caso de "región" es la llave primaria de la entidad "país". La clave primaria de la tabla es el ID de la entidad.

Para las relaciones 1:M solo basta agregar una nueva columna que incluya el identificador primario de la entidad "M" como llave foránea. Por ejemplo, "*Proyectos*" a demás de contar en las columnas con los atributos de la entidad a la que corresponde esta tabla, se van agregar columnas con llaves primarias de las tablas "*Clima*", "*Region*" y "*Estado*".

Las relaciones M:M deben contar con sus propias tablas. El nombre de las tablas va ser el nombre de las tablas que se relacionan, mientras que las columnas contendrán los identificadores primarios de las entidades relacionadas. Para la clave primaria de la tabla, es preferible crear un nuevo número identificatorio que permita identificar cada una de las relaciones de manera única.

Aun así, esto no es todo para crear la base de datos relacional en un sistema de gestión de base de datos, también es necesario especificar el tipo de dato que se va a almacenar en cada una de las columnas. Todos los ID serán definidos como "int" , es decir, un tipo de dato entero de longitud fija que puede almacenar valores numéricos enteros. Las columnas de las tablas '*Pais*", "*Region*", "*Clima*", "*Panel*", "*Cultivo*", "*Seguimiento*", "*Diseno*" y "*Estado*" que no sean ID serán del tipo "tinytext", utilizado para almacenar cadenas de texto cortas de hasta 255 caracteres. Las columnas "*Nombre*", "*Dueno*", "*Desarrollador*" de la tabla "*Proyectos*" también serán de tipo "tinytext" , mientras que "*Latitud*", "*Longitud*", "*Capacidad*", "*Superficie*", "*Altura*" y "*Altitud*" serán del tipo "float", un tipo de dato utilizado para representar números de punto flotante, es decir, números que pueden tener una

parte decimal variable. Por último, la columna '*Periodo_de_inicio*" sera del tipo "int".

La figura 4.2, muestra el modelo relacional modelado a partir del modelo entidad-relación de la figura 4.1. Además, se decidió agregar una columna extra a las tablas "*Clima*", "*Diseno*", "*Estado*" que no se contemplaba inicialmente en el modelo entidad-relación. Esta columna adicional se denomina "descripcion" y es del tipo "tinytext" , sirviendo como un texto explicativo para cada una de las categorías almacenadas en las respectivas tablas.

También cabe destacar que, a pesar de que los cultivos tienen una relación M:M con los proyectos, la forma en que se construyó la base de datos puede generar problemas al trabajar con Pandas, ya que los proyectos con dos tipos de cultivos aparecerán duplicados. La solución elegida para afrontar este desafío fue tener dos columnas: una para el cultivo principal y otra para el cultivo secundario.

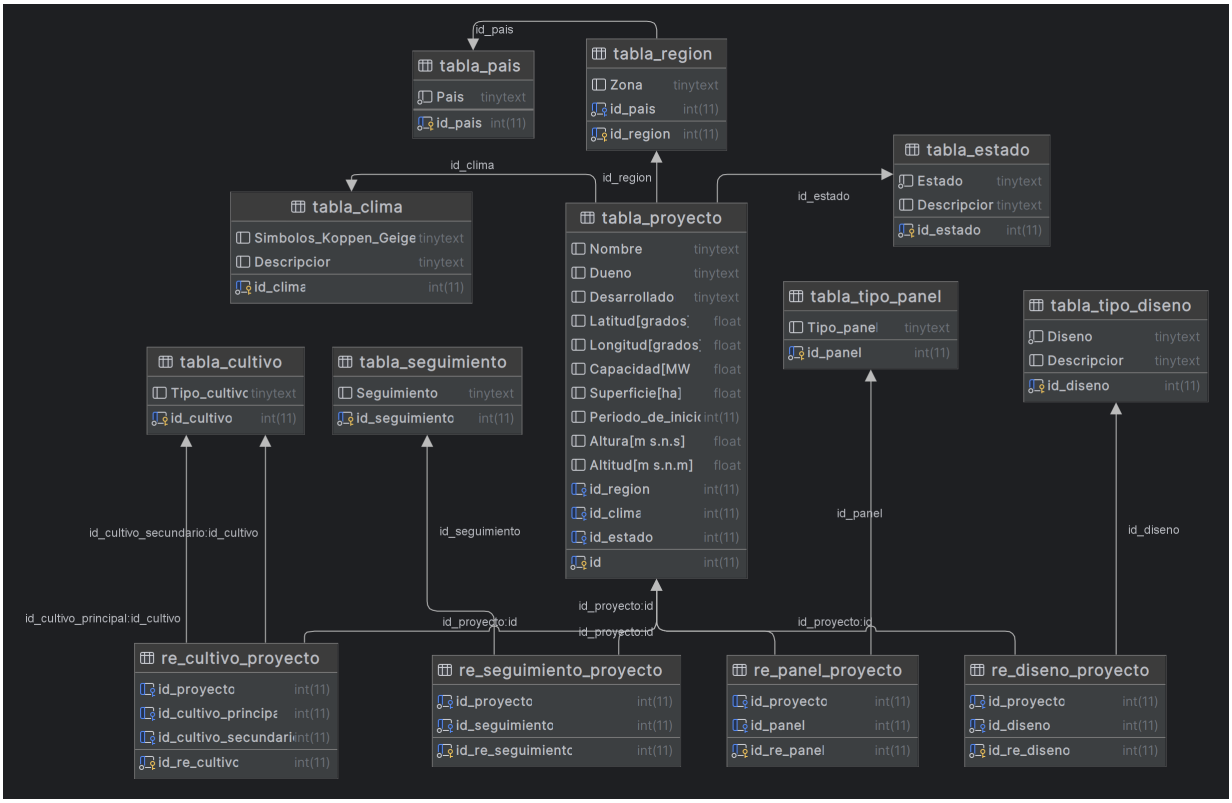

Figura 4.2: Modelo de la base de datos relacional (imagen realizada en DataGrip).

4.5. SIG agrivoltaico

Con la base de datos ya construida, se puede proceder con la elaboración del sistema de información geográfica (SIG) de proyectos agrivoltaicos. La idea tras el SIG es la visualización de la información de una manera que el usuario pueda interpretarla fácilmente e interactuar con ella. Es importante destacar que la herramienta fue completamente elaborada en lenguaje HTML, un lenguaje estándar utilizado para crear y diseñar páginas web estáticas.

La herramienta cuenta con un mapa interactivo que muestra el catastro de instalaciones agrivoltaicas alrededor del mundo y sus características más relevantes. El usuario puede agru-

par los proyectos según alguna de las siguientes características: capacidad, superficie, altura, altitud, clima, tipo de cultivo, tipo de panel, tipo de seguimiento y tipo de diseño. También se incluye la característica de "agrupación", que es el resultado del modelo de agrupamiento (ver sección 6.3). Cada característica tiene su propio mapa con marcadores que indican la ubicación de los proyectos, acompañados de una leyenda. Estos marcadores incluyen una pequeña ficha con la información del proyecto y están asociados a un color que representa la categoría a la que pertenece según la característica que se esté observando, como se ilustra en la figura 4.3.

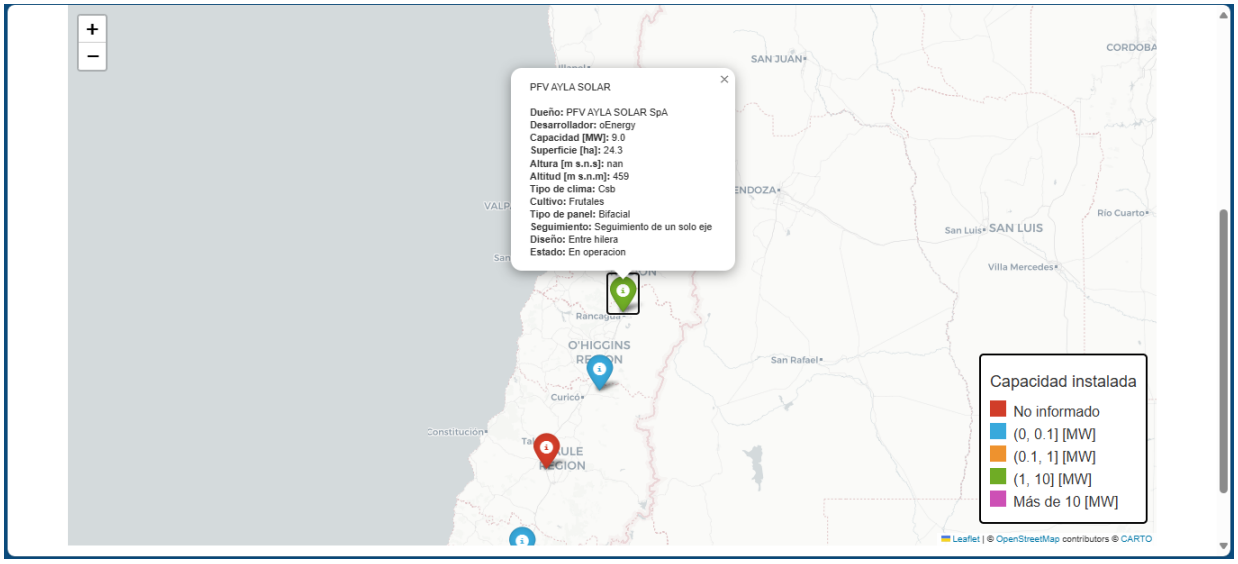

Figura 4.3: Marcador y ficha informativa de un proyecto agrivoltaico en el SIG desarrollado.

La herramienta utiliza un menú desplegable para seleccionar características en el mapa y permite acceder a todos los datos almacenados en una hoja de cálculo de Google, según lo requiera el usuario. También incluye un botón que redirige al usuario a un Google Forms para contribuir con información al mapa. Posteriormente, la información proporcionada en el Google Forms debe ser verificada y, si es apropiada, se agregará al mapa. En la figura 4.4 se muestra parte del SIG con todos los elementos mencionados anteriormente.

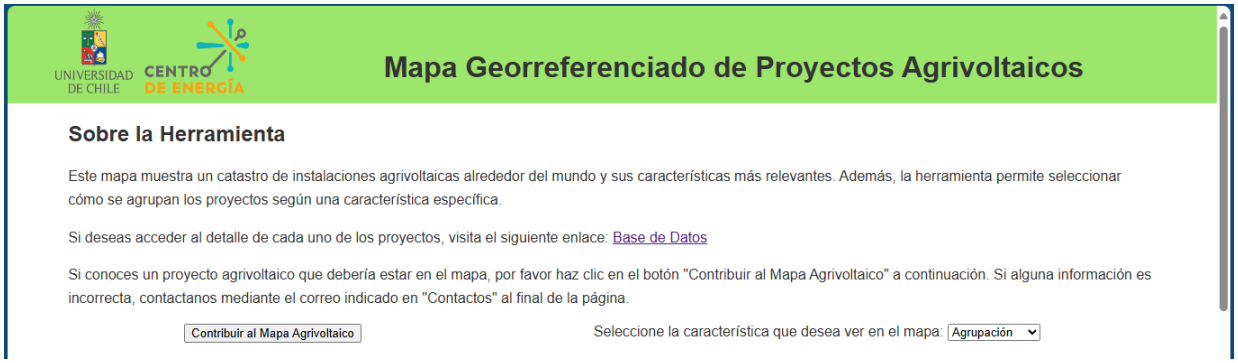

Figura 4.4: Parte de la página web con los elementos previamente mencionados.

En caso de que alguna información esté desactualizada o incorrecta, el usuario puede comunicarse a través de las vías específicas al final página web para informar del error.

Capítulo 5

Análisis exploratorio de datos

En este capítulo se abordan los resultados del análisis exploratorio del conjunto de datos construido en el capítulo anterior. Durante esta fase, se exploran y visualizan los datos para comprender su estructura, identificar patrones, tendencias y posibles relaciones entre variables. El código utilizado para el análisis exploratorio de los datos puede ser encontrado en el anexo D.

Durante la realización de un análisis exploratorio de datos, es común emplear técnicas como la estadística descriptiva, gráficos y visualizaciones, así como medidas de resumen para examinar la distribución y la variabilidad de los datos. Este enfoque ayuda a identificar valores atípicos, datos faltantes u otros problemas que puedan afectar la calidad de los resultados analíticos.

Una vez completado el análisis exploratorio de datos, se obtienen valiosas percepciones que pueden guiar el desarrollo de modelos analíticos más avanzados o estrategias de toma de decisiones. Estas percepciones pueden revelar relaciones entre variables, patrones ocultos en los datos o incluso nuevas preguntas de investigación.

5.1. Análisis univariado

El análisis univariado consiste en analizar cada variable de forma individual, utilizando diversas técnicas estadísticas y gráficas para explorar las características de una variable.

Cabe destacar que no todas las categorías del conjunto de datos construido serán analizadas, solo se examinarán aquellas que se utilizarán para el agrupamiento de los datos. El resto de las categorías servirán como características informativas para la realización del SIG. Las categorías a analizar son: capacidad, superficie, altura, altitud, clima, cultivo, diseño, panel y seguimiento. Entre ellas, capacidad, superficie, altura y altitud son variables numéricas, mientras que clima, cultivo, diseño, panel y seguimiento son variables categóricas.

5.1.1. Medidas resumen de las variables numéricas

El primer paso antes de cualquier estudio estadístico es acceder a la base de datos mediante SQLAlchemy y almacenar en un DataFrame de Pandas la consulta que retorne todas las categorías a analizar. Un DataFrame es una estructura de datos bidimensional se utiliza para almacenar y manipular datos tabulares.

En la tabla 5.1 se presenta un resumen estadístico de las columnas numéricas del DataFrame. La tabla muestra las principales estadísticas de los datos, incluyendo el conteo, la media, la desviación estándar, los valores mínimos y máximos, así como los percentiles.

	Capacidad [MW]	Superficie [ha]	Altura [m]	Altitud [m s. n. m.]
Conteo	578,000000	566,000000	121,000000	586,000000
Media	15,711068	49,494709	1,766612	319,494881
Desviación estándar	50,132941	175,235397	1,416412	290,615523
Valor mínimo	0,004800	0,002600	0,300000	-35,000000
25 %	1,400000	2,985000	0,710000	193,250000
50 %	2,2500000	4,670000	0,910000	289,000000
75 %	6,377500	12,172500	3,000000	343,750000
Valor máximo	550,000000	1902,060000	5,500000	1853,000000

Tabla 5.1: Resumen estadístico de las columnas numéricas del DataFrame.

A partir de este primer acercamiento, se pueden observar varias aspectos relevantes. En primer lugar, es necesario lidiar con los datos faltantes, ya que los algoritmos de agrupamiento no aceptan valores en blanco (nulos). Además, se puede observar que los valores máximos de las categorías capacidad, superficie y altitud se alejan significativamente del percentil 75 %, lo que sugiere la posible presencia de datos anómalos o fuera de norma (en inglés "outliers"). La importancia de detectar y manejar correctamente los datos anómalos radica en que puede mejorar el rendimiento y la precisión del modelo.

5.1.2. Histograma variables numéricas

Con el propósito de identificar la distribución de los datos, se elabora un histograma. Este paso resulta fundamental, dado que determinar si las variables numéricas siguen una distribución normal proporciona una valiosa información. Los algoritmos de agrupamiento tienden a desempeñarse mejor cuando las variables exhiben esta distribución. Sin embargo, en los casos en los que las variables no se ajustan a una distribución normal, se hace necesario aplicar transformaciones con el fin de aproximarlas lo más posible a dicha distribución. A continuación, en la figura 5.1 se presenta el histograma de las variables numéricas del DataFrame.

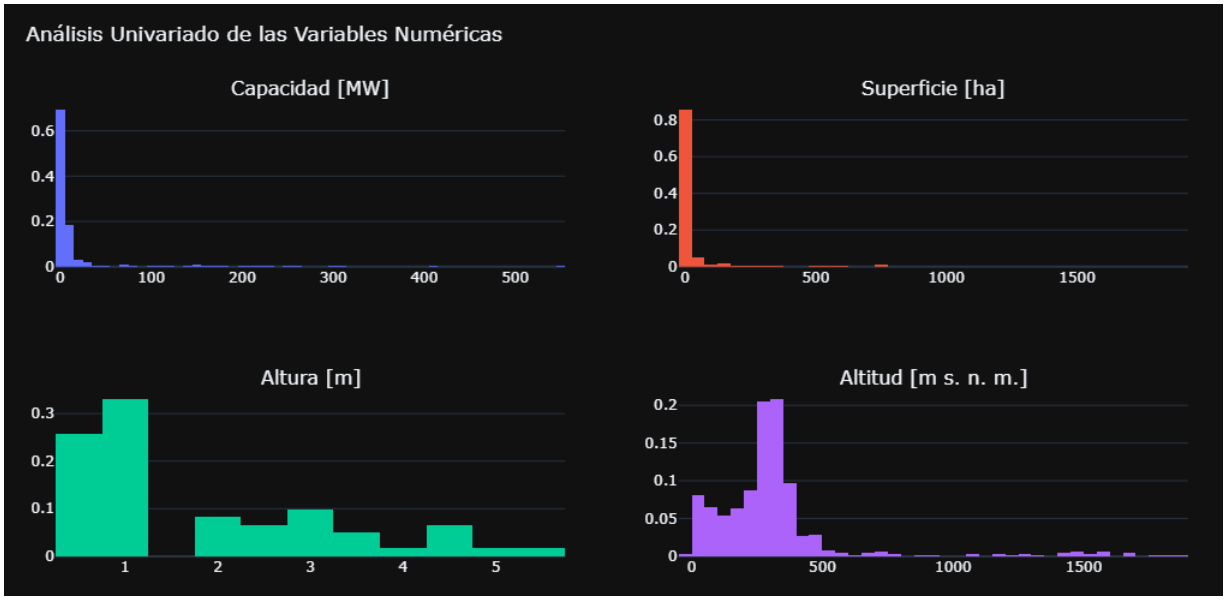

Figura 5.1: Histograma de las variables numéricas.

Cabe mencionar que los histogramas se encuentran normalizados, de manera que la suma de todas las barras sea igual a 1, lo que convierte al histograma en una representación de la función de densidad de probabilidad de la distribución de los datos.

A partir de los histogramas, se puede observar que gran parte de los datos se encuentran concentrados hacia el extremo izquierdo de la gráfica y que hay presencia de datos que se alejan considerablemente de la media. Este fenómeno era anticipado en los resultados mostrados en la tabla 5.1.

En relación con la potencia instalada de las instalaciones agrivoltaicas, se observa que la mayoría, aproximadamente el 70 %, tiene una capacidad inferior a 5 [MW]. Además, estas instalaciones suelen no ocupar extensiones de terreno mayores a 25 hectáreas, ya que el 85 % de los proyectos se encuentra por debajo de ese umbral. De igual forma, alrededor de la mitad de los proyectos que informaron la altura del panel respecto al suelo, presenta una altura inferior a 1,25 metros. Por último, la gran mayoría de los proyectos se encuentra a altitudes inferiores a 500 [m s. n. m.].

De igual forma para caracterizar las distribuciones, resultan útiles medidas como el skewness y la curtosis. El skewness, también conocido como asimetría, proporciona una indicación de la falta de simetría en la forma de la distribución. En esencia, nos revela si la distribución está desplazada hacia la derecha o hacia la izquierda con respecto a su valor central. Un skewness de cero implica que la mitad de los datos están por encima del valor central y la otra mitad por debajo, lo que denota una distribución equilibrada. Si el skewness es positivo, la distribución muestra una inclinación hacia la izquierda, mientras que un skewness negativo indica una inclinación hacia la derecha [35].

Por otro lado, la curtosis indica el grado de concentración de los datos en torno a la media. Una distribución con una alta curtosis presenta una concentración mayor de datos alrededor de la media y una cola más pronunciada, lo que sugiere la presencia de valores extremos

en la distribución. Por el contrario, una baja curtosis refleja una concentración menor de datos alrededor de la media y una cola menos pronunciada, indicando una menor presencia de valores extremos en la distribución [35].

A continuación, en la tabla 5.2 se presenta el skewness y curtosis de las variables numéricas.

Variables	Skewness	Curtosis
Capacidad	5.45	37.41
Superficie	6.14	48.07
Altura	0.98	-0.25
Altitud	3.06	10.85

Tabla 5.2: Skewness y curtosis de las variables numéricas del DataFrame.

En la práctica, se suelen utilizar valores bajos de skewness y curtosis para indicar que una distribución se asemeja a una distribución normal. Sin embargo, en este caso, no se cumple esta condición, por lo que será necesario aplicar transformaciones paramétricas que buscan aproximar una distribución arbitraria a una distribución normal.

5.1.3. Histograma variables categóricas

Los histogramas son una herramienta poderosa para visualizar la distribución de datos, especialmente en el caso de variables numéricas. Sin embargo, también resultan útiles para identificar la frecuencia de distintas categorías en un conjunto de datos y detectar posibles desequilibrios en la distribución de estas categorías.

Vale la pena señalar que los gráficos que se muestran a continuación representan la frecuencia de aparición de las distintas categorías, a diferencia de los histogramas de las variables numéricas que estaban normalizados.

Seguidamente, en la figura 5.2 se presenta el histograma de la variables categóricas clima, panel, seguimiento y diseño.

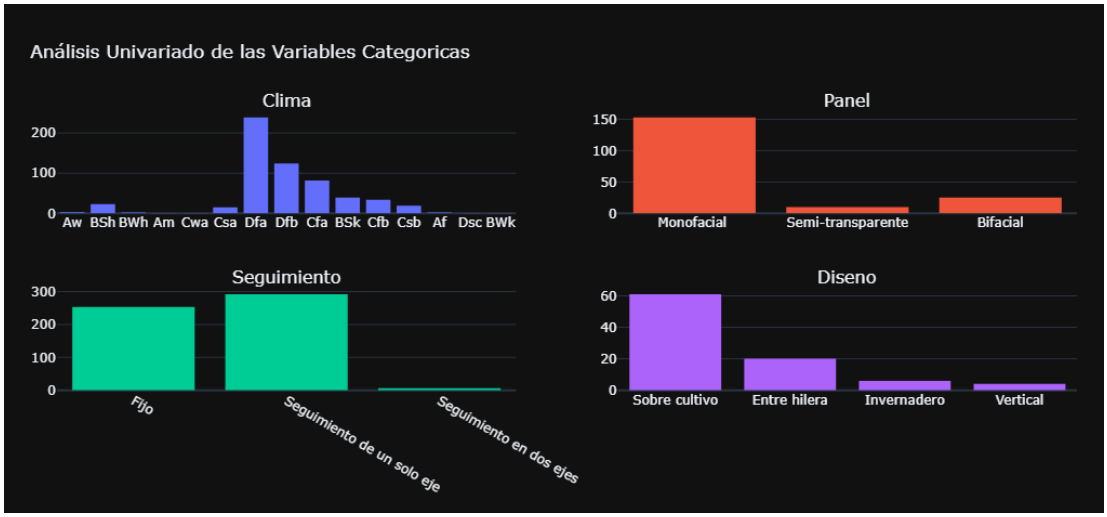

Figura 5.2: Histograma de las variables categóricas clima, panel, seguimiento y diseño.

A partir de estos histogramas, es posible notar que la mayoría de los proyectos agrivoltaicos analizados presentan climas del tipo Dfa, Dfb y Cfa principalmente. Estos climas se caracterizan por ser continentales (Dfa y Dfb) y templados (Cfa), sin estación seca y con veranos cálidos y templados. Además, se puede notar que, las dos categorías mas usada con respecto al seguimiento son: arreglo fijo y seguimiento en un solo eje.

Aunque en el diseño de las plantas agrivoltaicas se prefiere la disposición sobre los cultivos y el uso de paneles monofaciales, la mayoría de los proyectos no incluyen información sobre el diseño ni el tipo de panel. Esto se refleja en que la suma de la frecuencia de las categorías es considerablemente menor que el total de proyectos censados en ambos casos.

Por último, en la figura 5.3 se muestra el histograma de la variable categórica cultivo. Se observa que el principal cultivo utilizado en las plantas agrivoltaicas son las praderas destinadas a la alimentación del ganado, seguido por el cultivo de hortalizas.

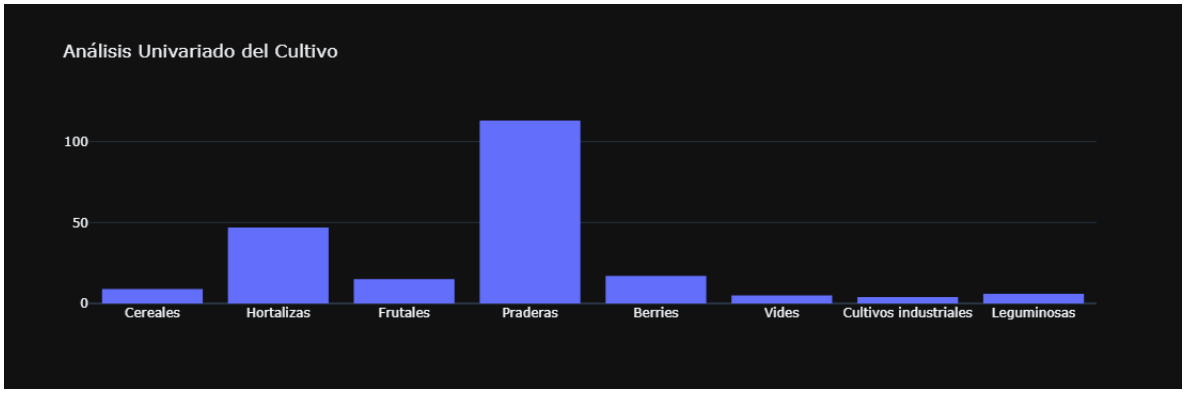

Figura 5.3: Histograma de las variable categórica cultivo.

5.2. Análisis multivariado

El análisis multivariado consiste en analizar múltiples variables simultáneamente, utilizando diversas técnicas estadísticas para entender las relaciones y patrones entre ellas. Este tipo de análisis permite explorar cómo las variables interactúan entre sí.

5.2.1. Correlación entre variables

El análisis de correlaciones es una técnica que permite identificar patrones y relaciones entre las distintas variables numéricas de un conjunto de datos. La correlación se puede medir utilizando diferentes métodos, uno de ellos es el coeficiente de correlación de Pearson, el cual es un índice que se utiliza para medir el grado de relación entre dos variables, siempre y cuando ambas sean cuantitativas y continuas. El valor del índice de correlación varía en el intervalo $[-1, 1]$, donde:

- Un valor cercano a 1 indica una fuerte correlación positiva, lo que significa que a medida que una variable aumenta, la otra también tiende a aumentar.
- Un valor cercano a -1 indica una fuerte correlación negativa, lo que significa que a medida que una variable aumenta, la otra tiende a disminuir.

- Un valor de cercano a 0 sugiere que no existe una correlación lineal entre las variables.

En la figura 5.4, se presenta un mapa de calor de la correlación entre las variables numéricas: capacidad, superficie, altura y altitud.

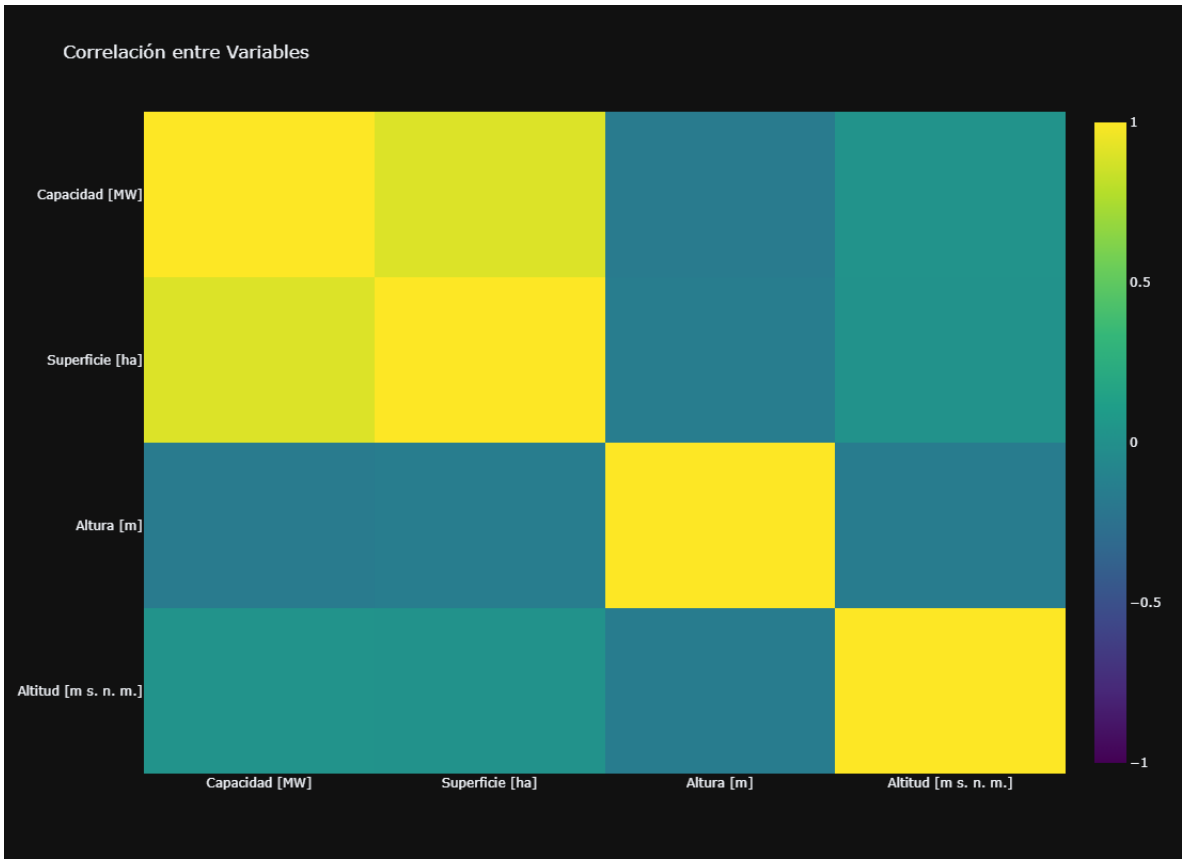

Figura 5.4: Correlación entre pares de variables numéricas.

En la tabla 5.3 se presenta el valor exacto del coeficiente de correlación de Pearson para las variables numéricas.

Variables	Capacidad	Superficie	Altura	Altitud
Capacidad	1	0,8966462	-0,1684083	0,02510324
Superficie	0,8966462	1	-0,1558334	0,01552788
Altura	-0,1684083	-0,1558334	1	-0,1598411
Altitud	0,02510324	0,01552788	-0,1598411	1

Tabla 5.3: Coeficiente de Pearson para las variables numéricas.

A partir de los resultados mostrados en la figura 5.4, es posible observar que las variables Capacidad y Superficie comparten el color amarillo, lo cual indica una correlación cercana a 1. En la tabla 5.3, se corrobora al observar que el valor de la correlación es de aproximadamente 0,90. Esto sugiere una fuerte correlación entre Capacidad y Superficie, por lo que una de estas variables podría ser desechada para la construcción del modelo. No descartar una de las variables introduciría ruido en el modelo.

Sin embargo, una alta correlación no implica causalidad, por lo que es importante entender bien cómo interactúan estas variables entre sí antes de eliminar alguna.

Dado que un aumento en la capacidad requiere la instalación de más paneles, lo que a su vez implica un mayor uso de superficie, es lógico que estas variables estén altamente relacionadas.

Por otro lado, el resto de los valores de correlación entre las variables es bajo, por lo tanto, no aportan ruido al modelo y no es necesario descartarlas.

5.2.2. Tablas de contingencia

Para el análisis de valores categóricos (categórico vs categórico) existen herramientas especializadas. Una de ellas es el uso de tablas de contingencia, también conocidas como tablas de dos entradas. Estas permiten calcular el número de ocurrencias de una variable para cada una de sus categorías en comparación con los valores de otra variable. Sin embargo, estos resultados deben tomarse con cautela, ya que no toda la base de datos se encuentra completamente caracterizada. De hecho, la mayor cantidad de datos faltantes se encuentra en las variables categóricas (ver sección 5.3), lo que podría dificultar la interpretación de los datos. En el anexo B se pueden revisar cada una de las tablas de contingencia.

Por otro lado, una existen mas alternativas para evaluar múltiples valores categóricos como la prueba de chi-cuadrado, que es una prueba estadística utilizada para determinar si existe una asociación significativa entre dos variables categóricas. Para el cálculo, es necesario utilizar las tablas de contingencia. A partir de estas tablas se calculan las frecuencias esperadas y el estadístico chi-cuadrado. La cantidad de filas y columnas de la tabla se usa para determinar los grados de libertad. Con el valor del estadístico chi-cuadrado calculado y los grados de libertad, se encuentra el valor de probabilidad en la tabla de distribución de chi-cuadrado. Este valor de probabilidad es una medida que se utiliza en estadística para determinar la significancia de los resultados de un análisis [36].

Si el valor de probabilidad es menor que el nivel de significancia (comúnmente 0,05), se rechaza la hipótesis nula, indicando que hay una asociación significativa entre las variables. La hipótesis nula establece que no hay asociación entre las variables, es decir, las variables son independientes. Si el valor de probabilidad es mayor o igual al nivel de significancia, no se rechaza la hipótesis nula, lo que sugiere que no hay evidencia suficiente para afirmar que existe una asociación entre las variables.

Sin embargo, se opta por no realizar esta prueba debido a que el tamaño de la muestra observado es pequeño. En este caso, es posible que la prueba de chi-cuadrado no tenga suficiente potencia para detectar una asociación real, incluso si esta existe. Esto se puede comprobar al observar que en las tablas no se cumple la suposición de que las frecuencias esperadas en cada categoría deben ser lo suficientemente altas, normalmente al menos 5 [37].

5.3. Identificación de valores faltantes

Los valores faltantes corresponden a datos que faltan en una o varias observaciones o variables del conjunto de datos. El primer paso es determinar la cantidad y la ubicación de estos valores.

Para determinar la cantidad, se cuenta el número de valores nulos o faltantes por categoría. En la tabla 5.4 se muestran los resultados de este conteo.

Variables	Valores Faltantes	Porcentaje de Datos Faltantes [%]
Capacidad	8	1,37
Superficie	20	3,41
Altura	465	79,35
Altitud	0	0,00
Clima	0	0,00
Cultivo	389	66,38
Panel	398	67,91
Seguimiento	35	5,97
Diseño	495	84,47

Tabla 5.4: Valores faltantes por cada una de las variables.

Para visualizar la ubicación y dispersión de los datos faltantes, se grafica en la figura 5.5 la matriz de nulidad de los datos, lo cual permite identificar los datos faltantes a través de barras. Cada espacio ausente en las barras indica la falta de un dato en esa posición específica.

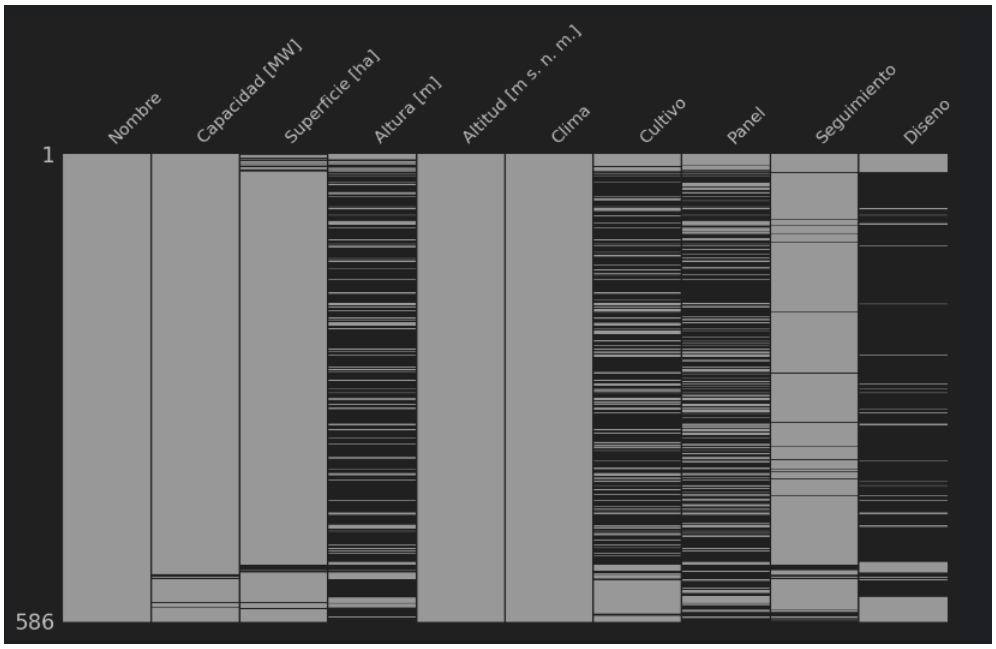

Figura 5.5: Ubicación de los datos faltantes en el DataFrame.

En la imagen 5.6 se muestran las correlaciones de nulidad entre pares de variables. Este tipo de gráfico solo presenta las variables que tienen datos faltantes. Las correlaciones varían

desde -1 hasta 1, donde -1 significa que las variables son excluyentes, es decir, la aparición de una hace que la otra esté ausente. Un valor de 1 corresponde a inclusión, lo que significa que la aparición de una hace que la otra también aparezca. Los valores cercanos a 0 (sin valor numérico en el gráfico) indican ausencia de relación de nulidad entre las variables.

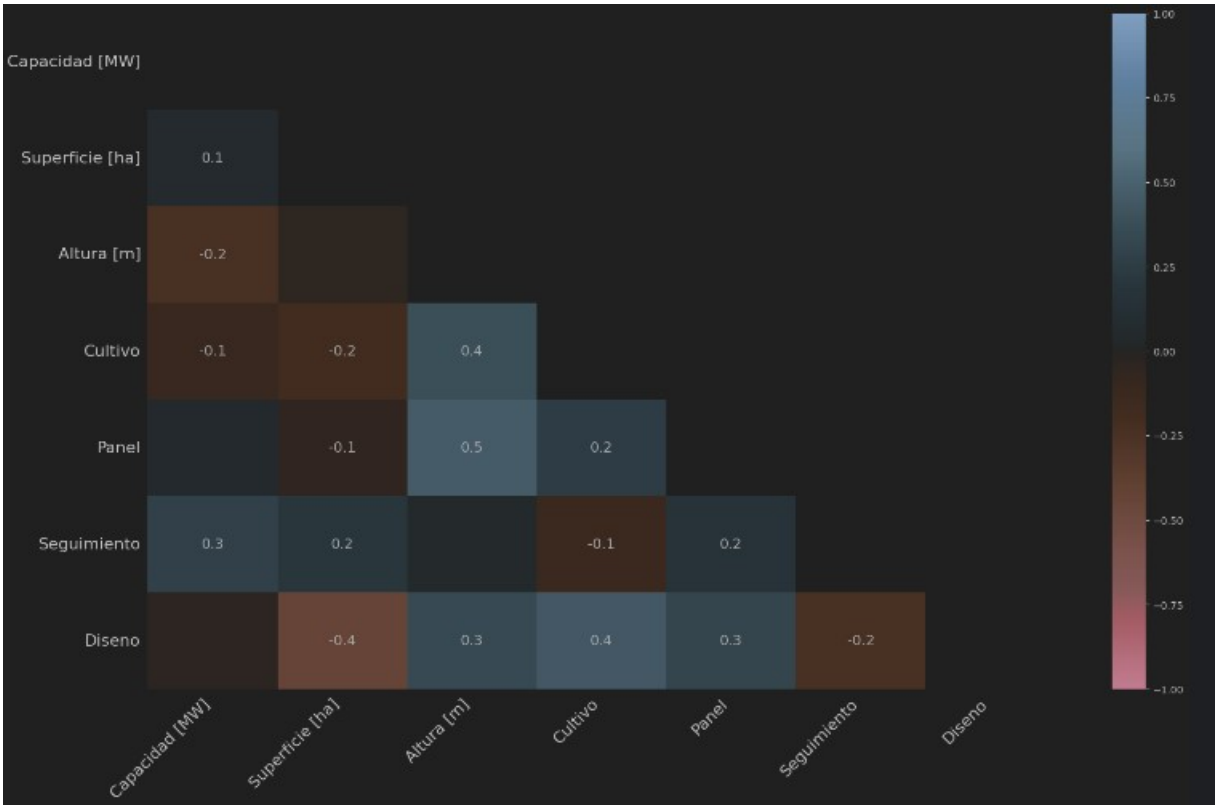

Figura 5.6: Correlaciones de nulidad entre pares de variables.

De los resultados obtenidos en la tabla 5.4, se puede observar que el porcentaje de datos faltantes en las categorías de altura, cultivo, panel y diseño es alto, especialmente en la categoría de diseño. En la figura 5.5, se observa que los datos faltantes de una variable en particular están esparcidos a lo largo de toda su columna. Por otro lado, en la figura 5.6, se puede ver que la pérdida de datos no puede ser explicada por ninguna otra característica del conjunto de datos.

La pérdida de datos puede deberse a que los sistemas de información geográfica (SIG) agrivoltaicos, al caracterizar los proyectos agrivoltaicos, omiten responder algunas preguntas. Un ejemplo de esto son los SIG inSPIRE Agrivoltaics Maps y el Mapa de Instalaciones Agrivoltaicas en Alemania, que son de naturaleza comunitaria. Asimismo, mucha información de los proyectos provienen de noticias y páginas web de empresas del sector agrivoltaico, las cuales pueden no tener un interés particular en proporcionar todas las características relevantes. Considerando lo anteriormente expuesto, los datos faltantes son completamente al azar.

Ahora bien, para tratar estos datos, lo más común es eliminar columnas o imputar datos. En el caso de la variable "diseño", en la que faltan muchos datos y que no es del todo informativa, dado que optar por colocar los paneles sobre el cultivo o entre las hileras repercute

en que los paneles se encuentren a una distinta altura respecto al suelo, algo que proporciona información similar a la que entrega la variable altura, es mejor eliminar esa variable del conjunto de datos.

Con respecto a las otras variables con una gran cantidad de datos faltantes que son más relevantes para el estudio, como altura, cultivo y panel, es mejor imputar los datos. Para el resto de las variables no mencionadas que también presentan datos faltantes, lo más adecuado es realizar una imputación de datos.

5.4. Preprocesamiento de datos

El preprocesamiento de datos es una etapa crucial en cualquier proyecto de análisis de datos o aprendizaje automático. En esta fase, se limpia y mejora la calidad de los datos para prepararlos adecuadamente para el modelo de aprendizaje automático. También se realizan transformaciones para asegurar que los valores de las columnas estén en escalas similares y tengan distribuciones relativamente cercanas a la distribución normal.

Para llevar a cabo el preprocesamiento, es fundamental contar con la información obtenida del análisis univariado, análisis multivariado e identificación de valores faltantes.

El primer paso es eliminar características irrelevantes o redundantes. Para este caso de estudio, la superficie se elimina debido a su alta correlación con la variable capacidad, y el diseño se descarta porque contiene muchos valores faltantes, lo que podría inducir un sesgo si se optara por imputar datos en esta variable, aparte de no ser tan relevante.

¿Qué sucede con la variable altura, una variable que también presenta numerosos datos faltantes? Antes de emplear algún método de imputación ofrecido por la librería scikit-learn, se debe considerar la norma alemana DIN SPEC 91434 [14] sobre instalaciones agrovoltaicas. Esta norma establece que los proyectos con diseño del tipo "sobre el cultivo" se consideran de este tipo siempre que los paneles se encuentren a más de $2,1$ metros sobre el suelo. En caso contrario, se consideran del tipo "entre hileras". Con base en esto, se imputa la categoría "sobre cultivos" a todos los proyectos agrivoltaicos con una altura superior a $2,1$ metros que no contaban con información sobre el diseño, y se asigna la categoría "entre hileras" a los proyectos con alturas menores a $2,1$ metros.

Aunque esto pueda parecer innecesario dado que la variable diseño no se utiliza en la construcción del modelo, se observa que varios proyectos informan su tipo de diseño pero no proporcionan la altura respecto al nivel del suelo. Por lo tanto, se procede a imputar las alturas de todos los proyectos con diseño del tipo "sobre cultivos" pero sin altura conocida, como el promedio de las alturas conocidas de proyectos con el mismo tipo de diseño (3,28 [m]). Este procedimiento se replica de igual manera para todos los proyectos con diseño del tipo "entre hileras", como el promedio de las alturas conocidas de proyectos con ese mismo tipo de diseño (0,78 [m]).

Este procedimiento se lleva a cabo con el objetivo de extraer la mayor cantidad de información posible de la variable diseño antes de su eliminación, ya que para las imputaciones posteriores, que se realizan con modelos más complejos, esta información ya no está disponi-

ble. Con esta estrategia, se redujo el número de datos faltantes en la variable altura de 465 a 429. Cabe destacar que este proceso solo modifica el DataFrame de pandas y no afecta la información almacenada en la base de datos ni en el SIG agrivoltaico.

Las variables, según su tipo—numéricas o categóricas—requieren tratamientos distintos. En el caso de las variables numéricas como capacidad, superficie y altura, el primer paso es escalarlas y normalizarlas. Dado que estas variables no siguen una distribución normal y contienen datos atípicos, se emplea "RobustScaler" de la librería scikit-learn. A diferencia de "StandardScaler", que normaliza los datos utilizando la media y la desviación estándar, RobustScaler emplea la mediana y el rango intercuartílico para manejar los datos atípicos de manera más eficaz.

A continuación, es necesario aproximar las distribuciones arbitrarias a una distribución gaussiana. Para lograr esto, se utiliza una transformación como la Yeo-Johnson, que se accede a través de la clase "PowerTransformer" de scikit-learn.

Una vez que los datos han sido escalados y normalizados, se puede proceder con la imputación de los valores faltantes. Para esto, se utiliza el método "KNNImputer" de scikit-learn. Este método emplea la técnica de los vecinos más cercanos para estimar los valores faltantes. Específicamente, el KNNImputer identifica los K vecinos más cercanos que tienen valores conocidos y utiliza una función de agregación, como la media, para imputar los valores ausentes. La metodología se basa en la premisa de que los valores faltantes pueden ser inferidos con precisión a partir de las observaciones más similares, proporcionando así una forma efectiva de manejar datos incompletos en el análisis y modelado. Para este proyecto, se consideraron los 5 vecinos más cercanos para la imputación.

Con los valores imputados, se puede considerar completado el preprocesamiento de los datos numéricos, los cuales ahora tienen la distribución mostrada en la figura 5.7.

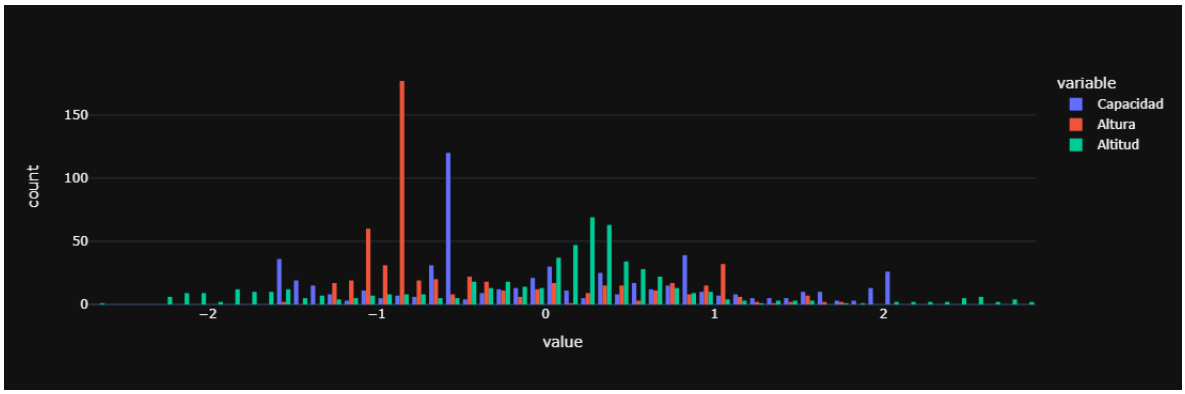

Figura 5.7: Histograma de las variables capacidad, superficie y altura ya preprocesadas.

Para el caso de las variables categóricas, como clima, cultivo, panel y seguimiento, lo primero es codificar los datos, ya que la mayoría de los algoritmos de aprendizaje automático trabajan con números y no con datos categóricos. Para esto, se utiliza "LabelEncoder", una clase de la biblioteca scikit-learn que convierte etiquetas de categorías en números enteros.

Una vez codificados los datos, se procede con la imputación de los valores faltantes. En

este caso, se utiliza el método "IterativeImputer" de scikit-learn. Este método funciona de manera iterativa, modelando cada característica con valores faltantes como una función de otras características y utilizando esas estimaciones para rellenar los valores faltantes. Las estimaciones se realizan con un modelo clasificador de bosque aleatorio ("RandomForestClassifier") de scikit-learn [38].

Un bosque aleatorio es un conjunto de árboles de decisión, donde cada árbol contribuye a la decisión final del clasificador. Un árbol de decisión es un modelo predictivo utilizado tanto para tareas de clasificación como de regresión. Se estructura como un árbol en el que cada nodo interno representa una pregunta o condición sobre una característica del conjunto de datos, cada rama representa el resultado de esa condición, y cada hoja (nodo terminal) representa una etiqueta de clase (en clasificación) o un valor continuo (en regresión).

Finalmente, una vez terminado el preprocesamiento, se obtiene un conjunto de datos limpio y completamente caracterizado, listo para ser agrupados. Esto significa que la nueva base de datos no presenta valores faltantes, ya que estos han sido reemplazados por estimaciones. Es importante resaltar nuevamente que la base de datos pos-imputación ya no incluye las columnas de superficie y diseño.

La base de datos pre-imputación y pos-imputación presentan grandes diferencias. En el caso de las variables numéricas pos-imputación, mostradas en la figura 5.7, ahora se concentran aproximadamente entre -3 y 3, asemejándose a una distribución normal, lo cual contrasta con las variables numéricas antes de la imputación, como se observa en la figura 5.1.

Para las variables categóricas pos-imputación, se presenta el histograma de su decodificación en la figura 5.8. Al comparar estas distribuciones con las pre-imputación, mostradas en la figura 5.2, se observa que las variables relacionadas con el tipo de panel, seguimiento y tipo de cultivo tienden a exacerbar la diferencia entre las categorías más comunes y las menos frecuentes. Esto podría ser indicativo de un sesgo debido a la disparidad inicial entre las categorías. Por otro lado, la variable categórica clima no muestra cambios, ya que inicialmente no presentaba datos faltantes.

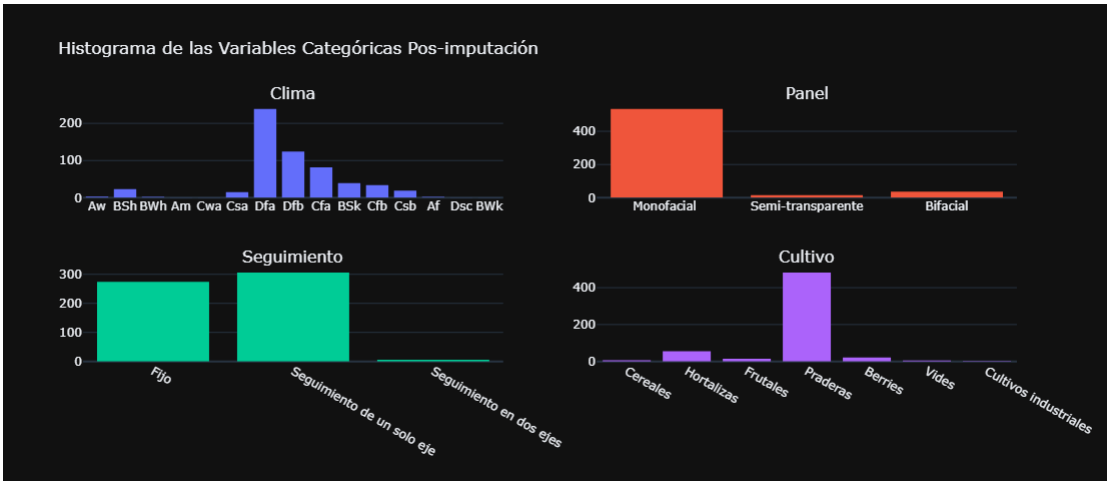

Figura 5.8: Histograma de las variables categóricas pos-imputación.

Capítulo 6

Análisis de identificación de agrupamientos

En este capítulo se aborda cómo se construye el modelo de agrupamiento, incluyendo la selección del número de grupos, las características compartidas por los grupos formados. Cabe destacar que para este caso de estudio se utilizaron dos métodos de agrupamiento de datos: k-prototypes y aglomerativo.

K-prototypes es una técnica con un enfoque particional, es decir, divide los datos en grupos sin solapamientos, de manera que cada dato pertenece a un solo grupo y no a otro. Por otro lado, el agrupamiento aglomerativo tiene un enfoque jerárquico, en el que los datos se agrupan estableciendo jerarquías entre ellos, organizando los datos en forma de un árbol.

Finalmente, se validan los modelos y se analiza cuál de estos métodos presenta un mejor desempeño. El código utilizado para el análisis de identificación de agrupamiento puede ser encontrado en el anexo D.

6.1. Selección de número adecuado de grupos

La selección del número adecuado de grupos es una fase crucial en el análisis de agrupamiento, un proceso común en el aprendizaje automático y la minería de datos. Determinar el número óptimo de grupos en un conjunto de datos puede afectar significativamente la calidad y la utilidad de los resultados del análisis.

Para la selección, se utilizan índices de evaluación de calidad de grupos, como se presenta en la sección 2.4.6. Sin embargo, primero es necesario explicar los parámetros que recibe el método k-prototypes de la biblioteca k-modes y el algoritmo aglomerativo de la biblioteca scipy.

6.1.1. K-prototypes

Este método requiere el número de grupos, el cual debe determinarse. También se necesita un método de inicialización, para este caso, se eligió el método Huang propuesto en [24]. Ade-

más, se debe establecer un parámetro para controlar la reproducibilidad de los resultados que involucran elementos aleatorios, esto asegura que la inicialización aleatoria de los centroides sea la misma en cada ejecución, lo que facilita la comparación de resultados y la depuración.

Dado que el número de grupos es lo que se está buscando, se debe iterar a través de un rango de grupos. Para este trabajo, se elige un rango de 2 a 20 grupos.

Una vez definidos estos parámetros, se entrena el modelo proporcionando el conjunto de datos preprocesado y especificando qué columnas son categóricas. El modelo asigna a cada proyecto una etiqueta de algún grupo formado, y estos resultados se almacenan para luego calcular la puntuación de la silueta en función del número de grupos. Para el cálculo, se utiliza el método "silhouette_score" de scikit-learn, que recibe como parámetros de entrada una matriz de distancias de los datos, precalculada mediante el método "gower_matrix" de la biblioteca Gower, junto con las etiquetas de los grupos. Dado que las distancias están precalculadas, también deben especificarse como parámetros de entrada para el cálculo de la puntuación de silueta. A continuación, en la figura 6.1 se muestra la puntuación de silueta para k-prototypes.

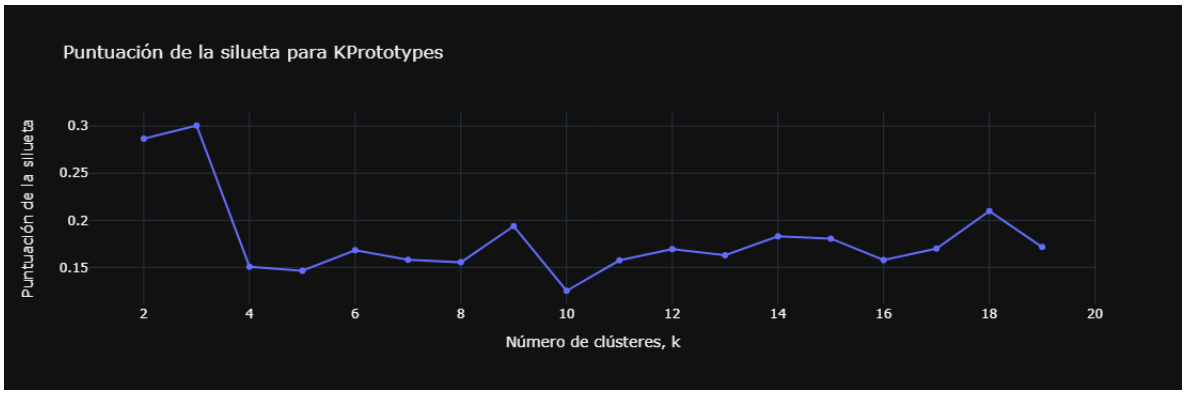

Figura 6.1: Puntuación de silueta según la cantidad k de grupos para el algoritmo k-prototypes.

A partir de esta figura, se observa que el pico se da cuando se tienen 3 grupos, y que aumentar la cantidad de grupos solo empeora el resultado. Por ahora, solo interesa identificar en qué puntos se producen el máximo, una análisis más detallado sobre la puntuación se aborda en la discusión sobre la validación del modelo. Cabe recordar que los índices de evaluación de calidad de grupo son útiles tanto para la selección del número de grupos como para la validación del modelo.

Para calcular el índice de Davies-Bouldin, se utiliza el método "davies_bouldin_score" de scikit-learn, el cual recibe como parámetros de entrada el DataFrame preprocesado y las etiquetas de los grupos. En la figura 6.2 se presenta el índice de Davies-Bouldin en función del número de grupos para el algoritmo k-prototypes.

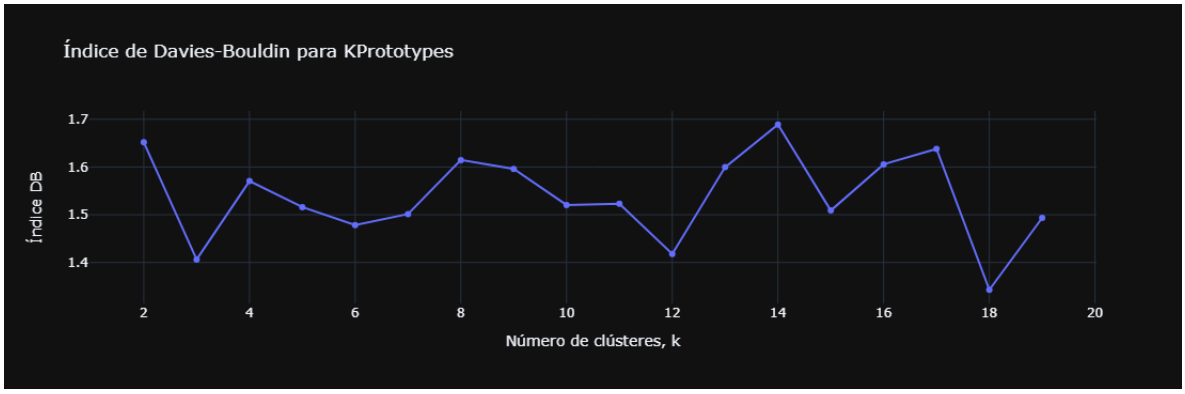

Figura 6.2: Índice de Davies-Bouldin según la cantidad k de grupos para el algoritmo k-prototypes.

A diferencia de la puntuación de silueta, que busca maximizar la puntuación, el índice de Davies-Bouldin busca minimizar el índice. Al observar la figura 6.2, se puede notar que el mínimo se produce para un k = 18.

El puntaje de silueta y el índice Davies-Bouldin sugieren diferentes números de grupos k. Sin embargo, es importante considerar que, aunque el índice Davies-Bouldin no tiene su valor mínimo cuando k = 3, el valor del índice es cercano al valor de k = 18. Por lo tanto, lo más adecuado sería seleccionar 3 grupos para la agrupación utilizando el algoritmo k-prototypes. Sin embargo, estos resultados son obtenidos con la semilla de generación de números aleatorios 42, por lo que es necesario comprobar la consistencia del resultado con otras semillas.

En este sentido, se replica el proceso de evaluación del número de grupos utilizando las semillas aleatorias 6 y 19. En ambos casos, la puntuación de silueta y el índice Davies-Bouldin indican que la cantidad óptima de grupos es 3, lo que confirma que la elección de k = 3 es acertada. Los gráficos con los índices de evaluación de calidad de grupo para ambas semillas pueden ser observados en el anexo C.

6.1.2. Agrupamiento jerárquico

Para realizar el agrupamiento jerárquico, se debe utilizar el método "linkage" de SciPy, el cual recibe como parámetros una matriz con las distancias de Gower y el tipo de enlace que se va a usar (ver sección 2.4.5).

Luego, se utiliza la función "fcluster" de la biblioteca SciPy para formar grupos a partir del resultado del agrupamiento jerárquico generado por la función linkage. Esta función permite cortar el dendrograma en un nivel específico para obtener un cierto número de grupos o definir grupos basándose en otros criterios. Los parámetros de entrada de la función son el resultado del agrupamiento jerárquico, un umbral cuyo significado depende del criterio elegido, y el criterio, que en este caso es el número máximo de grupos deseados. La función formará grupos hasta llegar al umbral.

La función fcluster retorna las etiquetas de los grupos. Una vez obtenidas estas etiquetas, el proceso de análisis de índices de evaluación de calidad de grupo es el mismo que el realizado con k-prototypes.

Se analizan los resultados para dos tipos de enlaces: completo y promedio, dado que son poco susceptibles a datos atípicos y ruidos.

Enlace completo

En la figura 6.3 se muestra la puntuación de silueta para el agrupamiento jerárquico de enlace completo, de se observa que la mejor puntuación de silueta se alcanza entre mayor sea la cantidad de grupos, alcanzando un máximo con k = 17.

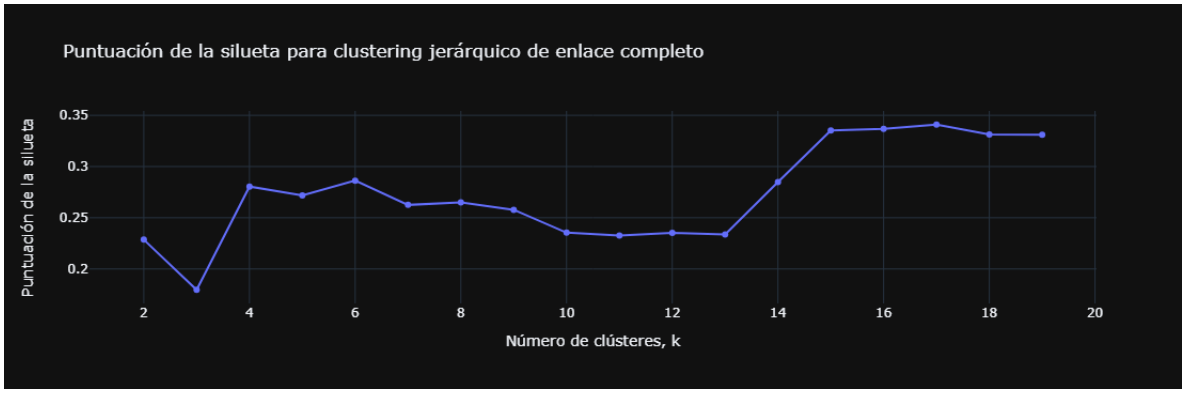

Figura 6.3: Puntuación de silueta según la cantidad k de grupos para el algoritmo de agrupación aglomerativa de enlace completo.

En la figura 6.4 se presenta el índice de Davies-Bouldin en función del número de grupos para el agrupamiento jerárquico de enlace completo, donde se puede observar que a medida que aumenta la cantidad de grupos disminuye el valor índice alcanzando el mínimo en k = 13.

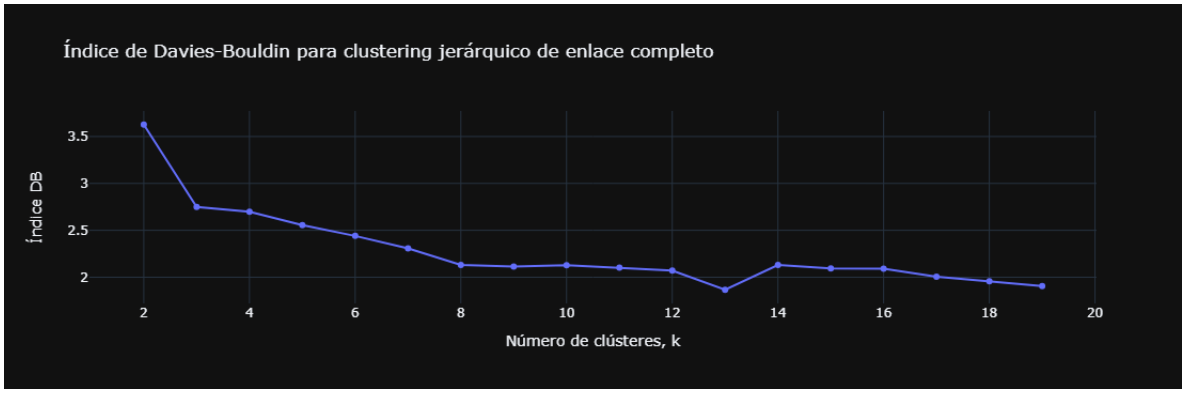

Figura 6.4: Índice de Davies-Bouldin según la cantidad k de grupos para el algoritmo de agrupación aglomerativa de enlace completo.

Se puede observar en ambas figuras que, a medida que aumenta el número de grupos, mejora la calidad de estos. Al escoger k = 17 se obtiene el mejor resultado en cuanto a puntuación de silueta, y el índice de Davies-Bouldin solo empeora ligeramente respecto al resultado obtenido con k = 13, , por lo que se elegirán 17 grupos. Sin embargo, esto puede ser un problema porque, si se tienen demasiados grupos en comparación con la cantidad de datos y como estos se agrupan, el modelo puede que esté sobreajustado, capturando ruido y detalles irrelevantes en lugar de las estructuras subyacentes verdaderas de los datos.

Enlace promedio

En la figura 6.5 se muestra la puntuación de silueta para el agrupamiento jerárquico de enlace promedio, donde se puede observar que el máximo se da en k = 4 y seguido de k = 2.

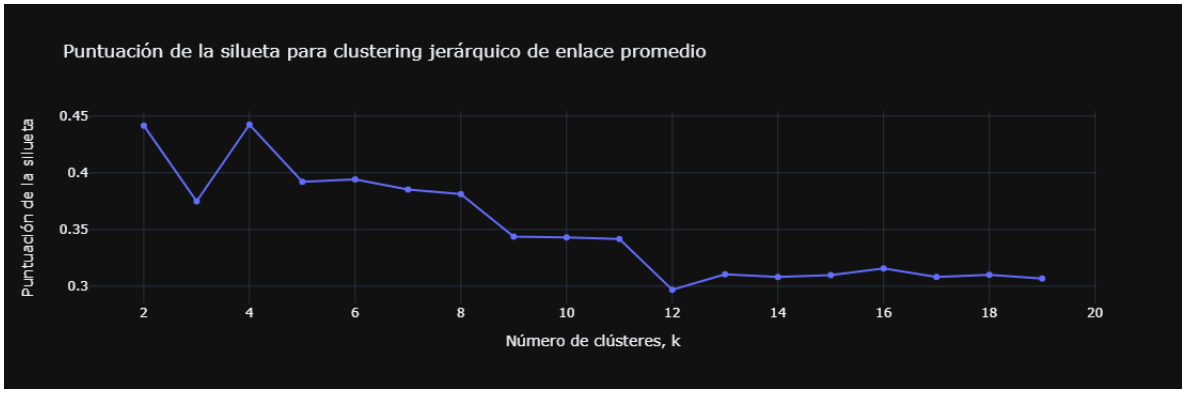

Figura 6.5: Puntuación de silueta según la cantidad k de grupos para el algoritmo de agrupación aglomerativa de enlace promedio.

En la figura 6.6 se presenta el índice de Davies-Bouldin en función del número de grupos para el agrupamiento jerárquico de enlace completo, donde a medida que aumenta la cantidad de grupos disminuye el valor índice.

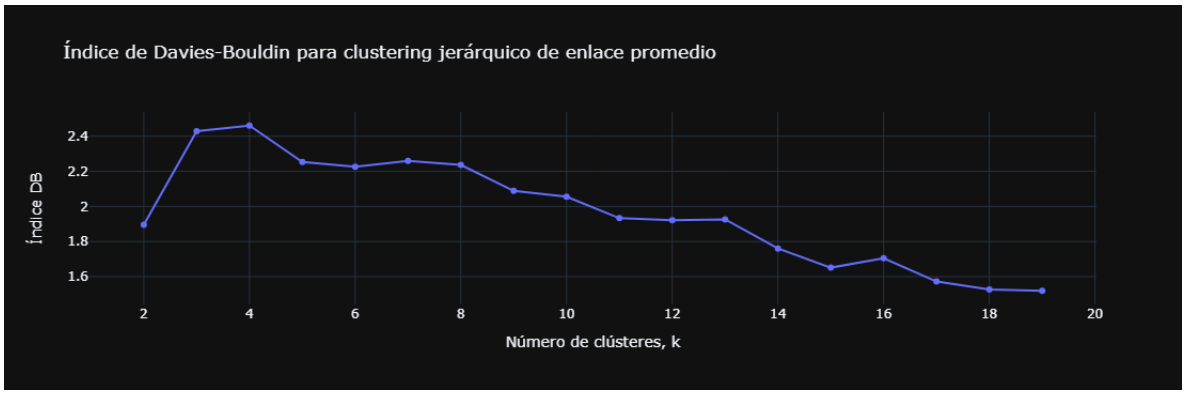

Figura 6.6: Índice de Davies-Bouldin según la cantidad k de grupos para el algoritmo de agrupación aglomerativa de enlace promedio.

Es importante recordar que el coeficiente de silueta se enfoca en la calidad interna de los grupos (cohesión y separación), indicando que con k = 4 los objetos están mejor emparejados dentro de su propio grupo y mal emparejados con los grupos vecinos.

Por otro lado, el índice de Davies-Bouldin con k = 20 sugiere que la calidad de los grupos es mejor, considerando la separación y compacidad. Dado que los índices de evaluación sugieren diferentes cantidades adecuadas de grupos, será necesario considerar otros factores, como el contexto del problema y el conocimiento del dominio.

En este caso, 20 grupos pueden parecer excesivos, dado que la mayoría de los datos se concentra en pocas categorías en las variables categóricas, y las variables numéricas también

parecen estar bien concentradas, como se observa en la sección 5.1. Por lo tanto, sería esperable que se formara una menor cantidad de grupos. En este contexto, la mejor elección de número de grupos es 2 o 4 en lugar de 20. Dado que la puntuación de silueta para k = 2 es similar a la de k = 4, pero su índice Davies-Bouldin es mucho menor, la mejor elección es k = 2.

6.2. Validación del modelo

Tal como se menciona en la sección anterior, los índices de evaluación de calidad de grupo no solo sirven como método para seleccionar el número adecuado de grupos, sino también como uno de los métodos para la validación de los modelos de agrupación. Una vez seleccionada la cantidad adecuada de grupos, se deben comparar los valores obtenidos para estos índices con el fin de escoger el modelo con mejor desempeño.

A continuación, en la tabla 6.1, se pueden observar los valores obtenidos para la puntuación de silueta y el índice Davies-Bouldin cuando se elige la cantidad adecuada de grupos según el algoritmo de agrupación. Cabe mencionar que los valores mostrados para el algoritmo k-prototypes son validos para la semilla de aleatoriedad 6. De igual manera, se observó que los resultados para otras semillas no varían de forma abrupta y se mantienen dentro de un margen pequeño.

Algoritmos de agrupamiento	Cantidad de grupos	Puntuación de silueta	Índice de Davies-Bouldin
K-prototypes	3	0,30320	1,37913
Jerárquico con enlace completo	17	0,34975	2,00717
Jerárquico con enlace promedio	2	0,044165	1,89557

Tabla 6.1: Valores de los índices de evaluación de calidad de grupos cuando se escoge la cantidad adecuada de grupos.

A partir de esta tabla, es posible notar que el algoritmo k-prototypes obtiene la mejor puntuación en el índice Davies-Bouldin, pero tiene la peor puntuación en el puntaje de silueta. Por otro lado, el algoritmo de agrupación jerárquica con enlace promedio logró la mejor puntuación de silueta, aunque su índice Davies-Bouldin no es tan bueno en comparación con k-prototypes. Finalmente, el algoritmo de agrupación jerárquica con enlace completo obtuvo una puntuación de silueta ligeramente mejor que k-prototypes, pero presenta el peor índice Davies-Bouldin entre los tres métodos evaluados.

Además, el algoritmo de agrupación jerárquica con enlace completo obtuvo la mayor cantidad adecuada de grupos, mientras que los otros métodos sugirieron un número mucho menor. Como se menciona anteriormente, una gran cantidad de grupos no se justifica al observar la distribución de los datos, lo que sugiere que este modelo podría estar sobreajustado. En contraste, el enlace promedio, utilizado en el algoritmo de agrupación jerárquica, mostró mejores resultados en ambos índices con una menor cantidad de grupos, lo que indica que el modelo tiene un mejor desempeño utilizando enlaces del tipo promedio.

Ahora bien, respecto a cuál método de agrupamiento es preferible entre k-prototypes y jerárquico con enlace promedio, se debe considerar qué métrica es más relevante para este caso de estudio. El objetivo del modelo es caracterizar zonas específicas donde no haya proyectos agrivoltaicos. Considerando esto, el índice Davies-Bouldin puede ser más relevante, dado que penaliza tanto la alta dispersión dentro de un grupo como la poca separación entre grupos. Esto es especialmente útil cuando se quiere asegurar que las zonas caracterizadas sean internamente consistentes (baja dispersión) y externamente diferenciadas (buena separación).

En resumen, mientras que un buen puntaje de silueta indica que los puntos están bien agrupados dentro de sus grupos y lejos de otros grupos, el índice de Davies-Bouldin proporciona que los grupos son compactos y distintos entre sí, lo cual es crítico para el objetivo de caracterización de zonas. Por lo tanto, para la construcción del modelo, es preferible utilizar el algoritmo k-prototypes.

6.3. Caracterización de los grupos formados

La selección del número de grupos y validación del modelo ha permitido identificar distintas agrupaciones de proyectos agrivoltaicos. Los resultados de este modelo se han representado en un mapa elaborado con Leaflet (ver figura 6.7), donde cada grupo está diferenciado por colores distintivos.

Figura 6.7: Mapa con el resultado de la agrupación de los proyectos agrivoltaicos.

Los proyectos se distribuyen en función de sus características compartidas, lo que facilita la visualización de patrones y tendencias geográficas. Es importante destacar que este mapa, elaborado con la agrupación de los proyectos, también forma parte del menú desplegable para la selección de características en la herramienta SIG agrivoltaica.

A continuación, se presenta una descripción detallada de cada grupo formado por el modelo.

6.3.1. Grupo 1

El primer grupo, compuesto por 94 proyectos y representado en azul en el mapa de la figura 6.7, se caracteriza por ser proyectos de baja capacidad instalada, generalmente menor a 1 [MW], con una capacidad promedio de 0,45 [MW]. Estos proyectos ocupan en promedio superficies de una hectárea y tienen paneles instalados a una altura de 3 metros respecto al suelo, aunque la mitad de los proyectos de este grupo no informan la altura de los paneles. La altitud promedio es de 240 metros sobre el nivel del mar. El clima en esta grupo suele ser templado, con tipos de clima Cfa (subtropical húmedo sin estación seca) y Cfb (oceánico templado), aunque una parte significativa de los datos muestra clima del tipo BSh (semiárido cálido). Los cultivos predominantes en este grupo son hortalizas y frutales. El diseño más común en este grupo es el de paneles instalados sobre el cultivo, y los paneles tienden a ser de arreglo fijo y monofaciales. En la tabla 6.2 se presenta un resumen de las características del grupo 1.

Variables	Valor promedio o más frecuente
Número de proyectos	94
Capacidad [MW]	0,45
Superficie [ha]	1
Altura [m]	3
Altitud [m s. n. m.]	240
Clima	Cfa y Cfb
Cultivo	Hortalizas y frutales
Panel	Monofacial
Seguimiento	Arreglo fijo
Diseño	Sobre cultivo

Tabla 6.2: Resumen de las características del grupo 1.

6.3.2. Grupo 2

El segundo grupo, compuesto por 100 proyectos y representado en amarillo en el mapa de la figura 6.7, se caracteriza por ser proyectos de gran escala con una capacidad instalada promedio de 55 [MW]. Estos proyectos ocupan en promedio superficies de 172 hectáreas, y la altura a la que se instalan los paneles suele no ser informada. La altitud promedio es de 95 metros sobre el nivel del mar. El tipo de clima más común en este grupo es Cfa (subtropical húmedo sin estación seca), aunque también incluye una importante proporción de otros tipos de clima, como BSk (semiárido frío), Cfb (oceánico templado), Dfa (continental templado sin estación seca) y Dfb (hemiboreal sin estación seca), entre otros. El cultivo principal en esta grupo son las praderas. En cuanto al diseño, no se ha detectado un patrón claro debido a la falta de datos. En cuanto al tipo de panel este suele ser monofacial (aunque una parte significativa de los proyectos no informa el tipo de panel) y cuenta con seguimiento en un eje. En la tabla 6.3 se presenta un resumen de las características del grupo 2.

Variables	Valor promedio o más frecuente
Número de proyectos	100
Capacidad [MW]	55
Superficie [ha]	172
Altura [m]	-
Altitud [m s. n. m.]	95
Clima	Cfa
Cultivo	Praderas
Panel	Monofacial
Seguimiento	Seguimiento en un eje
Diseño	-

Tabla 6.3: Resumen de las características del grupo 2.

6.3.3. Grupo 3

El tercer grupo, que abarca la mayoría de los datos (392 proyectos) y está representado en rojo en el mapa de la figura 6.7, se caracteriza por ser proyectos de tamaño intermedio, con una capacidad promedio de 10 [MW] y una superficie de 28 hectáreas. La altura a la que se instalan los paneles no es ampliamente informada, aunque el promedio es relativamente bajo, alrededor de 0,70 metros. La altitud promedio es de 395 metros sobre el nivel del mar. El clima en este grupo suele ser continental, con tipos Dfa (continental templado sin estación seca) y Dfb (hemiboreal sin estación seca). El cultivo principal en este grupo son las praderas. En cuanto al diseño, no se ha detectado un patrón claro debido a la falta de datos. El tipo de panel suele ser monofacial, y un poco más de la mitad de los proyectos cuentan con seguimiento en un eje, mientras que el resto tiene un arreglo fijo. En la tabla 6.4 se presenta un resumen de las características del grupo 3.

Variables	Valor promedio o más frecuente
Número de proyectos	392
Capacidad [MW]	10
Superficie [ha]	28
Altura [m]	0,7
Altitud [m s. n. m.]	395
Clima	Dfa y Dfb
Cultivo	Praderas
Panel	Monofacial
Seguimiento	Seguimiento en un eje y arreglo fijo
Diseño	-

Tabla 6.4: Resumen de las características del grupo 3.

6.3.4. Comparación entre los grupos formados

A partir de la caracterización de los grupos y su distribución en el mapa, se pueden observar algunas tendencias. El grupo 1 tiende a coincidir con proyectos piloto o de prueba,

lo que explica que sean pequeños en cuanto a capacidad instalada y superficie ocupada. Estos proyectos están mayormente distribuidos en Europa (excepto Países Bajos), India, Chile y en una cantidad significativa en la costa este de Estados Unidos.

En el caso del grupo 2, estos proyectos pueden ser considerados comerciales debido a su tamaño y la inversión requerida para proyectos de tal magnitud. El cultivo principal en este grupo son las praderas, y están ubicados principalmente en los Países Bajos, la costa este de Estados Unidos y el Estado de California. Sin embargo, al observar los proyectos en los Países Bajos de manera individual, se puede notar que sus características no se asemejan completamente a las del grupo 2. Esto puede deberse a que varias características no están informadas, y al imputar los datos, los algoritmos de imputación pudieron haber identificado similitudes con los proyectos de la costa este de Estados Unidos basándose en los pocos datos disponibles, como el clima. Esto resulta en una imputación de datos que no representa con precisión la realidad de los proyectos en los Países Bajos, donde por ejemplo la capacidad o la superficie utilizada por los proyectos es considerablemente menor al promedio del grupo.

Por último, el grupo 3 representa proyectos de capacidad y tamaño intermedios, que son considerablemente grandes para ser proyectos piloto o de estudio, lo que sugiere que probablemente también sean proyectos con fines comerciales. Estos proyectos están concentrados en el centro de Estados Unidos, en zonas donde predomina el clima continental o frío debido a la escasa influencia oceánica.

A continuación, en la tabla 6.5 se muestran las características de cada grupo formado.

Variables	Grupo 1	Grupo 2	Grupo 3
Número de proyectos	94	100	392
Capacidad [MW]	0,45	55	10
Superficie [ha]	1	172	28
Altura [m]	3	-	0,7
Altitud [m s. n. m.]	240	95	395
Clima	Cfa y Cfb	Cfa	Dfa y Dfb
Cultivo	Hortalizas y frutales	Praderas	Praderas
Panel	Monofacial	Monofacial	Monofacial
Seguimiento	Arreglo fijo	Seguimiento en un eje	Seguimiento en un eje y arreglo fijo
Diseño	Sobre cultivo	-	-

Tabla 6.5: Tabla comparativa de los grupos formados.

Algunos hallazgos que se pueden extraer de la tabla 6.5 y tendencias de los grupos son los siguientes:

- El grupo 3 es el que contiene la mayor cantidad de proyectos, mientras que los grupos 1 y 2 tienen un número similar de proyectos.
- Se observa una diferencia muy marcada entre la capacidad promedio instalada y la superficie promedio de cada grupo.
- Los proyectos pequeños o pilotos (grupo 1) prefieren una instalación de paneles a mayor altura, mientras que los proyectos con fines comerciales no.

- Los grupos 1 y 2 tienden a compartir tipos de climas similares, a diferencia del grupo 3.
- Los proyectos con fines comerciales (grupos 2 y 3), el tipo de cultivo más utilizado es la pradera, a diferencia del grupo 1, donde predominan las hortalizas y frutas.
- Un aspecto transversal en todos los grupos es el uso de paneles monofaciales.
- En cuanto al seguimiento, en el grupo 3 no se observa una tendencia clara sobre cuál es el más común, a diferencia de los grupos 1 y 2.

Capítulo 7

Conclusión

7.1. Síntesis de resultados

En el presente documento, se ha propuesto un sistema de información geográfica para evaluar soluciones agrivoltaicas. La metodología propuesta logró identificar y estudiar las características más relevantes de los proyectos agrivoltaicos analizados, junto con la construcción de un modelo de caracterización en base a técnicas de agrupamiento.

En la primera fase de la metodología, se desarrolla un sistema de información geográfica (SIG), una herramienta diseñada para visualizar los datos que caracterizan los proyectos agrivoltaicos. A partir de la recopilación de la información de cada proyecto, se observa que la gran mayoría de los proyectos georreferenciados pertenecen a Estados Unidos. Esto se debe a la existencia de una herramienta dedicada al catastro de los proyectos agrivoltaicos en ese país, algo que aún no se ha implementado o se encuentra en versiones preliminares en otros países.

Otro hallazgo importante al buscar información sobre proyectos agrivoltaicos es que las características mostradas dependen significativamente de la fuente consultada. Es difícil encontrar una fuente que proporcione información completa sobre todas las características definidas como relevantes. Una de las posibles razones para esta disparidad es que algunas empresas pueden ser reticentes a proporcionar información detallada de sus productos sin una cotización de por medio.

El SIG agrivoltaico es una página web desarrollada en lenguaje HTML, cuyo elemento central es un mapa con marcadores, generado mediante la biblioteca Leaflet. El usuario puede interactuar con el mapa, agrupando los proyectos según sus características y consultando las fichas informativas de cada uno. Esta herramienta es útil tanto para instituciones como para empresas con intereses en el sector agrivoltaico, ya que permite identificar proyectos similares y reconocer patrones en zonas geográficas específicas a partir de la experiencia internacional. Por ejemplo, al conocer el clima y la altitud de una zona específica, en la herramienta se puede buscar proyectos en áreas con características geográficas y climatológicas similares, y a partir de estos caracterizar la zona sin implementación de soluciones agrivoltaicas. Además, la herramienta proporciona una extensa base de datos que puede ser valiosa para la elaboración

de otros proyectos, como la construcción de modelos predictivos basados en la experiencia internacional.

En la segunda fase de la metodología, se estudiaron las variables más relevantes de los sistemas agrivoltaicos seleccionados. Para ello, se realiza un análisis univariado, un análisis multivariado y una identificación de valores faltantes.

Primero se realiza el análisis univariado, el cual revela que aproximadamente el 70 % de los proyectos agrivoltaicos tienen una capacidad inferior a 5 [MW], mientras que el 85 % no supera las 25 hectáreas de superficie. La mayoría de los proyectos se ubica a menos de 500 [m s. n. m.]. En cuanto a la altura de los paneles respecto al suelo, alrededor de la mitad de los proyectos se encuentran por debajo de 1.25 [m]. Sin embargo, este resultado debe interpretarse con cautela debido a que la altura con respecto al suelo presenta una cantidad significativa de datos faltantes.

El análisis univariado también mostró que los tipos de clima más comunes son Dfa (continental templado sin estación seca), Dfb (hemiboreal sin estación seca) y Cfa (subtropical húmedo sin estación seca). Por otro lado, aunque el tipo de panel presenta varios datos faltantes, se observa una clara tendencia hacia el uso de paneles monofaciales. En cuanto al tipo de seguimiento, no hay una preferencia clara entre los arreglos fijos y el seguimiento en un eje.

En el análisis multivariado, se observa un alto grado de correlación entre la capacidad instalada y la superficie ocupada por los proyectos, lo que sugiere una relación lineal entre ambas variables. Esto tiene sentido, ya que a medida que aumenta la cantidad de paneles, es decir, la capacidad instalada, también aumenta la superficie requerida. No se observaron correlaciones significativas entre el resto de las variables.

En cuanto a la identificación de valores faltantes, se detecta que hay características que no están siendo informadas por las fuentes consultadas, como la altura del panel respecto al suelo, el tipo de cultivo, el tipo de panel, el tipo de seguimiento y el tipo de diseño de la planta agrivoltaica.

En la tercera fase de la metodología, se determina que el método de agrupamiento que ofrece el mejor desempeño para caracterizar zonas específicas es el algoritmo K-prototypes. Para este algoritmo, se ha identificado que el número más adecuado de grupos es 3.

El grupo 1, con 94 proyectos, incluye proyectos de baja capacidad (0,45 [MW]) que ocupan en promedio 1 hectárea y tienen paneles instalados a 3 metros del suelo. La altitud promedio es de 240 metros sobre el nivel del mar, y el clima predominante es templado, con tipos Cfa (subtropical húmedo sin estación seca) y Cfb (oceánico templado). Los cultivos principales son hortalizas y frutales, y el diseño más común es con paneles monofaciales y arreglo fijo.

El grupo 2, compuesto por 100 proyectos, se caracteriza por proyectos de gran escala con una capacidad promedio de 55 [MW] y superficies de 172 hectáreas. La altitud promedio es de 95 metros, y el clima más común es Cfa (subtropical húmedo sin estación seca). Los cultivos predominantes son praderas, y los paneles suelen ser monofaciales con seguimiento en un eje.

El grupo 3, que incluye 392 proyectos, abarca proyectos de tamaño intermedio con una capacidad promedio de 10 [MW] y superficies de 28 hectáreas. La altura promedio de los paneles es de 0,70 metros, y la altitud promedio es de 395 metros sobre el nivel del mar. El clima es continental, con tipos Dfa (continental templado sin estación seca) y Dfb (hemiboreal sin estación seca). Los cultivos principales son praderas, y los paneles suelen ser monofaciales, con un poco más de la mitad de los proyectos con seguimiento en un eje y el resto con arreglo fijo.

Entre los resultados obtenidos también se incluye el mapa agrivoltaico de Chile, mostrado en el anexo A. Este mapa es un subproducto de la investigación, el cual no estaba inicialmente contemplado realizar en la propuesta metodológica. Es importante destacar que este constituye el primer mapa de proyectos agrivoltaicos en Chile y ha sido utilizado por Energy Partnership Chile-Alemania y Fraunhofer Chile para la elaboración de un "policy brief" (informe de políticas) sobre el uso compartido de suelos para la agricultura y la generación de energía solar fotovoltaica.

7.2. Evaluación de objetivos propuestos

En esta sección se revisitan los objetivos propuestos en la sección 1.2.

El primer objetivo es elaborar un mapa georreferenciado de proyectos agrivoltaicos. Este objetivo se ha cumplido en su totalidad, ya que la herramienta SIG desarrollada (ver sección 4.5) incluye una serie de mapas que agrupan los proyectos según sus características, además de que cada proyecto mostrado cuenta con una ficha que detalla sus características más relevantes.

El segundo objetivo es identificar y estudiar las características más relevantes de los sistemas agrivoltaicos seleccionados. Este objetivo también está cumplido, puesto que el análisis exploratorio de los datos ha logrado identificar patrones significativos y tendencias clave en las variables de interés. A través de métodos estadísticos, se ha podido observar cuáles son las variables más comunes, en qué rangos se distribuyen, y cómo se relacionan entre sí. Esto ha permitido una mejor comprensión de las dinámicas subyacentes en los sistemas agrivoltaicos.

El tercer objetivo corresponde a la construcción de un modelo para caracterizar zonas específicas donde no se han desarrollado proyectos agrivoltaicos. Este objetivo se ha cumplido, aunque presenta algunas limitaciones. En primer lugar, el modelo construido a partir del agrupamiento de datos permite caracterizar zonas sin implementación agrivoltaica, conociendo el tipo de clima y la altitud de una zona específica, es posible a través de la herramienta SIG, buscar proyectos que compartan características similares. A partir de esa búsqueda, se puede caracterizar la zona sin implementación agrivoltaica.

No obstante, como se menciona, presenta algunas limitantes. Por ejemplo, la herramienta enfrenta dificultades al agrupar proyectos con varios datos faltantes, como se observa en los proyectos ubicados en los Países Bajos. Otra limitación es que, para caracterizar una zona, es necesario que existan proyectos en áreas similares, de lo contrario, la caracterización no es posible.

Por último, el cuarto objetivo es validar el modelo, el cual está completo tanto por los índices de evaluación de calidad de grupo como por inspección visual. La validación se ha realizado utilizando métricas específicas, como la puntuación de silueta y el índice Davies-Bouldin, para evaluar la cohesión y separación de los clusters. Además, se ha llevado a cabo una revisión visual en la herramienta elaborada para comprobar la coherencia y la representación adecuada de los datos en el modelo.

En conclusión, a partir del trabajo realizado se obtuvo una herramienta que permite caracterizar y agrupar soluciones agrivoltaicas según sus características.

7.3. Trabajo Futuro

Con el objetivo de complementar este trabajo, se identificaron varios aspectos que podrían haberse abordado desde una perspectiva diferente, junto con sugerencias para dar continuidad al estudio:

- Para el análisis multivariado, se sugiere buscar métodos avanzados que puedan estudiar las posibles relaciones entre las variables numéricas y categóricas de manera más efectiva. Métodos como el análisis de componentes principales (PCA) para datos mixtos podrían ofrecer una visión más comprensiva de las interacciones entre variables.
- En la imputación de datos, se recomienda utilizar un método que trabaje simultáneamente con datos numéricos y categóricos, asegurando una imputación más precisa y coherente. Algoritmos basados en aprendizaje automático, como los bosques aleatorios, pueden ser particularmente útiles en este contexto. Sin embargo, este enfoque no se aplicó debido a que su implementación en código resultó ser más complicada de lo que se pensó inicialmente.
- Se sugiere continuar trabajando en la recopilación de información, especialmente en la obtención de datos faltantes de algunos proyectos. Tener información más detallada puede conducir a un modelo de caracterización más preciso y robusto. Para ello, se recomienda contactar a las empresas o propietarios detrás de estos proyectos.
- Se sugiere revisar constantemente si se han agregado nuevos proyectos agrivoltaicos a las bases de datos mencionadas en la sección 2.6, con el fin de mantener actualizada la herramienta. Asimismo, se recomienda buscar proyectos en nuevos países. Para ello, puede ser útil visitar medios de comunicación especializados en el sector fotovoltaico y agrivoltaico, como PV Magazine.
- Con el fin de aprovechar al máximo la base de datos construida, se recomienda analizar el uso de algoritmos de aprendizaje supervisado como lo son los algoritmos de clasificación y regresión. Algoritmos como las regresiones lineales y logísticas, máquinas de soporte vectorial (SVM), redes neuronales artificiales, y modelos de ensemble como el boosting y bagging, pueden permitir predecir variables basándose en la experiencia internacional.
- Se sugiere utilizar un algoritmo de agrupamiento mixta alternativo para evaluar si se pueden obtener mejores resultados en comparación con los presentados en este trabajo.
- Con respecto a la herramienta web, se recomienda que, al cambiar de característica, no se reemplace el mapa completo, sino que únicamente se modifiquen los colores de los marcadores. Esto no solo mejorará la experiencia del usuario al hacer las transiciones

más rápidas y suaves, sino que también reducirá la carga de procesamiento y el tiempo de carga del mapa.

- Se sugiere agregar una opción en la herramienta que permita mostrar un mapa topográfico y un mapa de la distribución de los tipos de clima según la clasificación de Köppen-Geiger, junto con los marcadores de cada uno de los proyectos.

Bibliografía

[1] Jones et al. - with major processing by Our World in Data., "Annual greenhouse gas emissions," 2024, [dataset]. Jones et al., "National contributions to climate change 2024.1" [original data].

[2] InSPIRE/Sites/Jacks Solar Garden | Open Energy Information, "OpenEI.org," accedido el 18 de mayo de 2024. [En línea]. Disponible: https://openei.org/wiki/InSPIRE/Sites/Jacks_Solar_Garden.

[3] scikit-learn, "2.3. Clustering — scikit-learn 0.20.3 documentation," accedido el 10 de Mayo de 2024. [En línea]. Disponible: https://scikit-learn.org/stable/modules/clustering.html.

[4] Biblioteca del Congreso Nacional (BCN), "Carbono neutralidad en el sector energético de Chile," accedido el 25 de octubre de 2023. [En línea]. Disponible: https://obtienearchivo.bcn.cl/obtienearchivo?id=repositorio/10221/32578/1/BCN_Carbononeutralidad_en_el_sector_energetico_Chile_15Oct._Rev._RT01_edPM.pdf.

[5] "Ministerio del medio ambiente 2022 5to informe bienal de actualización ante la convención marco de las naciones unidas sobre cambio climático," accedido el 20 de agosto de 2024. [En línea]. Disponible: https://cambioclimatico.mma.gob.cl/wp-content/uploads/2022/12/Informe_5IBA_2022.pdf.

[6] Ministerio del Medio Ambiente, "Informe del Inventario Nacional de Chile 2022: Inventario nacional de gases de efecto invernadero y otros contaminantes climáticos 1990-2020," accedido el 30 de noviembre de 2023. [En línea]. Disponible: https://snichile.mma.gob.cl/wp-content/uploads/2023/04/2022_IIN_CL.pdf.

[7] UNFCCC, "El Acuerdo de París," accedido el 30 de noviembre de 2023. [En línea]. Disponible: https://unfccc.int/es/acerca-de-las-ndc/el-acuerdo-de-paris.

[8] IPCC, "Calentamiento global de $1,5°C$," accedido el 30 de noviembre de 2023. [En línea]. Disponible: https://www.ipcc.ch/site/assets/uploads/sites/2/2019/09/IPCC-Special-Report-1.5-SPM_es.pdf.

[9] Naciones Unidas, "Desafío globales: Cambio climático," accedido el 18 de mayo de 2024. [En línea]. Disponible: https://www.un.org/es/global-issues/climate-change.

[10] C. Dupraz, H. Marrou, G. Talbot, L. Dufour, A. Nogier, and Y. Ferard, "Combining solar photovoltaic panels and food crops for optimising land use: Towards new agrivoltaic schemes," *Renewable Energy*, vol. 36, p. 2725–2732, 10 2011, [En línea]. Disponible: https://www.sciencedirect.com/science/article/abs/pii/S0960148111001194https://doi.org/10.1016/j.renene.2011.03.005.

[11] A. Guleed and K. Farid, "SHADING ANALYSIS OF AGRIVOLTAIC SYSTEMS," 2023, accedido el 5 de diciembre de 2023. [En línea]. Disponible: https://mdh.diva-portal.org/smash/get/diva2:1781356/FULLTEXT01.pdf.

[12] S. A. y Ganadero (SAG), "2011 (rectificada). pauta para estudio de suelos servicio agrÍcola y ganadero," [En línea]. Disponible: https://www.sag.gob.cl/sites/default/files/pauta-para-estudio-de-suelos--mod-2016.pdf.

[13] T. Sekiyama and A. Nagashima, "Solar sharing for both food and clean energy production: performance of agrivoltaic systems for corn, a typical Shade-Intolerant crop," *Environments*, vol. 6, no. 6, p. 65, Jun. 2019, [En línea]. Disponible: https://doi.org/10.3390/environments6060065.

[14] (PAS) DIN SPEC, "Agri-photovoltaic systems - Requirements for primary agricultural use," 2021, (DIN SPEC 91434:2021-05) [En línea]. Disponible: https://dx.doi.org/10.31030/3257526.

[15] M. Laub, L. Pataczek, A. Feuerbacher, S. Zikeli, and P. Högy, "Contrasting yield responses at varying levels of shade suggest different suitability of crops for dual land-use systems: A meta-analysis," *Agronomy Sustain. Develop.*, vol. 42, no. 3, Jun. 2022, [En línea]. Disponible: https://doi.org/10.1007/s13593-022-00783-7.

[16] S. Neupane Bhandari, S. Schlüter, W. Kuckshinrichs, H. Schlör, R. Adamou, and R. Bhandari, "Economic feasibility of agrivoltaic systems in food-energy nexus context: Modelling and a case study in niger," *Agronomy*, vol. 11, no. 10, p. 1906, Sep. 2021, [En línea]. Disponible: https://doi.org/10.3390/agronomy11101906.

[17] Víctor Olaya, *Sistemas de información geográfica.* Autoedición, 2014.

[18] G. B. Korte, *The GIS book*, 4th ed. Santa Fe: OnWord Press, 1997.

[19] F. J. Martínez López and A. Gallegos Ruiz, *Programación de datos relacionales.* Madrid, España: Rama Ed., 2017.

[20] L. Hueso Ibáñez, *Bases de datos.* Madrid, España: Rama Ed., 2014.

[21] J. R. Capacho Portilla and W. Nieto Bernal, *Diseño de base de datos.* Barranquilla, Colombia: Univ. Del Norte, 2017.

[22] M. Harmouch, "17 types of similarity and dissimilarity measures used in data science," Apr. 2021, accedido el 10 de Mayo de 2024. [En línea]. Disponible: https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science.

[23] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, vol. 27, no. 4, p. 857, Dec. 1971, [En línea]. Disponible: https://doi.org/10.2307/2528823.

[24] Z. Huang, "Data Mining and Knowledge Discovery," *Discovery*, vol. 2, no. 3, p. 283–304, 1998, [En línea]. Disponible: https://doi.org/10.1023/a:1009769707641.

[25] I. Meza, "Clase 12: Clustering," 2024, gitHub repository, Accedido el 12 de Mayo de 2024.[En línea]. Disponible: https://github.com/MDS7202/MDS7202/blob/main/clases/2024-01/12_Clustering.ipynb.

[26] scikit-learn, "Selecting the number of clusters with silhouette analysis on KMeans clustering," accedido el 12 de Mayo de 2024.[En línea]. Disponible: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py.

[27] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, p. 224–227, Apr. 1979, [En línea]. Disponible: https://doi.org/10.1109/tpami.1979.4766909.

[28] P. Badilla, "Clase 16 - Manejo de Valores Faltantes," 2022, accedido el 25 de Junio de 2024.[En línea]. Disponible: https://github.com/MDS7202/MDS7202/blob/main/clases/2022-01/16_Manejo_de_Valores_Faltantes/Clase_16-Manejo_de_Valores_Faltantes.ipynb.

[29] NREL, "Information | Open Energy Information," accedido el 25 de octubre de 2023. [En línea]. Disponible: https://openei.org/wiki/Information.

[30] ——, "InSPIRE/Agrivoltaics Map | Open Energy Information," accedido el 25 de octubre de 2023. [En línea]. Disponible: https://openei.org/wiki/InSPIRE/Agrivoltaics_Map.

[31] NSEFI, "About Us - NSEFI | NSEFI - National Solar Energy Federation of India," accedido el 30 de noviembre de 2023. [En línea]. Disponible: https://nsefi.in/about-us/.

[32] ——, "Agrivoltaics Map | Agrivoltaics website," accedido el 30 de noviembre de 2023. [En línea]. Disponible: https://www.agrivoltaics.in/agripv-map-of-india.

[33] Fraunhofer ISE, "Agrivoltaic Facilities in Germany. Agri-Photovoltaik: Chance für Landwirtschaft und Energiewende," 2024, accedido el 8 de julio de 2024. [En línea]. Disponible: https://agri-pv.org/en/community/agrivoltaic-facilities-in-germany/.

[34] F. Rubel, K. Brugger, K. Haslinger, and I. Auer, "The climate of the european alps: Shift of very high resolution köppen-geiger climate zones 1800–2100," *Meteorologische Zeitschrift*, vol. 26, pp. 115–125, 04 2017, [En línea]. Disponible: https://koeppen-geiger.vu-wien.ac.at/pdf/Paper_2017.pdf.

[35] I. Meza, "Clase 9 - Análisis Exploratorio de Datos," 2024, accedido el 16 de Mayo de 2024.[En línea]. Disponible: https://github.com/MDS7202/MDS7202/blob/main/clases/2024-01/09_Analisis_Exploratorio_de_Datos.ipynb.

[36] R. Drennan, V. González Fernández, "Estadística para arqueólogos: Un enfoque de sentido común," 2019, [En línea]. Disponible: https://www.digitaliapublishing.com/a/60794b.

[37] M. L. McHugh, "The chi-square test of independence," *Biochemia Medica*, vol. 23, pp. 143–149, 06 2013, [En línea]. Disponible: https://doi.org/10.11613/bm.2013.018.

[38] D. J. Stekhoven and P. Buhlmann, "Missforest–non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, pp. 112–118, 10 2011, [En línea]. Disponible: https://doi.org/10.1093/bioinformatics/btr597.

Anexo A

Mapa Agrivoltaico de Chile

Mapa agrivoltaico que muestra los proyectos censados en Chile. Este mapa es parte del "policy brief" elaborado por Energy Partnership Chile-Alemania, el cual aborda el uso compartido de suelos para la agricultura y la generación de energía fotovoltaica [1].

[1] Acceso al policy brief.

Anexo B

Tablas de Contingencia

A continuación, se presentan las tablas de contingencia para cada par de variables categóricas.

Cultivo Clima	Berries	Cereales	Cultivos industriales	Frutales	Hortalizas	Praderas	Vides
Af	0	0	0	0	1	0	0
Am	0	0	0	0	1	0	0
Aw	0	1	0	0	1	0	0
BSh	0	3	0	4	10	5	0
BSk	0	1	0	1	5	20	0
BWh	0	0	0	0	2	0	0
BWk	0	0	0	0	1	0	0
Cfa	1	0	0	0	2	33	0
Cfb	11	1	2	4	2	4	1
Csa	0	1	0	2	2	4	3
Csb	1	0	0	3	5	4	1
Cwa	0	0	0	0	1	0	0
Dfa	0	0	0	0	7	17	0
Dfb	1	0	0	0	2	25	0
Dsc	0	0	0	0	1	0	0

Tabla B.1: Tabla de contingencia para las variables clima y cultivo.

Panel Clima	Bifacial	Monofacial	Semi-transparente
Af	1	0	0
Am	0	1	0
Aw	0	2	0
BSh	1	15	2
BSk	3	12	2
BWh	0	2	0
Cfa	5	22	1
Cfb	4	0	3
Csa	1	3	0
Csb	4	8	2
Cwa	0	1	0
Dfa	3	56	0
Dfb	2	31	0
Dsc	1	0	0

Tabla B.2: Tabla de contingencia para las variables clima y panel.

Seguimiento Clima	Fijo	Seguimiento de un solo eje	Seguimiento en dos ejes
Af	0	1	0
Am	1	0	0
Aw	3	0	0
BSh	16	6	1
BSk	8	29	0
BWh	2	0	0
BWk	0	0	1
Cfa	32	43	4
Cfb	10	5	0
Csa	4	9	0
Csb	11	8	0
Cwa	1	0	0
Dfa	102	130	0
Dfb	62	61	0
Dsc	1	0	0

Tabla B.3: Tabla de contingencia para las variables clima y seguimiento.

Diseño Clima	Entre hilera	Invernadero	Sobre cultivo	Vertical
Am	1	0	0	0
Aw	1	0	2	0
BSh	3	1	12	1
BSk	3	1	3	0
BWh	1	0	0	0
BWk	0	0	1	0
Cfa	0	1	6	0
Cfb	4	2	18	1
Csa	0	0	7	1
Csb	4	1	6	1
Cwa	1	0	0	0
Dfa	2	0	3	0
Dfb	0	0	3	0

Tabla B.4: Tabla de contingencia para las variables clima y diseño.

Panel Cultivo	Bifacial	Monofacial	Semi-transparente
Berries	1	1	0
Cereales	0	5	0
Cultivos industriales	2	0	0
Frutales	2	3	4
Hortalizas	6	24	5
Praderas	4	38	0
Vides	1	0	0

Tabla B.5: Tabla de contingencia para las variables cultivo y panel.

Seguimiento Cultivo	Fijo	Seguimiento de un solo eje	Seguimiento en dos ejes
Berries	3	1	0
Cereales	5	1	0
Cultivos industriales	0	2	0
Frutales	4	7	0
Hortalizas	31	9	2
Praderas	45	63	0
Vides	2	2	0

Tabla B.6: Tabla de contingencia para las variables cultivo y seguimiento.

Diseño Cultivo	Entre hilera	Invernadero	Sobre cultivo	Vertical
Berries	0	1	9	0
Cereales	2	0	5	0
Cultivos industriales	0	0	2	0
Frutales	2	1	11	0
Hortalizas	10	2	20	1
Praderas	2	0	2	0
Vides	0	0	4	1

Tabla B.7: Tabla de contingencia para las variables cultivo y diseño.

Seguimiento Panel	Fijo	Seguimiento de un solo eje	Seguimiento en dos ejes
Bifacial	8	16	0
Monofacial	91	57	5
Semi-transparente	8	2	0

Tabla B.8: Tabla de contingencia para las variables panel y seguimiento.

Diseño Panel	Entre hilera	Invernadero	Sobre cultivo	Vertical
Bifacial	2	1	8	3
Monofacial	11	0	27	0
Semi-transparente	0	4	4	0

Tabla B.9: Tabla de contingencia para las variables panel y diseño.

Diseño Seguimiento	Entre hilera	Invernadero	Sobre cultivo	Vertical
Fijo	12	4	30	3
Seguimiento de un solo eje	4	0	15	0
Seguimiento en dos ejes	0	0	6	0

Tabla B.10: Tabla de contingencia para las variables seguimiento y diseño.

Anexo C

Índices de evaluación de calidad de grupo para k-prototypes

En este anexo se presentarán los gráficos obtenidos para la puntuación de silueta y el índice de Davies-Bouldin con las semillas aleatorias 6 y 19.

Cuando la semilla encargada de la aleatoriedad es 6, se obtiene que la cantidad adecuada de grupos es 3, tal como se observa en la figura C.1 que muestra la puntuación de silueta y en la figura C.2 que presenta el índice Davies-Bouldin.

Cuando la semilla encargada de la aleatoriedad es 19, se obtiene que la cantidad adecuada de grupos también es 3, tal como se observa en la figura C.3 que muestra la puntuación de silueta y en la figura C.4 que presenta el índice Davies-Bouldin.

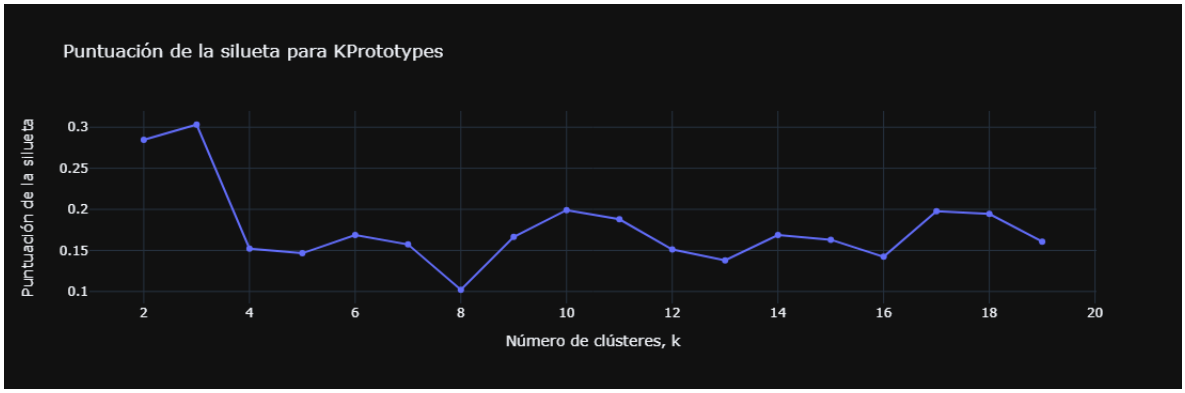

Figura C.1: Puntuación de silueta según la cantidad k de grupos para el algoritmo k-prototypes, cuando la semilla aleatoria es 6.

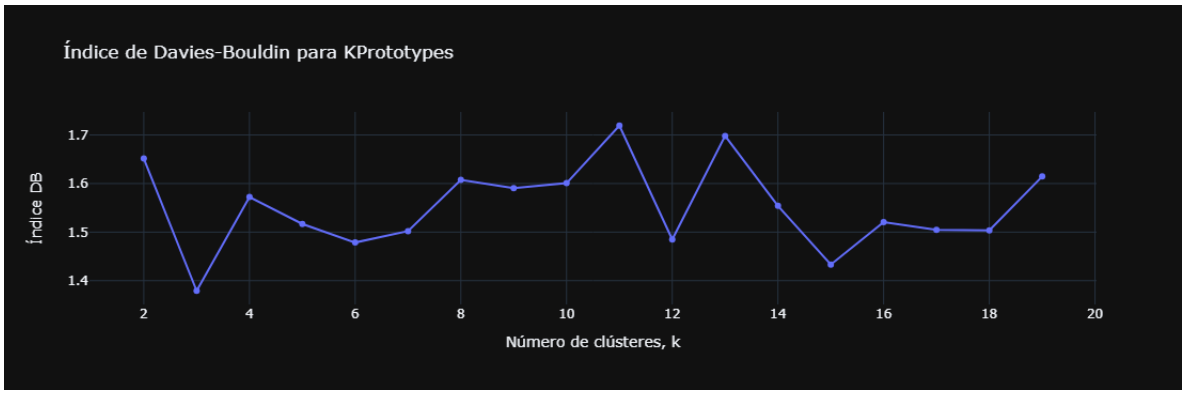

Figura C.2: Índice de Davies-Bouldin según la cantidad k de grupos para el algoritmo k-prototypes, cuando la semilla aleatoria es 6.

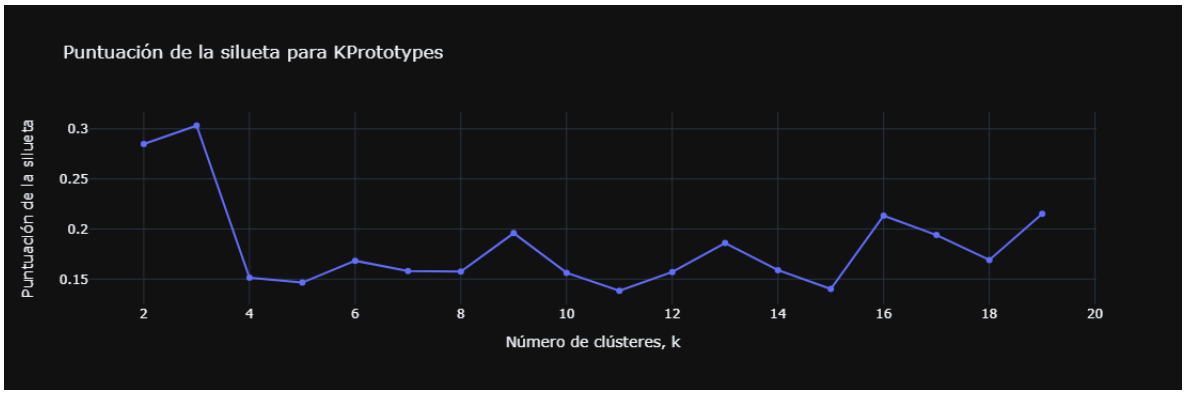

Figura C.3: Puntuación de silueta según la cantidad k de grupos para el algoritmo k-prototypes, cuando la semilla aleatoria es 19.

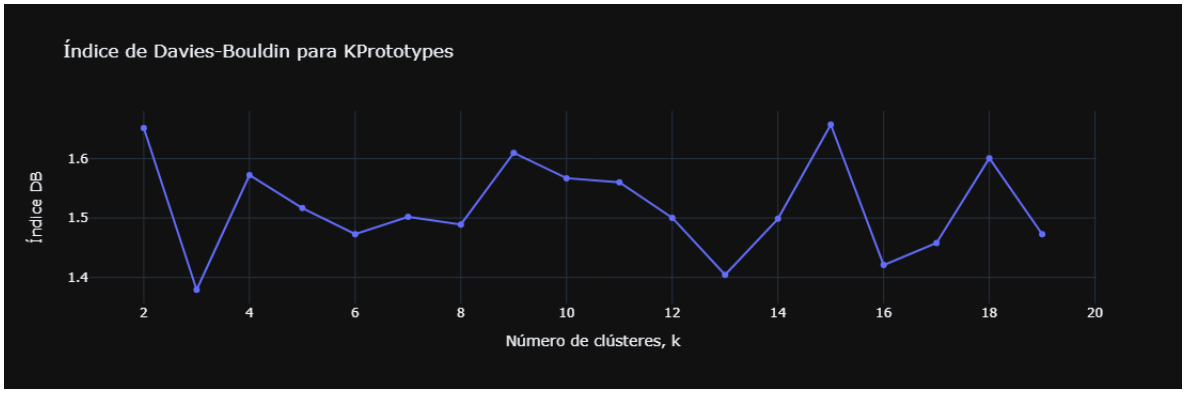

Figura C.4: Índice de Davies-Bouldin según la cantidad k de grupos para el algoritmo k-prototypes, cuando la semilla aleatoria es 19.

Anexo D

Repositorio

En el presente trabajo se incluye como anexo un repositorio digital que contiene diversos archivos relevantes para el desarrollo de esta investigación. La finalidad de este repositorio es proporcionar un acceso fácil y centralizado a los recursos utilizados y generados durante el proceso de investigación, facilitando su consulta y revisión por parte de los interesados.

El repositorio contiene los siguientes elementos:

- Una copia de seguridad (en inglés "dump") de los datos almacenados, para restaurar o transferir los datos a otra base de datos (Dump_Database.sql).
- Catastro de los proyectos agrivoltaicos (Catastro.xlsx).
- El notebook utilizado para el análisis de los datos y la construcción del modelo (EDA_Modelo.ipynb).
- Una carpeta con todos los archivos HTML necesarios para ejecutar la herramienta SIG en un navegador. La herramienta como tal se encuentra en el archivo denominado "Herramienta_SIG.html".

Este repositorio se encuentra accesible a través del siguiente enlace: `https://drive.google.com/drive/folders/1xrSWVhz9LkaUOFXobP6KcoXhTyAHlosL?usp=drive_link`