



UNIVERSIDAD DEL MAGDALENA
Facultad de Ingeniería
Programa de Ingeniería Electrónica

Proyecto de Trabajo de Grado

**Sistema para la selección de módulos fotovoltaicos
comercializados en Colombia basado en la
comparación de datos ambientales e información
técnica proporcionada por los fabricantes**

Alexander Jesús Arroyo Granados
Manuel Alfonso Manjarres Rivera



**Sistema para la selección de módulos
fotovoltaicos comercializados en Colombia
basado en la comparación de datos
ambientales e información técnica
proporcionada por los fabricantes**

Alexander Jesús Arroyo Granados

Manuel Alfonso Manjarres Rivera



UNIVERSIDAD DEL MAGDALENA

SANTA MARTA, COLOMBIA

29 de noviembre de 2022

Prefacio

Este trabajo de grado se presentó como parte de los requisitos para optar al grado profesional como Ingeniero Electrónico, de la Universidad del Magdalena y no ha sido presentado previamente para la obtención de otro título en esta Universidad u otras. El mismo contiene el plan de investigación que se llevó a cabo al interior del grupo de investigación MAGMA Ingeniería, adscrito al programa de ingeniería electrónica de la Universidad del Magdalena, durante el período comprendido entre el mes de **junio** de **2022** y el mes de **octubre** de **2022**, bajo la dirección de **Víctor José Olivero Ortiz** y los co-directores **Carlos Arturo Robles Algarín** y **Diego Andrés Restrepo Leal**.

Alexander Jesús Arroyo Granados

alexanderarroyojg@unimagdalena.edu.co

Manuel Alfonso Manjarres Rivera

manuelmanjarresar@unimagdalena.edu.co

FACULTAD DE INGENIERÍA

UNIVERSIDAD DEL MAGDALENA

Santa Marta, 29 de noviembre de 2022

Resumen

En esta investigación se plantea el desarrollo de un sistema experto artificial capaz de recomendar un módulo fotovoltaico considerando la información técnica entregada por el fabricante y las condiciones ambientales presentes en diversas zonas geográficas de la región caribe colombiana. Esto, con el objetivo de facilitar la selección del módulo que mejor se adapte a las condiciones de trabajo. Por lo cual, se propone implementar métodos web scraping para la extracción de la información de especificaciones técnicas dadas por el fabricante sobre cierta cantidad de módulos fotovoltaicos que son comercializados en el país, que con la integración de procedimientos de ingeniería de características para el tratamiento de los datos permitan obtener conjuntos de datos óptimo para el entrenamiento de modelos inteligentes.

Además, se hace uso de los conjuntos de datos para entrenar modelos inteligentes de clasificación capaces de predecir valores desconocidos dentro de un grupo establecido, y relacionarlos con uno o varios módulos fotovoltaicos que se adapten a las condiciones de temperatura e irradiación. Finalmente, se desarrolla una interfaz web que permite a los usuarios interactuar con el modelo inteligente y obtener resultados de los paneles que mejor se adaptan según sean sus condiciones iniciales de temperatura e irradiación, y observar las características relevantes de dichos resultados.

Índice general

Índice de tablas	5
Índice de figuras	6
1. INTRODUCCIÓN	8
Referencias	12
2. WEB SCRAPING	15
2.1. Extracción de la información de los módulos fotovoltaicos	15
2.2. Identificación y selección de la fuente de los datos	17
2.3. Extracción de los datos utilizando técnicas de Web Scraping	22
2.4. Resultados	26
2.5. Conclusiones	33
Referencias	35
3. INGENIERÍA DE CARACTERÍSTICAS	37
3.1. Procesamiento y caracterización de los datos	37
3.2. Datos de temperatura e irradiación	39
3.3. Estimación de la generación de energía	42
3.4. Etiquetando los datos	45

<i>ÍNDICE GENERAL</i>	4
3.5. Resultados	49
3.6. Conclusiones	50
Referencias	52
4. ENTRENAMIENTO DE UN MODELO INTELIGENTE	54
4.1. Selección del modelo de aprendizaje	54
4.2. Entrenamiento del modelo de aprendizaje	58
4.3. Resultados	62
4.4. Conclusiones	62
Referencias	64
5. IMPLEMENTACIÓN DE UNA APLICACIÓN WEB	66
5.1. Funcionamiento general	66
5.2. BackEnd	67
5.3. FrontEnd	68
5.4. Despliegue de la aplicación web	70
5.5. Conclusiones	71

Índice de tablas

2.1. Datos extraídos del sitio Web es.ensolar.com	24
2.2. Pruebas realizadas a los módulos fotovoltaicos	33
3.1. Estructura de la tabla	40
4.1. Modelos entrenados	62

Índice de figuras

1.1. Diagrama de etapas del proyecto	10
2.1. Vista principal del sitio es.ensolar.com	18
2.2. Lista de fabricantes seleccionados	19
2.3. Módulos solares del fabricante Jinko Solar	20
2.4. Información STC del módulo FV Tiger Pro 72HC 530-550	21
2.5. Información NOCT y térmica del módulo FV Tiger Pro 72HC 530-550	21
2.6. Diagrama de funcionamiento de Scrapy Framework	22
2.7. Secciones seleccionadas con la expresión xpath correspondiente	23
2.8. Secciones seleccionadas con la expresión xpath correspondiente	24
2.9. Cartel de aviso de bloqueo del sitio Web es.ensolar.com	25
2.10. Cantidad de módulos fotovoltaicos por fabricantes	27
2.11. Cantidad de módulos fotovoltaicos monocristalinos por fabricante	28
2.12. Cantidad de módulos fotovoltaicos policristalinos por fabricante	29
2.13. Cantidad de módulos fotovoltaicos Bifacial por fabricante	30
2.14. Cantidad de módulos fotovoltaicos Perc por fabricante	30
2.15. Cantidad de módulos fotovoltaicos capa fina y bifacial hjt por fabricante	31
2.16. Cantidad de módulos fotovoltaicos capa bifacial Perc por fabricante	31
2.17. Cantidad de módulos fotovoltaicos por tecnología	32
2.18. Distribución de las potencias de los módulos fotovoltaicos	33

3.1. Vista principal del Climate Change Knowlege Portal	40
3.2. Vista principal del PVGIS tool	41
3.3. Universidad del Magdalena vista desde el mapa PVGIS tool	41
3.4. Diagrama de flujo del proceso de cálculo de la generación de energía	44
3.5. Pairplot de las columnas mes, año, temperatura, irradiación y gene- ración	46
3.6. Elbow curve	47
3.7. Diagrama de flujo del proceso de etiquetado de los datos	48
3.8. Datos agrupados por el algoritmo K-means	50
4.1. Road map de scikit-learn	56
4.2. Distribución de la irradiación	57
4.3. Distribución de la temperatura	57
4.4. Distribución de la potencia	58
4.5. SDG Classfier (logistic-regression)	59
4.6. SDG Classfier (linear-SVC)	60
4.7. SDG Classifier usando search grid	61
5.1. Estructura general	67
5.2. BackEnd con FastAPI y MariaDB	68
5.3. Vista principal de la aplicación web	69
5.4. Resultados de la petición	70
5.5. Información STC del módulo FV	70

Capítulo 1

INTRODUCCIÓN

La creciente popularidad de las energías limpias, el cumplimiento de los objetivos de desarrollo sostenible (ODS), la transición energética que se espera ocurra en los próximos años y el uso de las energías renovables como objetivo para los países del mundo [1] han impulsado el uso de energías alternativas, las cuales provienen de recursos ilimitados como el viento y el sol, como es el caso de la energía eólica y la energía solar fotovoltaica. Sin embargo, la implementación de proyectos con este tipo de energías se enfrenta a diversos retos para su correcta operación. En el caso de la energía solar fotovoltaica, la energía que se puede generar a partir de un módulo fotovoltaico a lo largo de su vida útil varía según su eficiencia y las condiciones ambientales a las que se ve expuesto. Es por esta razón que los fabricantes de módulos fotovoltaicos describen su comportamiento mediante pruebas de tipo STC (Standard Test Conditions) y PTC (PVUSA Test Conditions) realizadas bajo condiciones controladas [2]. Esta información es puesta de forma detallada en hojas técnicas y publicadas en sus páginas web oficiales, aportando así a un mayor conocimiento de la tecnología y a su vez generando una cantidad de datos e información que necesita ser analizada.

Los datos pueden ser generados por la interacción del ser humano con el mundo

digital o bien por dispositivos IoT [3] obteniendo diariamente alrededor de 2.5 quintillones de datos en el planeta, y se estima que en los últimos años se han generado por lo menos el 90 % de los datos de toda la historia [4]. Dicha cantidad de datos ha tomado una importancia significativa en la toma de decisiones. Debido a esto se han creado técnicas para examinar y extraer la información de páginas web, una de ellas, por ejemplo, el web scraping, usada comúnmente para: la creación de bases de datos con información de contactos, relacionar diversas fuentes online y reunir datos con fines de investigación [5] [6]. Estas técnicas son comúnmente desarrolladas sobre lenguajes de programación con una alta difusión y adopción de librerías e información disponible como lo es Python [7], Además, estas técnicas van de la mano con herramientas para el manejo de estructuras de datos, en las cuales se pueden almacenar desde pequeñas hasta grandes cantidades de información listas para el procesamiento [8]. El aprovechamiento de estos recursos tecnológicos es clave para la creación de nuevas herramientas capaces de ayudarnos a tomar decisiones basadas en el análisis de datos, sin embargo, el manejo de big data presenta entre sus retos: la forma en cómo están estructurados los datos y la capacidad del hardware para almacenar y procesar información [9]. Por otro lado, debido al gran volumen de datos, es necesario seleccionar adecuadamente los conjuntos data frames que contengan la información idónea para que el modelo se ajuste [10].

Al existir una alta cantidad de modelos de predicción, la comparación de estos demuestra que, según las variables de los factores estudiados, las métricas de desempeño pueden variar de acuerdo con lo esperado [11]. Cabe destacar que, para el pronóstico de energía solar fotovoltaica se han implementado en diferentes estudios, diferentes técnicas de predicción, que mayormente se han basado en modelos estadísticos, artificial neural network (ANN) y técnicas más complejas como extreme learning machine (ELM). No obstante, las redes neuronales demostraron un mejor desempeño en esta aplicación [12]. Por otro lado, el uso de estas tecnologías y en su aplicación directa en plantas de energías renovables [13] [14] es fundamental para el

aprovechamiento de estos recursos naturales, y su correcta explotación para mejorar la eficiencia energética, así mismo, estas técnicas de predicción pueden ser aplicadas al diagnóstico de fallas de los sistemas fotovoltaicos [15], haciendo sistemas más robustos y escalables.

El desarrollo de esta investigación tiene como principal objetivo: Desarrollar un sistema para la selección de módulos fotovoltaicos comercializados en Colombia basado en la comparación de datos ambientales e información técnica proporcionada por los fabricantes. Para ello, se dividió el proyecto en 3 objetivos específicos los cuales son:

1. Seleccionar estrategias adecuadas para la extracción de información de las páginas oficiales de los fabricantes de módulos fotovoltaicos.
2. Desarrollar un modelo de selección que contraste la información normalizada de los fabricantes teniendo en cuenta las variables ambientales del caribe colombiano.
3. Desplegar una aplicación web que permita consultar recomendaciones sobre los módulos fotovoltaicos comercializados en Colombia.

A su vez, cada uno de estos objetivos se dividieron en cuatro fases importantes para el proyecto y se ven representadas de forma general por el siguiente diagrama (figura 1.1).

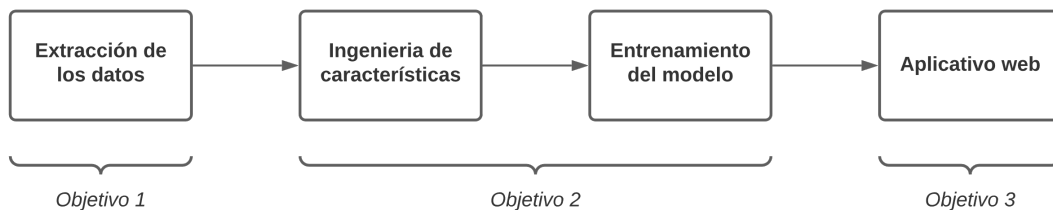


Figura 1.1: Diagrama de etapas del proyecto

En el capítulo 2 se realiza la extracción de los datos de módulos fotovoltaicos comercializados en Colombia, para esto, se hace uso de métodos y técnicas de web scraping para automatizar el proceso de la extracción. Además, se hace uso de herramientas de visualización de datos para conocer la distribución y comportamiento general de los datos obtenidos.

En el capítulo 3 se aplican técnicas de ingeniería de características para obtener nuevos datos por medio de modelos matemáticos, y posteriormente se usa un algoritmo de aprendizaje no supervisado para clasificar los resultados dentro de diferentes grupos según sean sus características.

En el capítulo 4 se hace uso de los grupos de datos obtenidos en el capítulo 3 para entrenar un modelo de aprendizaje supervisado, basado en diferentes algoritmos lineales de clasificación. Una vez entrenados se realiza una breve comparación de los resultados obtenidos y se selecciona el modelo con mejor rendimiento. Los modelos son entrenados con el fin de clasificar datos de temperatura. Irradiación y potencia del módulo dentro de uno de los grupos.

En el capítulo 5 se desarrolla una interfaz web que puede ser utilizada por los usuarios para interactuar con el modelo entrenado y obtener como resultado los 5 módulos solares que mejor se adapten a las condiciones de temperatura, irradiación y potencia entregadas, por otro lado, el usuario puede observar las características en condiciones STC de los módulos resultantes. Cabe resaltar que el enfoque de esta investigación está basada en datos del Magdalena, Colombia.

Finalmente, se obtuvo un conjunto de datos con más de 275 mil datos basado en la generación de los módulos fotovoltaicos según la irradiación y temperatura promedios del Magdalena durante 15 años, que fue usado para el entrenamiento de un modelo de machine learning que arrojó un 90.18% en la precisión al momento de clasificar, y que fue utilizado base para el aplicativo web propuesta en esta investigación.

Referencias

- [1] Sandi Knez, Goran Šimić, Anica Milovanović, Sofia Starikova, and Franc Željko Županič. Prices of conventional and renewable energy as determinants of sustainable and secure energy development: regression model analysis. *Energy, Sustainability and Society*, 12(1):1–12, 2022. 8
- [2] Mark Durrenberger. Stc vs. ptc: Why solar panel testing matters. <https://acortar.link/WnuNoY>, December 2015. (Accessed on 10/29/2022). 8
- [3] Toly Chen and Yu-Cheng Wang. Hybrid big data analytics and industry 4.0 approach to projecting cycle time ranges. *The International Journal of Advanced Manufacturing Technology*, 120(1):279–295, 2022. 9
- [4] Michael Völske, Janek Bevendorff, Johannes Kiesel, Benno Stein, Maik Fröbe, Matthias Hagen, and Martin Potthast. Web archive analytics. *arXiv preprint arXiv:2107.00893*, 2021. 9
- [5] Byron Vladimir Montúfar Chiquito and Ivonne Paulina Mullo Rodríguez. Creación de una aplicación web para la extracción de datos de páginas web acerca de procesos de la matriculación vehicular utilizando la técnica de web scraping. B.S. thesis, 2021. 9
- [6] David Mathew Thomas and Sandeep Mathur. Data analysis by web scraping using python. In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 450–454. IEEE, 2019. 9

- [7] Rachel Krohn and Tim Weninger. Library adoption in public software repositories. *Journal of Big Data*, 6(1):1–19, 2019. 9
- [8] Dazhi Yang, Dennis van der Meer, and Joakim Munkhammar. Probabilistic solar forecasting benchmarks on a standardized dataset at folsom, california. *Solar Energy*, 206:628–639, 2020. 9
- [9] Shubhra Kanti Karmaker, Md Mahadi Hassan, Micah J Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. Automl to date and beyond: Challenges and opportunities. *ACM Computing Surveys (CSUR)*, 54(8):1–36, 2021. 9
- [10] Sahar Zare, Zahra Meidani, Maryam Ouhadian, Hosein Akbari, Farid Zand, Esmaeil Fakharian, and Roxana Sharifian. Identification of data elements for blood gas analysis dataset: a base for developing registries and artificial intelligence-based systems. *BMC Health Services Research*, 22(1):1–20, 2022. 9
- [11] Souhaila Chahboun and Mohamed Maaroufi. Principal component analysis and machine learning approaches for photovoltaic power prediction: a comparative study. *Applied Sciences*, 11(17):7943, 2021. 9
- [12] Gabriel de Freitas Viscondi and Solange N Alves-Souza. A systematic literature review on big data for solar photovoltaic electricity generation forecasting. *Sustainable Energy Technologies and Assessments*, 31:54–63, 2019. 9
- [13] Huaizhi Wang, Yangyang Liu, Bin Zhou, Canbing Li, Guangzhong Cao, Nikolai Voropai, and Evgeny Barakhtenko. Taxonomy research of artificial intelligence for deterministic solar power forecasting. *Energy Conversion and Management*, 214:112909, 2020. 9
- [14] V Prema, Mahajan Sagar Bhaskar, Dhafer Almakhlles, N Gowtham, and K Uma Rao. Critical review of data, models and performance metrics for wind and solar power forecast. *IEEE Access*, 2021. 9

- [15] Sravankumar Jogunuri and FT Josh. Artificial intelligence methods for solar forecasting for optimum sizing of pv systems: A review. *Res. J. Chem. Environ*, 24:174–180, 2020. 10

Capítulo 2

WEB SCRAPING

2.1. Extracción de la información de los módulos fotovoltaicos

Para implementar técnicas que permitan extraer información de internet simulando el comportamiento humano en la navegación por la World Wide Web se deben tener en cuenta diferentes aspectos en los métodos de extracción de los datos como: la estructura de la web, la organización de la información dentro de la página y la variabilidad en la ubicación de los datos en diferentes sitios web. Existen distintas librerías y frameworks para el uso de técnicas y métodos de web scraping que permiten adaptarse a este tipo de escenarios, siendo el [framework de scrapy](#) uno de los ampliamente usados para la extracción de big data gracias a su efectividad en arquitectura y flujo de datos, que adicionalmente permite su integración con otras tecnologías.

Los métodos de web scraping ofrecen una amplia gama de aplicaciones en la solución de distintos problemas en diferentes áreas de estudio, ofreciendo un marco común en la extracción de datos para entrenar modelos de aprendizaje, estudios y modelos estadísticos, estudios de mercados, benchmarking y optimización de precios,

procesos de investigación, así como también es usado para determinar el comportamiento humano en la web, entre otros.

A lo largo de los años, la información obtenida a través de estas técnicas han sido importantes en diversos estudios e implementaciones, las cuales han arrojado interesantes resultados. En [1] se utilizan técnicas de web scraping para extraer titulares y declaraciones de la red social Reddit con el fin de entrenar modelos de aprendizaje de machine learning que identifiquen el sentimiento que expresa la noticia, en [2] se aplica para crear un buscador de artículos científicos basado en palabras clave, dicho buscador realiza una navegación a través de diferentes sitios web e identifica coincidencia en los títulos de artículos, finalmente se extrae información como título, autores, referencias y es almacenada en una base de datos que puede ser descargada en diferentes formatos como JSON y CSV. Por otro lado, en [3] los autores usan estas técnicas para encontrar relación entre diferentes páginas web y cómo estas se conectan por medio de hipervínculos. Es posible saber cuántos de estos te llevan a los mismos sitios web, encontrando así una relación directa entre ellos. El uso de estas técnicas, como se mencionó anteriormente, tiene una gran aplicabilidad en los estudios de mercado en los que se busca conocer el estado de hoteles [4] y la ubicación de páginas web de empresas [5], entre otras cosas. Asimismo, En [6] comprueban patrones de diseño en páginas web, lo que facilita hacer minería de datos a sitios que tienen estructuras HTML y CSS muy diferente, sin embargo, en el documento se observa que a pesar de estas diferencias los patrones tienden a ser comunes en la ubicación título y el cuerpo principal del sitio web. Por otro lado, en el ámbito científico estas técnicas también usadas para la obtención de datos en algunos campos como las ciencias sociales y la producción de información [7], en el área de la psicología para entender el comportamiento humano [8] e incluso para conocer los requerimientos actuales para trabajar en áreas como la informática biométrica [9].

En el desarrollo de un proyecto con web scraping se deben tener en cuenta algunos

aspectos legales y éticos sobre la información que se extrae, es decir, no se debe hacer scraping a imágenes y/o videos que contengan derechos de autor o datos personales de los usuarios como nombres, ID, direcciones, etc. Respetar los tiempos entre una solicitud y otra, no sobrecargar la página de peticiones, realizar estos procedimientos durante las horas de menor tráfico del sitio y restringir el acceso al archivo Robots.txt, el cual indica a los buscadores a que URLs del sitio no deben ingresar.

2.2. Identificación y selección de la fuente de los datos

En el proceso de selección del sitio web que agrupa la información general de los módulos fotovoltaicos comercializados en Colombia, del cual se obtienen los datos usando técnicas y métodos de web scraping, se deben considerar diferentes aspectos como: la distribución de los datos y organización y estructura HTML y CSS de la página web. Durante este proceso se extrajo información del sitio web es.ensolar.com (figura 2.1) el cual es un directorio de empresas y productos de la industria de energía solar; cuenta con una amplia gama de fabricantes a nivel mundial, buena organización y estandarización de los datos, fácil navegación y disponibilidad de la información de cada uno de los productos que se encuentran en stock.



Figura 2.1: Vista principal del sitio es.enfsolar.com

Una vez seleccionada la página web, se agruparon los fabricantes teniendo en cuenta dos criterios simples: que sus productos se comercializaran en Colombia (específicamente módulos solares) y que hicieran parte del directorio seleccionado. Lo que dejó como resultado un total de 23 fabricantes (figura 2.2).

LISTA DE FABRICANTES
AE Solar
Amerisolar
CSI Solar
CSUN
GCLSI
Hyundai Energy Solutions
JA Solar
Jinko Solar
Kyocera
LG Electronics
LONGi Solar
Luxen Solar Energy
Osda Solar
ReneSola
Restar Solar Energy
Risen Energy
Sharp
Solartech Power
Suntech Power
Trina Solar
Yingli Green Energy
Znshine Solar

Figura 2.2: Lista de fabricantes seleccionados

El sitio web proporciona una organización sistemática de los productos ofertados por cada fabricante, estandarizando la ubicación de cada modelo de módulo fotovoltaico (figura 2.3). Adicionalmente, cada uno de los enlaces redirecciona a la información técnica proporcionada por el fabricante, esto evita descargar la hoja de datos para acceder a la información bajo condiciones Standard Test Conditions

(STC), Nominal operating Cell Temperature (NOCT) y sus características térmicas e información general como nombre, modelo y potencia (figura 2.4 y 2.5)

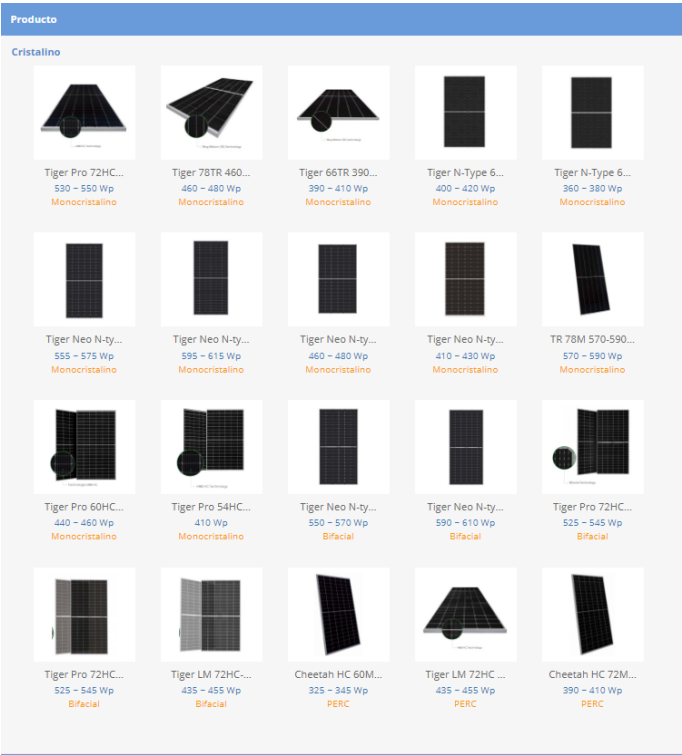


Figura 2.3: Módulos solares del fabricante Jinko Solar

No. de Modelo	JKM530M-72HL4/JKM530M-72HL4-V	JKM535M-72HL4/JKM535M-72HL4-V	JKM540M-72HL4/JKM540M-72HL4-V	JKM545M-72HL4/JKM545M-72HL4-V	JKM550M-72HL4/JKM550M-72HL4-V
Garantía					
Garantía de Producto	12 Años				
Garantía de energía	25 años de 84,8% potencia de salida				
Características Eléctricas en STC					
Potencia Máxima (Pmax)	530 Wp	535 Wp	540 Wp	545 Wp	550 Wp
Tensión en el Punto de Máxima Potencia (Vmax)	40,56 V	40,63 V	40,7 V	40,8 V	40,9 V
Corriente en el punto de máxima potencia (Imax)	13,07 A	13,17 A	13,27 A	13,36 A	13,45 A
Tensión en Circuito Abierto (Voc)	49,26 V	49,34 V	49,42 V	49,52 V	49,62 V
Corriente de cortocircuito (Isc)	13,71 A	13,79 A	13,85 A	13,94 A	14,03 A
Eficiencia	20,55 %	20,75 %	20,94 %	21,13 %	21,33 %
Tolerancia de Potencia (+)	+ 3 %	+ 3 %	+ 3 %	+ 3 %	+ 3 %
Condiciones de prueba estándar (STC): Masa de aire AM 1,5, radiación 1000W/m², temperatura de célula 25°C.					

Figura 2.4: Información STC del módulo FV Tiger Pro 72HC 530-550

Características Eléctricas en NOCT					
Potencia Máxima (Pmax)	394 Wp	398 Wp	402 Wp	405 Wp	409 Wp
Tensión en el Punto de Máxima Potencia (Vmax)	37,84 V	37,91 V	38,08 V	38,25 V	38,42 V
Corriente en el punto de máxima potencia (Imax)	10,42 A	10,5 A	10,55 A	10,6 A	10,65 A
Tensión en Circuito Abierto (Voc)	46,5 V	46,57 V	46,65 V	46,74 V	46,84 V
Corriente de cortocircuito (Isc)	11,07 A	11,14 A	11,19 A	11,26 A	11,33 A
Temperatura	45±2 °C				
	Temperatura en condiciones normales de operación (NOCT): 800W/m², AM 1,5, velocidad del viento de 1m/s, temperatura ambiente de 20°C				
Características Térmicas					
Rango de Temperatura	-40–85 °C				
Coefficiente de Temperatura de Pmax	-0,35 %/°C				
Coefficiente de Temperatura de Voc	-0,28 %/°C				
Coefficiente de Temperatura de Isc	0,048 %/°C				

Figura 2.5: Información NOCT y térmica del módulo FV Tiger Pro 72HC 530-550

2.3. Extracción de los datos utilizando técnicas de Web Scraping

Existen diversas librerías y frameworks para la implementación de web scraping, [Beautiful Soup](#), [Selenium](#) y [Scrapy framework](#) son, por ejemplo, algunos de los más famosos. Están desarrollados y soportados para el lenguaje de programación de alto nivel [Python](#), quien tiene una alta difusión, una gran comunidad, excelente documentación y ofrece una gran versatilidad en casi cualquier área de desarrollo. Esto permite la creación de programas sencillos y complejos para la extracción de datos de la web. En este trabajo de investigación se usó el [framework de scrapy](#), este cuenta con funcionalidades y ventajas sobre otras tecnologías previamente mencionadas, algunas de estas ventajas son: la ejecución asíncrona de sus procesos, múltiples solicitudes al mismo tiempo y la posibilidad de delimitar acciones. Adicionalmente, esta herramienta permite la implementación de un CrawlSpider, programa capaz de extraer enlaces de la página web con el fin de entrar a cada uno de ellos de forma recursiva e incluso de forma infinita (figura 2.6).

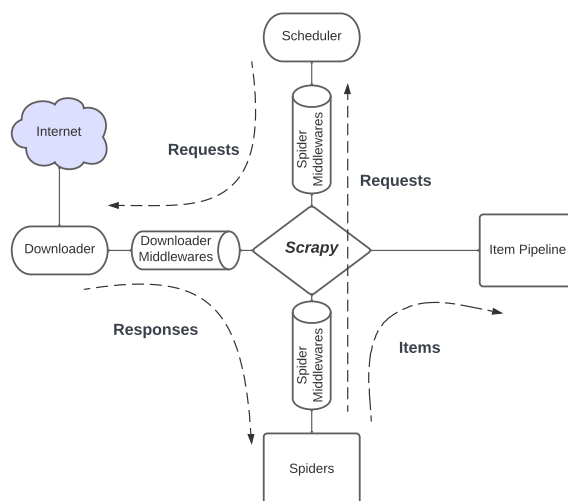


Figura 2.6: Diagrama de funcionamiento de Scrapy Framework

El funcionamiento de ENF_scraper está basado en un CrawlSpider de scrapy framework, lo que hace posible el uso de expresiones Xpath para extraer los enlaces de cada uno de los módulos fotovoltaicos de cada fabricante previamente seleccionado, obteniendo internamente una lista de enlaces (figura 2.7).

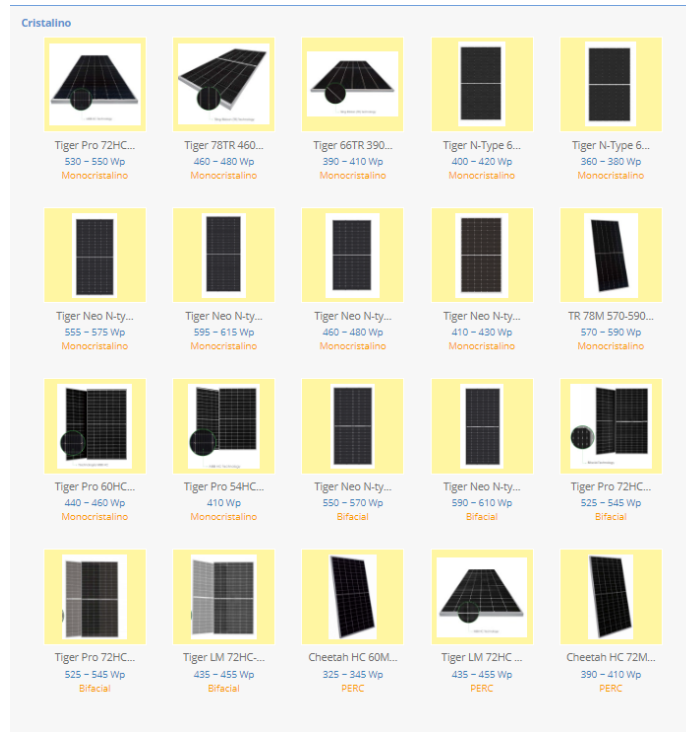


Figura 2.7: Secciones seleccionadas con la expresión xpath correspondiente

El spider de scrapy accede a cada enlace para extraer la información técnica del módulo fotovoltaico, para ello, hace uso nuevamente de expresiones Xpath que corresponden a la información en condiciones Standard Test Conditions (STC), Nominal operating Cell Temperature (NOCT), sus características térmicas e información general como nombre, modelo y potencia contenidas dentro del HTML o CSS (figura 2.8 y Tabla 2.1). La información extraída a partir de la expresión Xpath tiene el formato de una lista de Python la cual, además, contiene información “basura” que no se desea como resultado. Una vez la información es obtenida por el Spider se realiza

un proceso de estandarización encargado de separar, limpiar, y organizar los datos de forma estructurada. Al mismo tiempo que esto sucede, los datos estructurados son enviados a los pipelines encargados de almacenarlos en un archivo Excel.

No. de Modelo	JKM530M-72HL4/JKM530M-72HL4-V	JKM535M-72HL4/JKM535M-72HL4-V	JKM540M-72HL4/JKM540M-72HL4-V	JKM545M-72HL4/JKM545M-72HL4-V	JKM550M-72HL4/JKM550M-72HL4-V
Garantía					
Garantía de Producto	12 Años				
Garantía de energía	25 años de 84,8% potencia de salida				
Características Eléctricas en STC					
Potencia Máxima (Pmax)	530 Wp	535 Wp	540 Wp	545 Wp	550 Wp
Tensión en el Punto de Máxima Potencia (Vmax)	40,56 V	40,63 V	40,7 V	40,8 V	40,9 V
Corriente en el punto de máxima potencia (Imax)	13,07 A	13,17 A	13,27 A	13,36 A	13,45 A
Tensión en Circuito Abierto (Voc)	49,26 V	49,34 V	49,42 V	49,52 V	49,62 V
Corriente de cortocircuito (Isc)	13,71 A	13,79 A	13,85 A	13,94 A	14,03 A
Eficiencia	20,55 %	20,75 %	20,94 %	21,13 %	21,33 %
Tolerancia de Potencia (+)	+ 3 %	+ 3 %	+ 3 %	+ 3 %	+ 3 %
	Condiciones de prueba estándar (STC): Masa de aire AM 1,5, radiación 1000W/m², temperatura de célula 25°C.				
Características Eléctricas en NOCT					
Potencia Máxima (Pmax)	394 Wp	398 Wp	402 Wp	405 Wp	409 Wp
Tensión en el Punto de Máxima Potencia (Vmax)	37,84 V	37,91 V	38,08 V	38,25 V	38,42 V
Corriente en el punto de máxima potencia (Imax)	10,42 A	10,5 A	10,55 A	10,6 A	10,65 A

Figura 2.8: Secciones seleccionadas con la expresión xpath correspondiente

Basic information	Standard test conditions	Nominal operating cell temperature	Thermal rating
Company name	Maximun power	Maximun power	Operating temperature range
PV module name	Voltage at maximun power	Voltage at maximun power	Temperature coefficient of Pmax
PV module model	Current at maximun power	Current at maximun power	Temperature coefficient of Voc
Tecnology	Open circuit voltage	Open circuit voltage	Temperature coefficien of Isc
	Short circuit current	Short circuit current	
	PV module efficiency	Temperature	
	Power tolerance		

Tabla 2.1: Datos extraídos del sitio Web es.ensolar.com

El funcionamiento de ENF_scraper tiene algunas restricciones programadas que

vale la pena resaltar:

1. Hará solo 2 peticiones como máximo al mismo tiempo.
2. Hará peticiones cada 10 segundos.
3. Cada petición se realizará con un User-Agent diferente.
4. Todos los datos serán almacenados en un archivo de excel con el nombre “enf-solar_datasheet.xlsx”.

Las tres primeras se toman como medida preventiva para evitar la detección del bot por parte de la página, esto debido a la falta de proxies. Si el proceso se realiza de forma rápida y sin usar proxies para encubrir la actividad, es posible que la página bloquee de forma permanente la IP desde la cual se ejecuta el programa, tal como sucedió en el desarrollo de esta sección (figura 2.9).

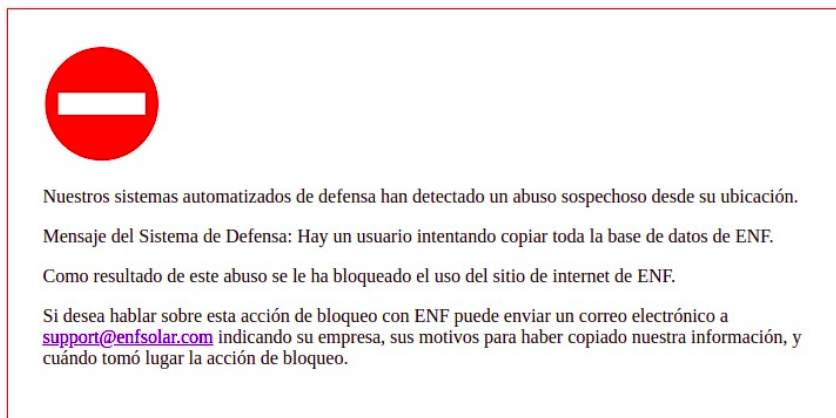


Figura 2.9: Cartel de aviso de bloqueo del sitio Web es.enfsolar.com

El desarrollo de esta sección se realizó exclusivamente con fines académicos y de investigación. Los procedimientos realizados se encuentra de forma detallada en el repositorio de [GitLab](#) en el cual se explica paso a paso cómo funciona ENF_scraper (*ENF_scraper: repositorio en línea*).

2.4. Resultados

Para obtener los datos de módulos fotovoltaico de los fabricantes que comercializan sus productos en Colombia, se ejecutó ENF_scraper. Tardó aproximadamente dos horas en realizar el proceso debido a las restricciones previamente mencionadas, conexión a internet y limitaciones del hardware. Como resultado se obtuvo un archivo de Excel con 2.191 módulos solares y sus características técnicas. Una vez obtenidos los datos se encuentran listos para realizar un análisis estadístico de su composición con el fin de entender cómo deben ser normalizados los datos posteriormente.

Se hizo uso de la librería [pandas](#) para la manipulación de los datos y [Power BI](#) para la representación gráfica de los mismos. Dentro de los datos obtenidos se observa que los fabricantes Znshine Solar, AE Solar y Yingli Green Energy ocupan los primeros puestos con mayor cantidad de módulos solares en el mercado local (figura 2.10), también se encuentran diversas tecnologías como, monocristalino, policristalino, bifacial, entre otras (figuras 2.11, 2.12, 2.13, 2.14, 2.15, 2.16). Los módulos con tecnología monocristalina son los de mayor presencia (figura 2.17), además, se encuentran en un rango de potencia desde los 100 W hasta los 675 W (figura 2.18). Por otro lado, se observa que el 100 % de los fabricantes aplican pruebas STC a los módulos fotovoltaicos, mientras que solo el 78 % de estos cuenta con información sobre pruebas NOCT (Tabla 2.2).

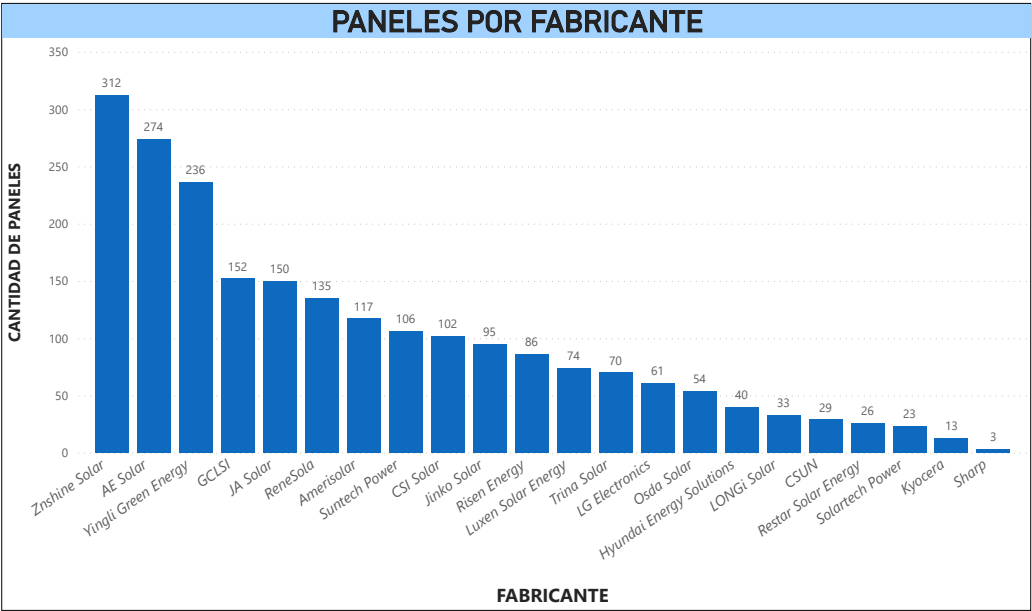


Figura 2.10: Cantidad de módulos fotovoltaicos por fabricantes

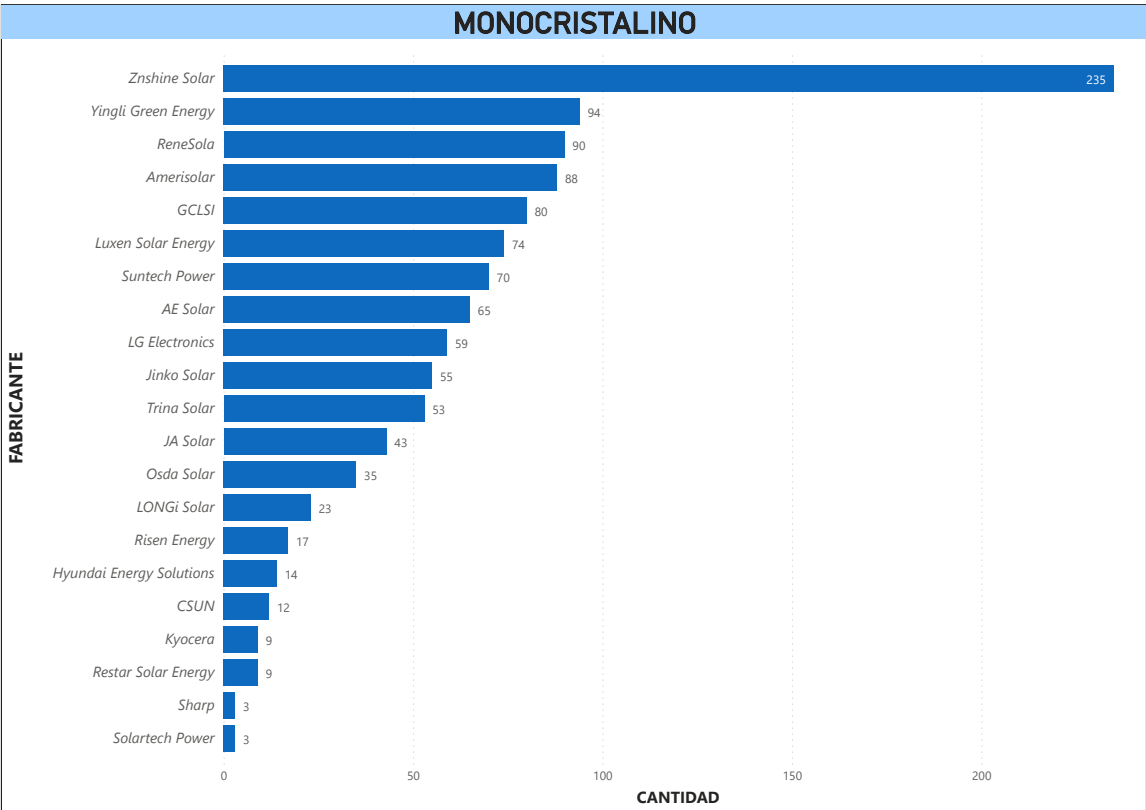


Figura 2.11: Cantidad de módulos fotovoltaicos monocristalinos por fabricante

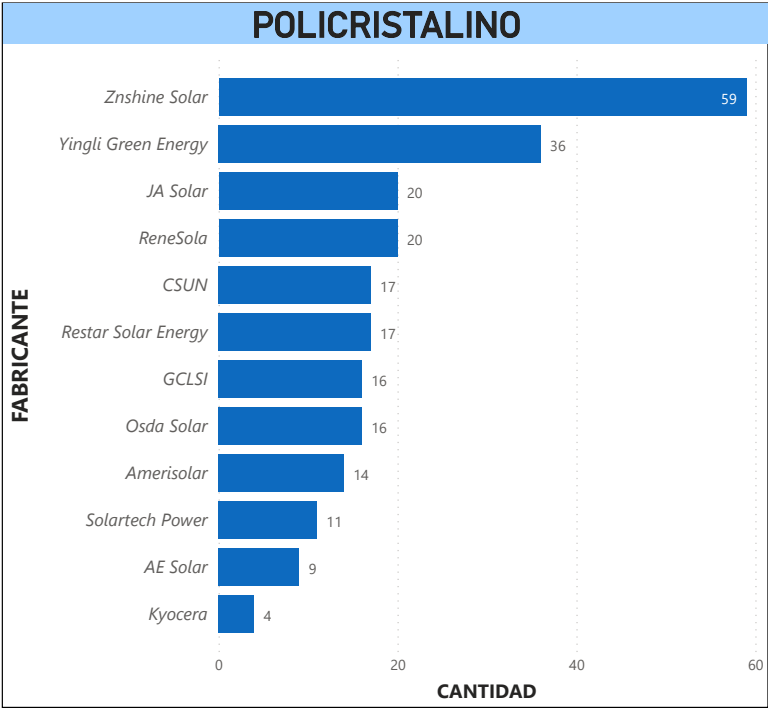


Figura 2.12: Cantidad de módulos fotovoltaicos policristalinos por fabricante

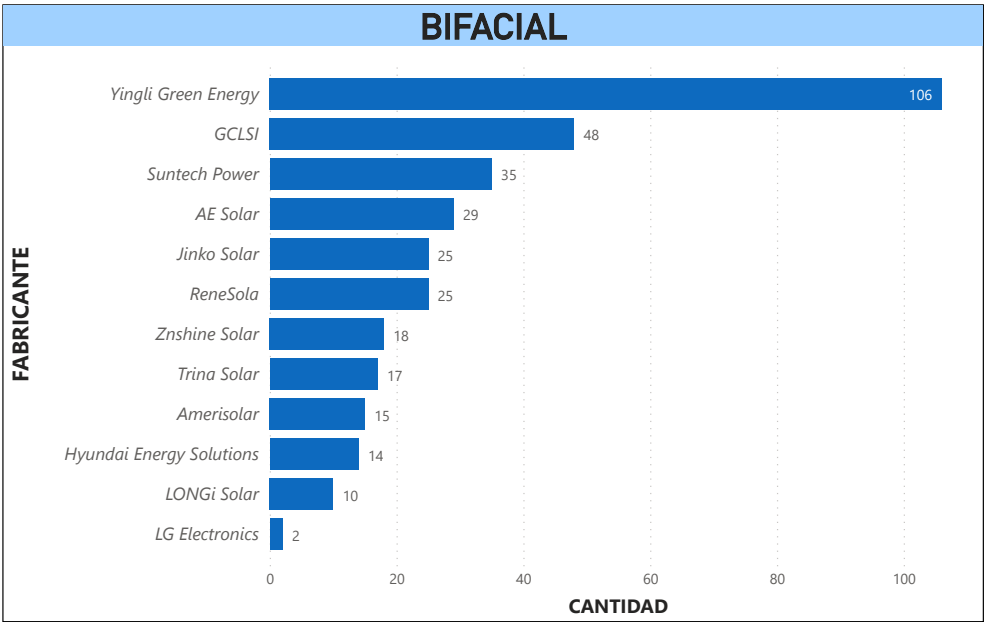


Figura 2.13: Cantidad de módulos fotovoltaicos Bifacial por fabricante

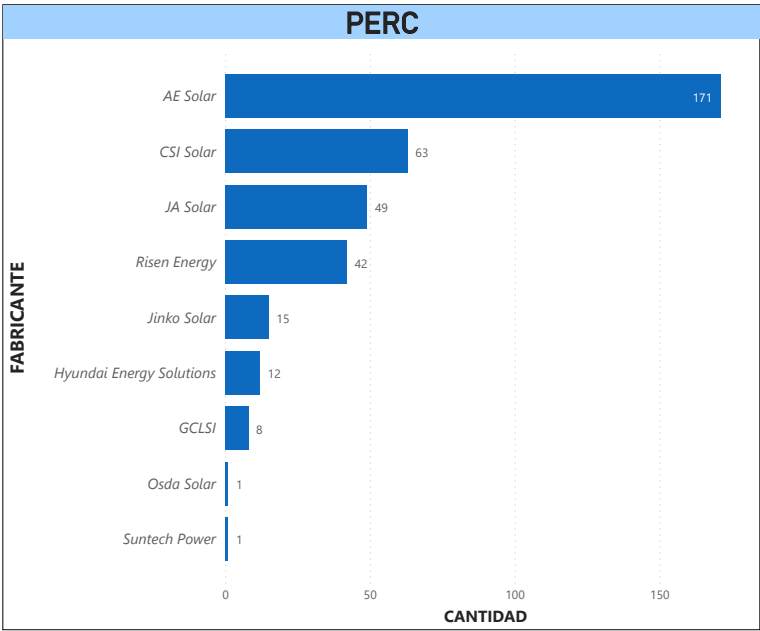


Figura 2.14: Cantidad de módulos fotovoltaicos Perc por fabricante

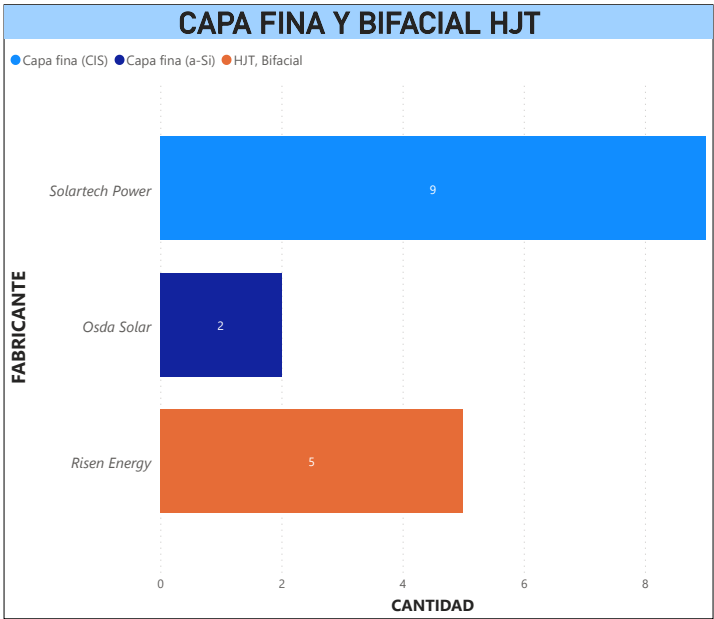


Figura 2.15: Cantidad de módulos fotovoltaicos capa fina y bifacial hjt por fabricante

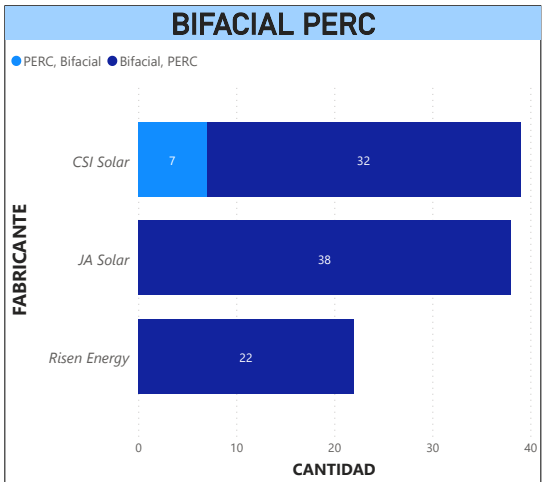


Figura 2.16: Cantidad de módulos fotovoltaicos capa bifacial Perc por fabricante

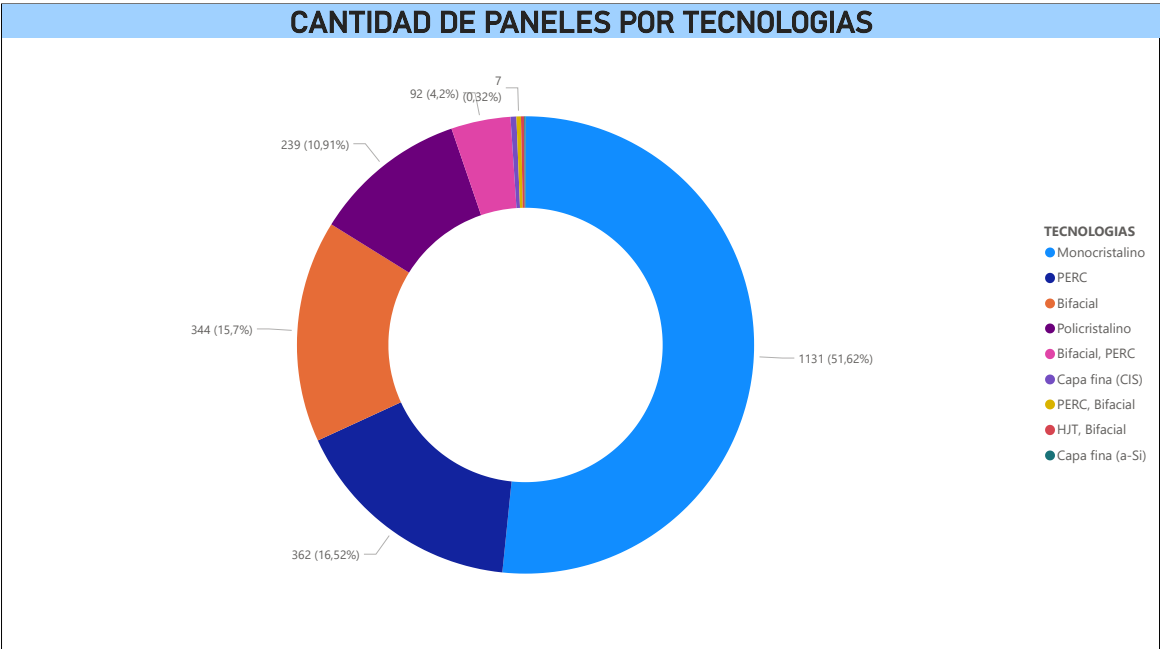


Figura 2.17: Cantidad de módulos fotovoltaicos por tecnología

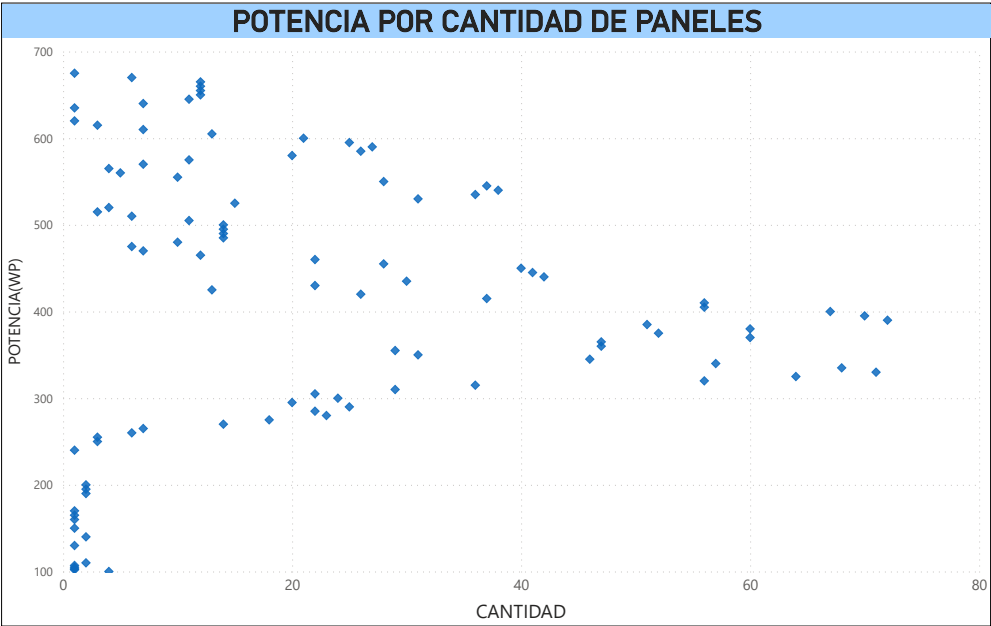


Figura 2.18: Distribución de las potencias de los módulos fotovoltaicos

Tipo de prueba	Total	Total %
Standard test conditions	2191	100.000
Nominal operating cell temperature	1709	78.001

Tabla 2.2: Pruebas realizadas a los módulos fotovoltaicos

2.5. Conclusiones

La extracción de datos de la web puede llegar a ser compleja según sea la fuente de la cual son obtenidos; sin embargo, estos procesos se pueden llevar a cabo de forma sistemática y ordenada con la aplicación de técnicas y métodos de web scraping y haciendo uso de un framework que permitan su aplicación sobre lenguajes de programación como Python, el cual fue usado en el desarrollo de esta investiga-

ción. Asimismo, la integración de estas tecnologías con herramientas de conversión de formato ha facilitado la creación de datasets con información sobre módulos fotovoltaicos comercializados en Colombia.

El desarrollo de diferentes procedimientos y técnicas aplicadas en este estudio ha permitido disponer de herramientas para el preprocesamiento y representación de los datos obtenidos, que como resultado más importante facilita el entendimiento y la distribución de los datos.

De esta forma se pueden identificar los datos de módulos fotovoltaicos comercializados en Colombia con potencial para la implementación en proyectos de análisis de datos, machine learning, redes neuronales, entre otros, con un principal interés en la aplicación para investigación en el mejoramiento de la generación de energía eléctrica renovable en la zona caribe colombiana.

Referencias

- [1] Parneet Kaur. Sentiment analysis using web scraping for live news data with machine learning algorithms. *Materials Today: Proceedings*, 65:3333 – 3341, 2022. 16
- [2] Wahaj Salem Alkaberi, Reem Hamed Aljuhani, and Huda Mohamed Alamoudi. Web scraper application for extracting scientific journals data. page 220 – 224, 2021. 16
- [3] Jesse S. Sayles, Ryan P. Furey, and Marilyn R. ten Brink. How deep to dig: effects of web-scraping search depth on hyperlink network analysis of environmental stewardship organizations. *Applied Network Science*, 7(1), 2022. Cited by: 0; All Open Access, Gold Open Access. 16
- [4] Saram Han and Christopher K Anderson. Web scraping for hospitality research: Overview, opportunities, and implications. *Cornell Hospitality Quarterly*, 62(1):89–104, 2021. 16
- [5] Anita Thonipara, Rolf Sternberg, Till Proeger, and Lukas Haefner. Digital divide, craft firms’ websites and urban-rural disparities—empirical evidence from a web-scraping approach. *Review of Regional Research*, pages 1–31, 2022. 16
- [6] Hamza Salem and Manuel Mazzara. Pattern matching-based scraping of news websites. volume 1694, 2020. Cited by: 4; All Open Access, Bronze Open Access. 16

- [7] Alex Luscombe, Kevin Dick, and Kevin Walby. Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*, 56(3):1023–1044, 2022. 16
- [8] Felix Speckmann. Web scraping: A useful tool to broaden and extend psychological research. *Zeitschrift für Psychologie*, 229(4):241, 2021. 16
- [9] Jürgen Schedlbauer, Georgios Raptis, and Bernd Ludwig. Medical informatics labor market analysis using web crawling, web scraping, and text mining. *International Journal of Medical Informatics*, 150:104453, 2021. 16

Capítulo 3

INGENIERÍA DE CARACTERÍSTICAS

3.1. Procesamiento y caracterización de los datos

Para entrenar modelos de aprendizaje usando algoritmos de machine learning o deep learning es fundamental que los datos se encuentren caracterizados.

Comúnmente los datos en la web se extraen en bruto, es decir, a simple vista no hay ninguna relación entre ellos, no comparten una escala o formato e incluso podrían no tener ninguna característica que los agrupe. Por esta razón la ingeniería de características es un paso fundamental antes de entrenar cualquier modelo de aprendizaje, esta se encarga de extraer y transformar los datos sin procesar y caracterizarlos de tal forma que puedan ser usados para entrenar modelos. Existen diversos pasos a seguir cuando se desea caracterizar datos y generalmente estos dependen de la forma en la que se encuentra la información, ya sea texto, imágenes, tablas u otros formatos. Por otro lado, también influye para qué serán usados dichos datos, ya sea para entrenar modelos de clasificación, regresiones lineales, regresiones temporales, análisis estadísticos u otros. Es decir, todo el proceso se realiza con el fin

de preparar los datos para entrenar uno o más modelos según sea el requerimiento.

Aunque los pasos a realizar no son siempre los mismos, existen diversas tecnologías y lenguajes que ayudan su realización, es claro que el uso de las herramientas que ofrecen los lenguajes de programación es una de las formas más óptimas de llevar a cabo estos procesos de vital importancia en la caracterización de la información.

Diferentes estudios se han adelantado para establecer cómo las técnicas de ingeniería de características afectan a los modelos inteligentes. Algunos de estos estudios han encontrado que, a pesar de disponer del mismo dataset de datos, según sean caracterizados, algunos modelos de clasificación ven afectada su eficiencia en la predicción y se puede ver representado como una mejoría, o por el contrario, se puede ver representado con una precisión mucho menor a la esperada [1]. De igual forma la importancia de elegir la manera en la que se preparan los datos varía mucho según el método analítico, los objetivos para los que se requieren y las técnicas de aprendizaje y software específicos por el cual deben ser analizados [2], por lo tanto, influye en que el experto artificial se comporte tal como lo harían los expertos naturales, desde cuáles son las características que tiene en cuenta hasta cuál sería el resultado de un diagnóstico [3].

Es claro que el uso de la ingeniería de características tiene un importante papel en cualquier proceso en que se desee entrenar un modelo inteligente, y sus aplicaciones son, por ejemplo, el procesamiento de datos del historial de crédito de personas latinoamericanas para el uso en modelos inteligentes capaces de predecir una puntuación de solvencia económica [4] o la caracterización de datos financieros que alimentan un modelo que predice el precio de las acciones [5]. No obstante, la mayoría de los datos del mundo real son impuros y no siempre están perfectamente organizados [6], lo que quiere decir que su procesamiento no es tan sencillo y por el contrario de lo que se pensaría, estos casos son muy comunes, por lo que existen técnicas de caracterización exclusiva para solventar este tipo de problemas [7]. Del mismo modo, existen retos en la caracterización de grandes volúmenes de datos ya

sea por la forma en como están estructurados o la capacidad de procesamiento del hardware y su disponibilidad de memoria para almacenar tal volumen de información [8]. En la práctica, se ha comprobado que la limpieza y la preparación de los datos suponen aproximadamente el 80 % del esfuerzo total de ingeniería de datos.

3.2. Datos de temperatura e irradiación

La temperatura e irradiación son dos de los factores más importantes al momento de estimar la eficiencia y generación de un módulo fotovoltaico. Muchos modelos matemáticos hacen uso de esta información para calcular una generación o eficiencia del módulo tal como veremos en la siguiente sección. Durante esta sección se hará un enfoque especial sobre la obtención de datos de temperatura e irradiación, además de los pasos a seguir para transformarlos a datos estructurados en un formato específico con el fin de facilitar su uso en secciones futuras.

Existen diferentes sitios web que ofrecen información pública de la temperatura, en Colombia, por ejemplo, existen diferentes estaciones meteorológicas que brindan información ambiental del punto donde estén ubicadas, un claro caso es [IDEAM](#). También existen otros sitios web que brindan esta información como el [Climate Change Knowledge Portal](#) e incluso algunas APIs como [meteomatics](#) también brindan esta información. En el desarrollo de esta investigación se usaron datos obtenidos de World Bank Group en su portal web “[Climate Change Knowledge Portal](#)” (figura 3.1), el cual es una herramienta de uso gratuito que brinda información sobre variables como temperatura y precipitación. La preferencia de esta herramienta se basa en su amplio rango de datos que corresponde a un registro que inicia en el año 1970 y finaliza en 2021 y que contiene información de diferentes regiones del mundo, en este caso se seleccionó la región Magdalena, Colombia. El archivo original se procesó para obtener un registro mes a mes durante un período de 15 años (2005 a 2020) del promedio de la temperatura máxima en $^{\circ}C$ registrada en el departamento del

Magdalena, Colombia.

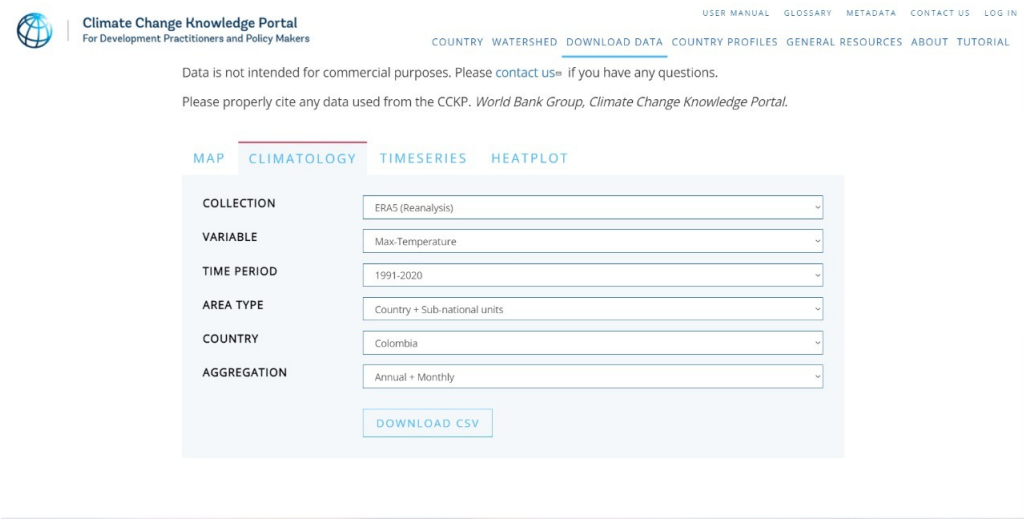


Figura 3.1: Vista principal del Climate Change Knowlege Portal

Al igual que la temperatura, la irradiación fue obtenida por medio de una página web, en este caso los datos fueron entregados por la herramienta web de EU Science Hub llamada [PVGIS tool](#), la cual es de uso gratuito y brinda datos de irradiación solar y rendimiento energético de una ubicación geográfica en específico (figura 3.2). La ubicación seleccionada para el desarrollo de esta investigación fue la Universidad del Magdalena (figura 3.3) en Santa Marta, Colombia, el archivo original descargado correspondió a un archivo JSON con información mes a mes correspondiente a la irradiación normal directa en $kWh/m^2/mes$ durante 15 años (2005 a 2020), este archivo se procesó para cambiar su formato a una tabla. Una vez los datos de temperatura e irradiación fueron adquiridos se unieron en una tabla con la siguiente estructura: (Tabla 3.1)

mes	Año	Irradiación	Temperatura
-----	-----	-------------	-------------

Tabla 3.1: Estructura de la tabla

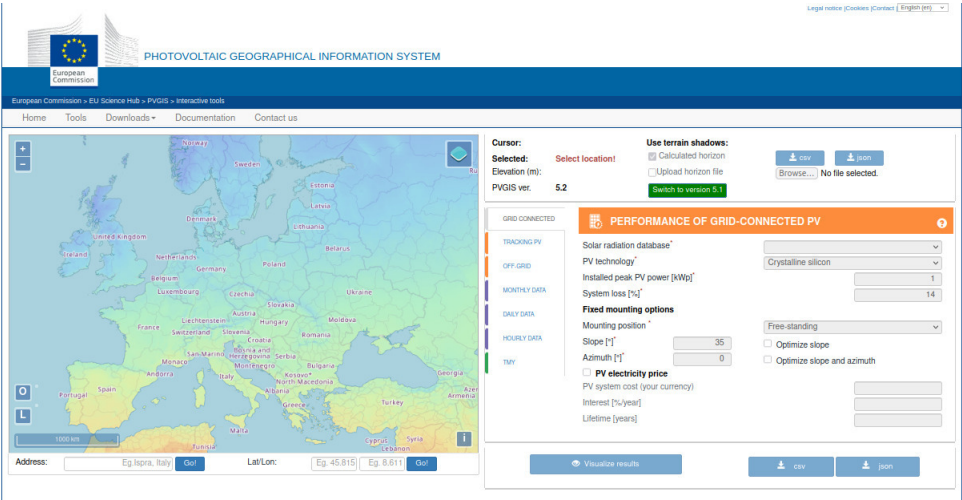


Figura 3.2: Vista principal del PVGIS tool



Figura 3.3: Universidad del Magdalena vista desde el mapa PVGIS tool

3.3. Estimación de la generación de energía

En el proceso de entrenamiento de un modelo de aprendizaje que permita determinar que módulo fotovoltaico se adapta mejor a ciertas condiciones ambientales, es necesario saber como se comportaría un módulo fotovoltaico bajo diferentes condiciones de irradiación y temperatura, específicamente en 15 años (2005 a 2020). Para ello es preciso usar modelos matemáticos que modelen el comportamiento de cada uno de los módulos fotovoltaicos bajo las condiciones de irradiación y temperatura a las que hubieran sido sometidas durante los diferentes meses del año en el rango de tiempo ya especificado.

El modelo matemático usado en esta investigación es el modelo de Ortiz que usa como parámetros principales información general en condiciones STC (Standard test condition) del módulo fotovoltaico, la temperatura, la irradiancia y el factor de ajuste de la curva (b), el cual se calcula por medio de una aproximación. El módulo fotovoltaico se modeló mediante la ecuación (3.1).

$$I(V) = \frac{I_x}{1 - e^{\left(\frac{-1}{b}\right)}} \left[1 - e^{\left(\frac{V}{bV_x} - \frac{1}{b}\right)} \right] \quad (3.1)$$

Donde V_x e I_x corresponden al voltaje en circuito abierto y corriente en corto circuito, y están representadas por la ecuación (3.2) y (3.3).

$$V_x = s \frac{E_i}{E_{iN}} TC_V (T - T_N) + s V_{m\acute{a}x} - s (V_{m\acute{a}x} - V_{m\acute{i}n}) e^{\left(\frac{E_i}{E_{iN}} \ln\left(\frac{V_{m\acute{a}x} - V_{oc}}{V_{m\acute{a}x} - V_{m\acute{i}n}}\right)\right)} \quad (3.2)$$

$$I_x = p \frac{E_i}{E_{iN}} [I_{sc} + TC_i (T - T_N)] \quad (3.3)$$

Además, el factor de ajuste de la curva se calculó para cada uno de los módulos fotovoltaicos mediante la ecuación (3.4)

while $|b_{n+1} - b_n| > \varepsilon$

$$b_{n+1} = \frac{V_{op} - V_{oc}}{V_{oc} \cdot \ln \left[1 - \frac{I_{op}}{I_{sc}} \cdot \left(1 - \exp \left(\frac{-1}{b_n} \right) \right) \right]} \quad (3.4)$$

Para una mayor claridad es necesario saber que, s es el número de módulos FV en serie; p es el número de módulos FV en paralelo; E_i es la irradiación solar; E_{iN} es 1000 W/m^2 ; T es la temperatura de funcionamiento; T_N es $25 \text{ }^\circ\text{C}$; TC_v es el coeficiente de temperatura de la tensión; TC_i es el coeficiente de temperatura de la corriente; V_{oc} es la tensión en circuito abierto; I_{sc} es el coeficiente de temperatura de la corriente; TC_v es el coeficiente de temperatura de la tensión; V_{max} es $103\% V_{oc}$; y V_{min} es $85\% V_{oc}$. [9] [10] [11]

El modelo descrito anteriormente fue codificado en el lenguaje de programación [Python](#), con el cual se creó una clase que representa al módulo fotovoltaico y que además realiza todos los cálculos necesarios para la estimación de la generación de cada uno de los módulos. Una vez codificado, el modelo se usó de manera iterativa sobre cada uno de los módulos fotovoltaico para calcular la generación bajo diferentes condiciones de temperatura e irradiación (figura 3.4), lo que dejó como resultado 180 datos de generación por cada módulo, es decir, una generación por cada mes durante 15 años y en total más de 275.000 datos, los cuales fueron organizados en un archivo de excel usando [Pandas](#). Cabe resaltar que, a algunos datos entregados al modelo se les realizó una operación matemática para obtener un valor equivalente en las unidades deseadas, este es el caso por ejemplo de la irradiación en $kWh/m^2/mes$ el cual se llevó a una irradiancia en w/m^2 , y una vez realizado el cálculo se volvió a unidades de generación de energía eléctrica kWh/mes .

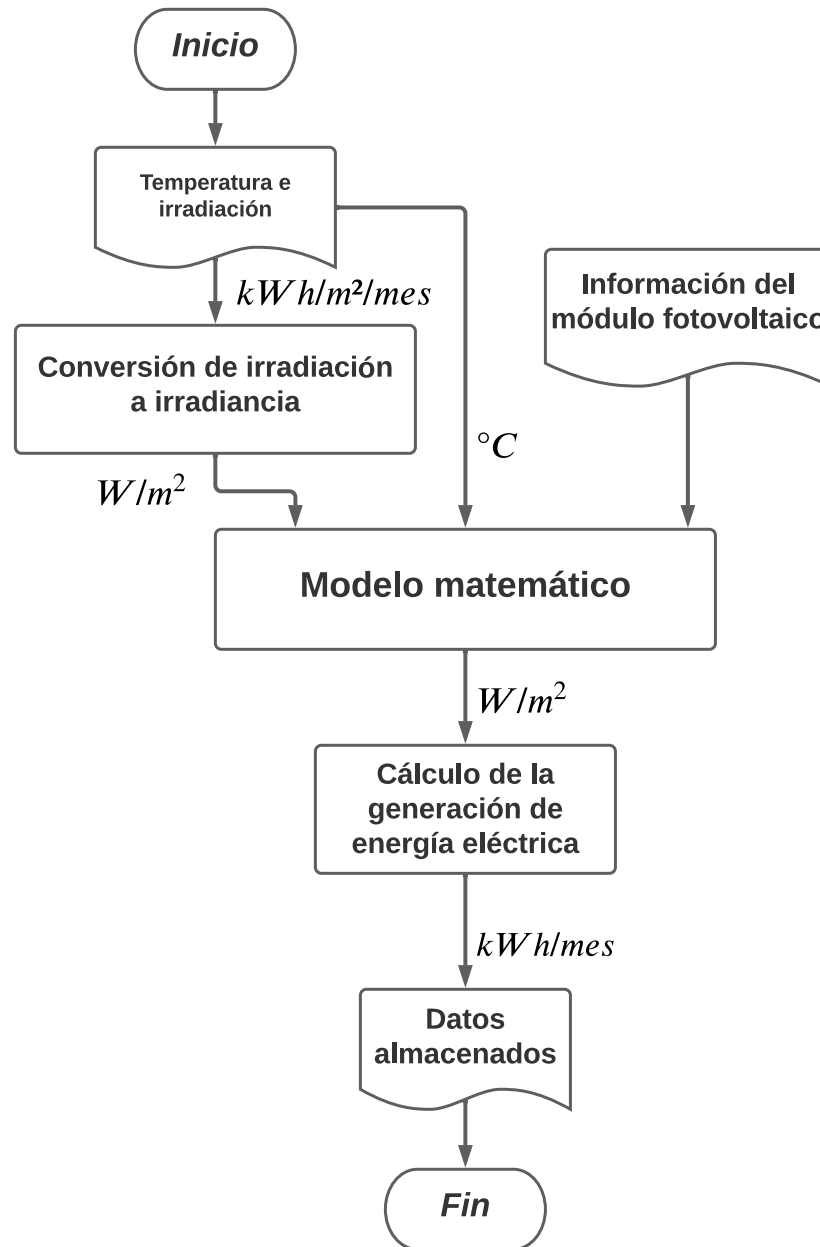


Figura 3.4: Diagrama de flujo del proceso de cálculo de la generación de energía

3.4. Etiquetando los datos

Una forma sencilla de observar el comportamiento de los datos es a través de las gráficas, en este caso se usó un [Pairplot](#) (figura 3.5) para entender como se relacionan algunas características, cuál es su comportamiento e incluso ver cuáles pueden llegar a ser óptimas para un futuro modelo de aprendizaje, sin embargo, la relación entre las características no llega a ser evidente debido a que para un solo módulo fotovoltaico se tiene un total de 180 valores, y para cada par irradiación-temperatura tenemos más de 1500 (la cantidad de módulos que se encuentran en el dataset). Además, los datos no cuentan con una etiqueta que permita clasificarlos en grupos diferentes según sean sus características, y realizar el proceso de etiquetado de forma manual para la cantidad de datos en el dataset resultaría complejo, lo que tomaría una cantidad de tiempo exagerada. En este tipo de casos la mejor opción es optar por herramientas que faciliten y realicen el trabajo de clasificar los datos, este sería un claro ejemplo para usar un algoritmo de clasificación no supervisado que encuentre relación entre los datos y los agrupe, tal como lo hace el algoritmo de agrupamiento [K-means](#) el cual fue implementado en esta investigación haciendo uso de [Python](#) y la librería [Scikit-learn](#).

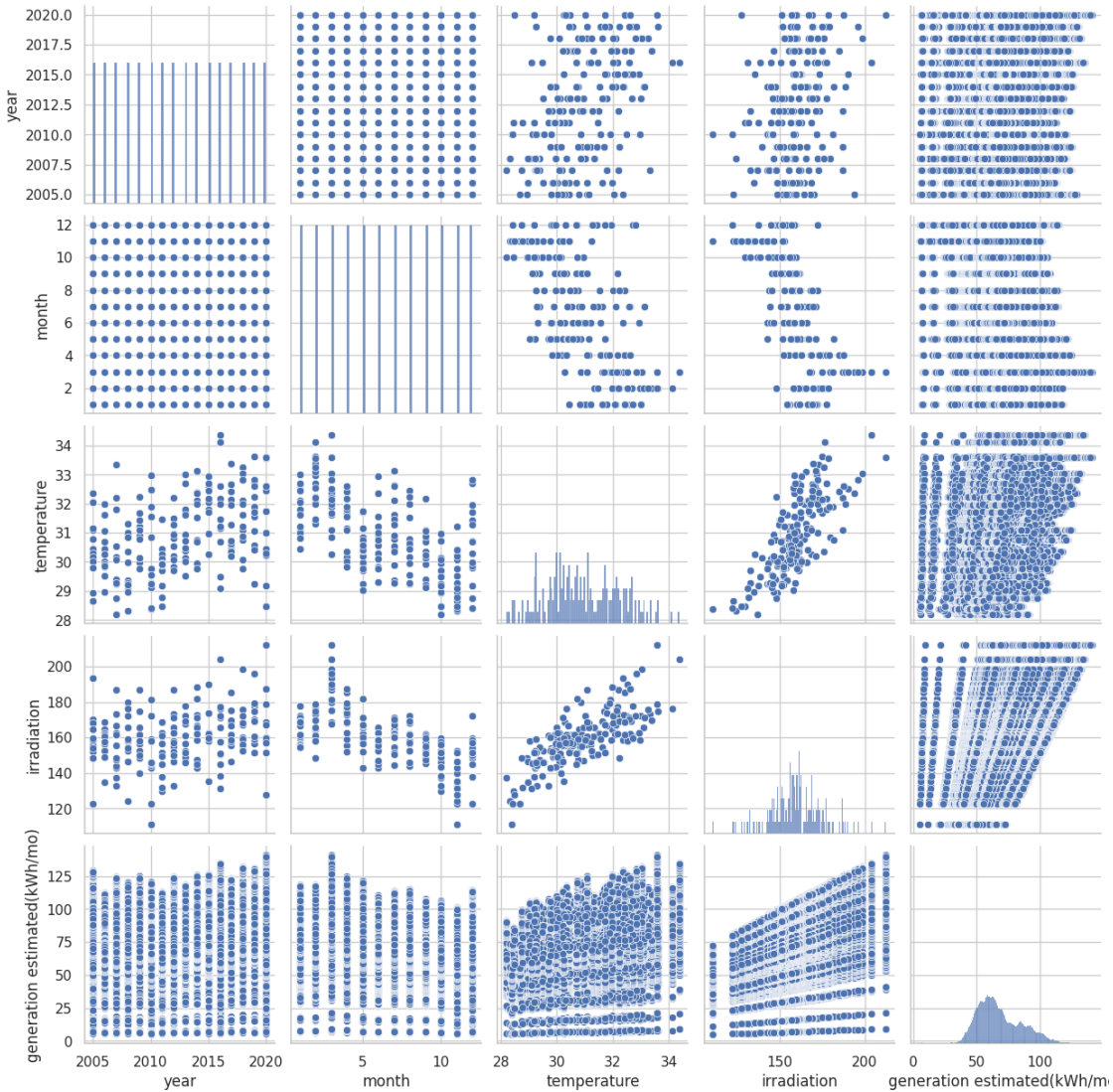


Figura 3.5: Pairplot de las columnas mes, año, temperatura, irradiación y generación

Se seleccionaron las columnas temperatura, irradiación y generación como principales características y posteriormente se hizo uso de la técnica Elbow curve que nos ayuda a determinar que cantidad de grupos es la óptima según nuestros datos (figura 3.6), en este caso el número de grupos seleccionados fue 5. Una vez definido el número de grupos se entrenó el modelo K-means y el resultado del proceso fue la

predicción de cada dato dentro de un grupo, es decir, una etiqueta. Estas etiquetas fueron adicionadas a los datos sin etiquetar para obtener un nuevo dataset con datos etiquetados (figura 3.7).

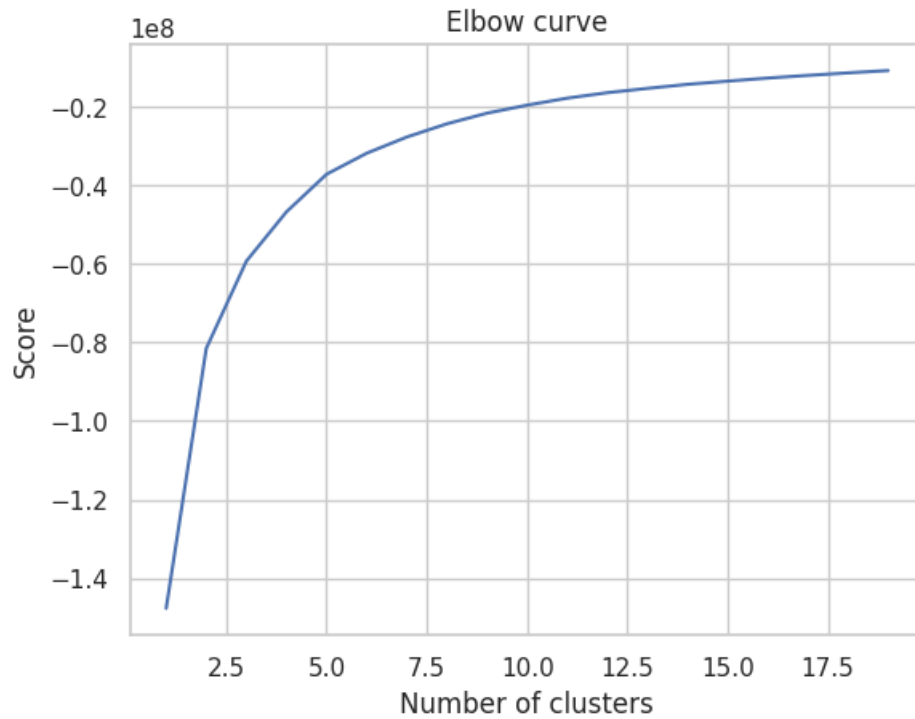


Figura 3.6: Elbow curve

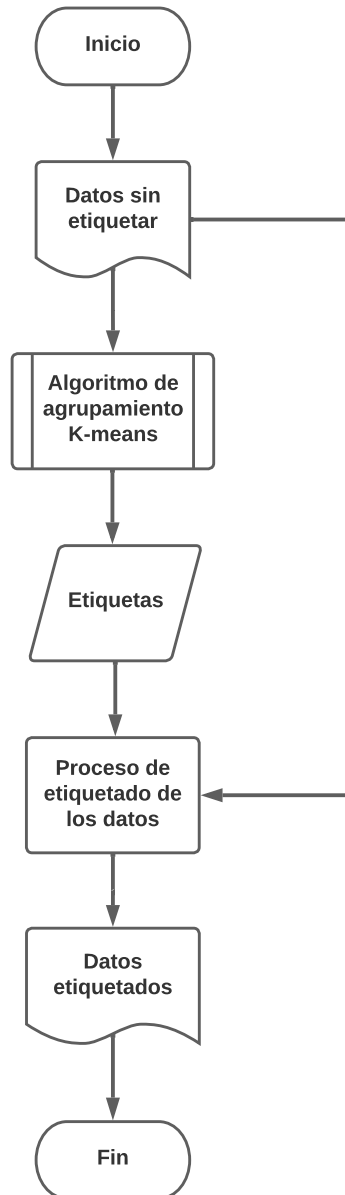


Figura 3.7: Diagrama de flujo del proceso de etiquetado de los datos

El desarrollo de esta sección se realizó exclusivamente con fines académicos y de investigación. Los procedimientos realizados se encuentran de forma detallada en

el repositorio de [GitLab](#) en el cual se explica paso a paso cómo se procesaron y etiquetaron los datos (*ingeniería de características:repositorio en línea*).

3.5. Resultados

Para obtener los datos de temperatura e irradiación necesarios para estimar la generación de energía en un período de 15 años de cada uno de los módulos fotovoltaicos. Se hizo uso de herramientas gratuitas como el [PVGIS tool](#) y el [WORLD BANK GROUP](#), además, se procesaron los datos obtenidos en el [capítulo anterior](#) para obtener solo un conjunto de datos que contara con toda la información necesaria, descartando así que algún módulo fotovoltaico no tuviera información que fuera de importancia. Como resultado de lo anterior se obtuvo un dataset con datos mes a mes desde el 2005 al 2020 de temperatura e irradiación y asimismo una clase de python que representa un modelo matemático de módulos fotovoltaicos, el cual se usó para calcular valores de generación y posteriormente crear un dataset con 180 datos de generación por cada uno de los módulos y que en total suman más de 275.000 valores.

Por otro lado, se hizo uso de la librería de aprendizaje automático de software gratuito [Scikit-learn](#) para agrupar los datos en 5 grupos, los cuales se graficaron usando colores para diferenciarlos entre sí (figura 3.8). Es pertinente aclarar que debido a la naturaleza de la gráfica en 3D no se aprecian los centroides de cada grupo. También permitió crear un nuevo [dataset](#) de datos etiquetados.

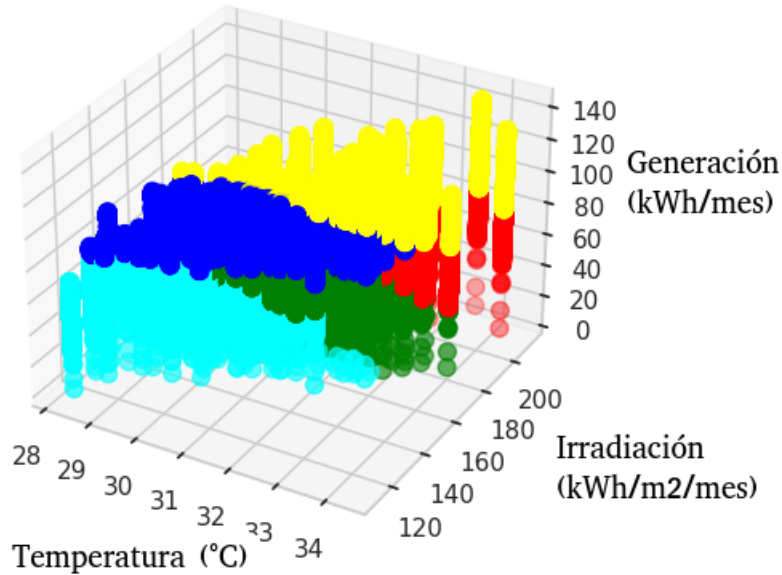


Figura 3.8: Datos agrupados por el algoritmo K-means

3.6. Conclusiones

El proceso de caracterización con diversos datos que no presentan una relación aparente, pueden llevarse a cabo de forma sistemática y ordenada con la aplicación de técnicas de ingeniería de características como la transformación por medio de modelos matemáticos para obtener nuevos valores relevantes y la visualización de los datos de forma cruzada para conocer su distribución y comportamiento. Asimismo, el uso de herramientas como algoritmos de machine learning facilitan el etiquetado de los datos para el proyecto que no solo requiere el reconocimiento de la forma en la que pueden agruparse los datos, sino también considerar aspectos adyacentes como los futuros modelos de machine learning que serán entrenados.

El desarrollo de diferentes procedimientos y técnicas aplicadas en este estudio ha permitido disponer de herramientas para la caracterización de los datos obtenidos,

que como resultado más importante facilita el entrenamiento de modelos de machine learning en este y futuros estudios.

De esta forma, se pueden visualizar los grupos a los que corresponden cada uno de los módulos fotovoltaicos, resultado de especial interés que guiará en la futura selección de un algoritmo de machine learning que haga posible clasificar nuevos datos dentro de uno de los grupos.

Referencias

- [1] Sjoerd Boeschoten, Cagatay Catal, Bedir Tekinerdogan, Arjen Lommen, and Marco Blokland. The automation of the development of classification models and improvement of model quality using feature engineering techniques. *Expert Systems with Applications*, 213, 2023. Cited by: 0; All Open Access, Hybrid Gold Open Access. 38
- [2] Zahraa Said Abdallah, Lan Du, and Geoffrey I Webb. Data preparation., 2017. 38
- [3] Sahar Zare, Zahra Meidani, Maryam Ouhadian, Hosein Akbari, Farid Zand, Esmaeil Fakharian, and Roxana Sharifian. Identification of data elements for blood gas analysis dataset: a base for developing registries and artificial intelligence-based systems. *BMC Health Services Research*, 22(1):1–20, 2022. 38
- [4] Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, and Manuel Graña. On the combination of graph data for assessing thin-file borrowers’ creditworthiness. *Expert Systems with Applications*, 213:118809, 2023. 38
- [5] Wen Long, Zhichen Lu, and Lingxiao Cui. Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, 164:163–173, 2019. 38
- [6] Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6):375–381, 2003. 38

- [7] Ayman Taha, Ali S. Hadi, Bernard Cosgrave, and Susan McKeever. A multiple association-based unsupervised feature selection algorithm for mixed data sets. *Expert Systems with Applications*, 212, 2023. Cited by: 0. 38
- [8] MH Padgavankar, SR Gupta, et al. Big data storage and challenges. *International Journal of Computer Science and Information Technologies*, 5(2):2218–2223, 2014. 39
- [9] Carlos Robles-Algarín, Diego Restrepo-Leal, and Adalberto Ospino Castro. Data from multimodal functions based on an array of photovoltaic modules and an approximation with artificial neural networks as a scenario for testing optimization algorithms. *Data in brief*, 27:104669, 2019. 43
- [10] Carlos Robles Algarín, Deimer Sevilla Hernández, and Diego Restrepo Leal. A low-cost maximum power point tracking system based on neural network inverse model controller. *Electronics*, 7(1):4, 2018. 43
- [11] Eduardo Ivan Ortiz Rivera. Modeling and analysis of solar distributed generation. *Ph. D. Thesis*, 2006. 43

Capítulo 4

ENTRENAMIENTO DE UN MODELO INTELIGENTE

4.1. Selección del modelo de aprendizaje

El uso de modelos inteligentes trae consigo el importante reto de entrenar con datos obtenidos ya sea de una base de datos, usando métodos de web scraping o realizando la recolección por medio de sistemas de monitoreo. A medida que las aplicaciones o actividades requieren el respaldo de un sistema inteligente, la demanda de nuevos algoritmos y modelos entrenados aumentará, creando la necesidad de implementar mecanismos que faciliten el entrenamiento y utilización según sean los datos de los que debe aprender el modelo.

Uno de los sectores que viene impulsando fuertemente la realización de modelos inteligentes, es la salud, lo que le convierte en un atractivo importante dadas las políticas de conservación de la vida. Esto ha generado la implementación de diversos estudios en los que los modelos inteligentes se convierten en una herramienta que facilita los diagnósticos y prevención de algunas enfermedades como la leucemia [1], el Covid-19 [2] y lesiones en la piel [3], todo esto basado en la clasificación de imágenes.

Sin embargo, el uso de el algoritmo de clasificación es ampliamente usado en otros sectores para clasificar texto por categorías [4], para la clasificación de dispositivos IOT según sea su tráfico de datos [5], para clasificar sectores según sean sus características sísmicas [6] e incluso en la industria 4.0 la aplicación de sistemas inteligentes ayuda a mejorar los tiempos de ciclo de trabajo [7]. Por otro lado, el sector energético no se queda atrás, y en la creciente popularidad de las energías limpias los modelos inteligentes se han convertido en una herramienta de gran ayuda, por lo que han puesto gran empeño en diseñar sistemas inteligentes de monitoreo de fallas en los sistemas fotovoltaicos [8] [9] y modelos de predicción para generación de energía de un sistema fotovoltaico [10] [11].

En la selección de un modelo de aprendizaje es necesario tener en cuenta la estructura de los datos con los cuales se entrenará el modelo. En el caso particular de esta investigación, la forma en la que se procesaron y caracterizaron los datos tuvo gran influencia en la selección de un algoritmo supervisado de clasificación, es decir, un algoritmo que catalogue una entrada dentro de uno de los 5 grupos previamente realizados.

Teniendo en cuenta que existen diferentes algoritmos de clasificación y diferentes tecnologías con las cuales se pueden implementar, llega a ser complejo escoger un modelo que se adapte a las necesidades. Debido a esto se hizo uso del road map de [Scikit-learn](#) (figura 4.1) el cual se siguió según las condiciones que se presentan en la ruta, como la cantidad de datos con los que se cuenta y si estos se encuentran etiquetados o no. Al seguir esta ruta, se llegó al algoritmo de clasificación [SGD_classifier](#), el cual es un algoritmo basado en el descenso de gradiente estocástico y que además, se puede implementar con diferentes clasificadores lineales como el SVM y logistic regression, entre otros.

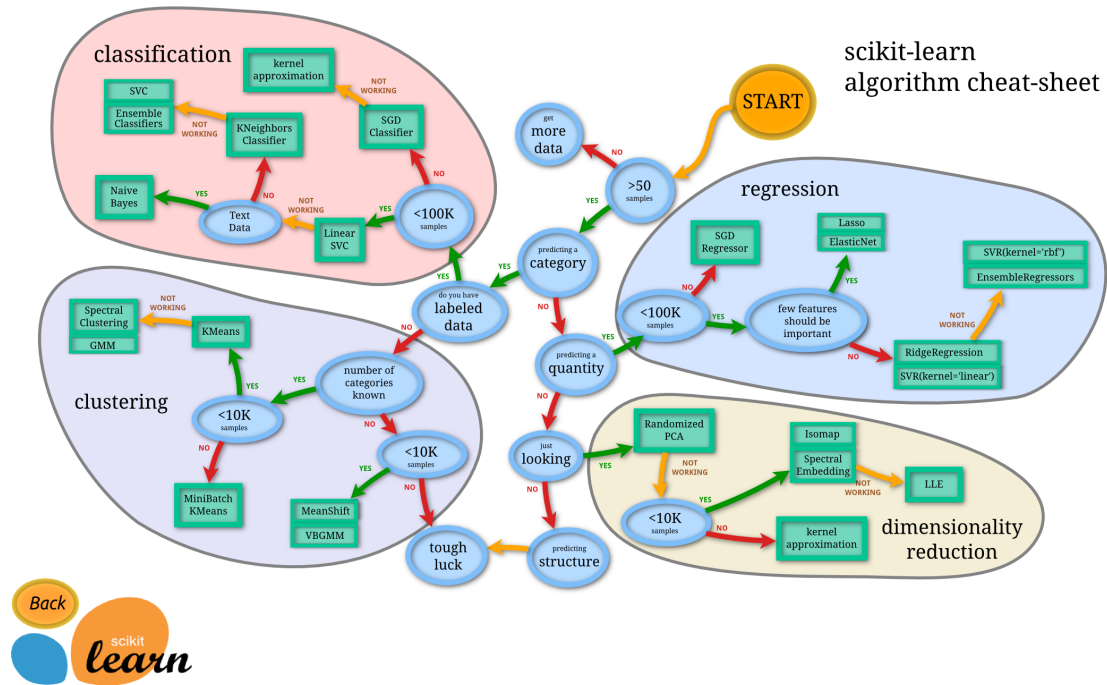


Figura 4.1: Road map de scikit-learn

Como paso final en el procesamiento de los datos antes de entrenar cualquier modelo, se seleccionaron las características de mayor relevancia, en este caso se seleccionaron las columnas de temperatura, irradiación y potencia del módulo fotovoltaico. Estas tres características tienen un efecto directo tanto en la eficiencia como en la generación del módulo fotovoltaico y es la principal razón por las que fueron tomados. Una vez determinadas las características, es importante estandarizar los datos con los que se entrenará y validará el modelo de clasificación, dado que para este, es mucho más sencillo procesar datos escalados en un único formato y rango. Por consiguiente, para seleccionar el algoritmo de estandarización es necesario conocer la distribución los datos. Se graficaron los datos de cada una de las características seleccionadas (figuras 4.2 , 4.3 y 4.4).

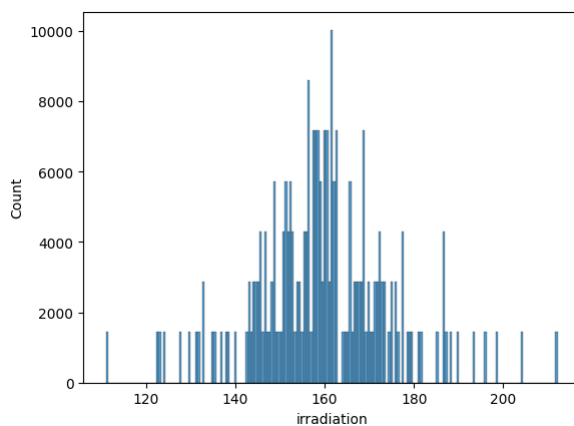


Figura 4.2: Distribución de la irradiación

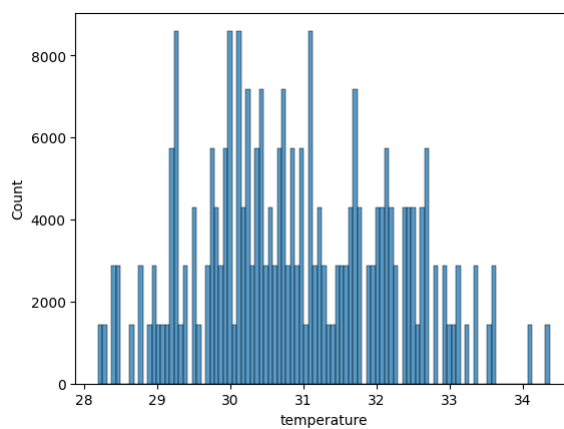


Figura 4.3: Distribución de la temperatura

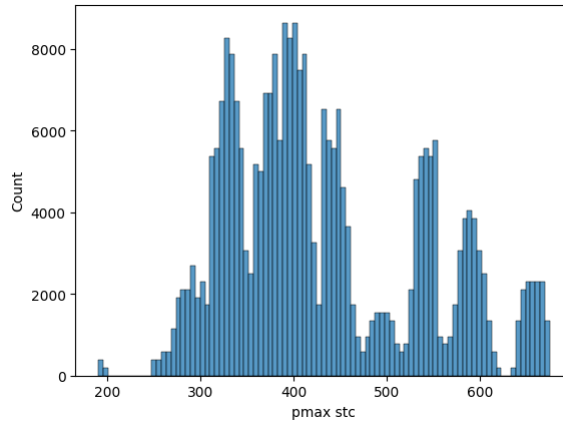


Figura 4.4: Distribución de la potencia

De las figuras 4.2 , 4.3 y 4.4 se observa que los datos se ajustaron a una distribución normal, y en consecuencia se utilizó el algoritmo standard scaler de la biblioteca de Python [Scikit-learn](#), que está optimizado para escalar datos que se distribuyen normalmente.

4.2. Entrenamiento del modelo de aprendizaje

Dado que para el algoritmo [SDG_calssifier](#) existen diferentes clasificadores lineales se tomó la decisión de usar el [Linear Support Vector Classifier](#) y el [Logistic regression](#) para comparar su precisión al entrenarlos por separado usando el mismo conjunto de datos, el cual fue dividido en 70 % de los datos para entrenamiento y 30 % para los datos de validación.

El primer clasificador fue el logistic regressor (figura 4.5) el cual se entrenó usando los hiperparámetros que [Scikit-learn](#) usa por defecto. Esto dejó como resultado una precisión del 89.70 %. Por otro lado, el clasificador linear support vector classifier (figura 4.6) también entrenado usando sus hiperparámetros por defecto dejó como resultado una precisión del 89.43 %. Estas pruebas iniciales se realizaron con el fin de escoger un clasificador que ofreciera la mejor precisión para posteriormente intentar

obtener mejores resultados. Cabe resaltar que, aunque los resultados no parecen ser muy diferentes, ese pequeño porcentaje puede ser significativo para el modelo de aprendizaje y la aplicación.

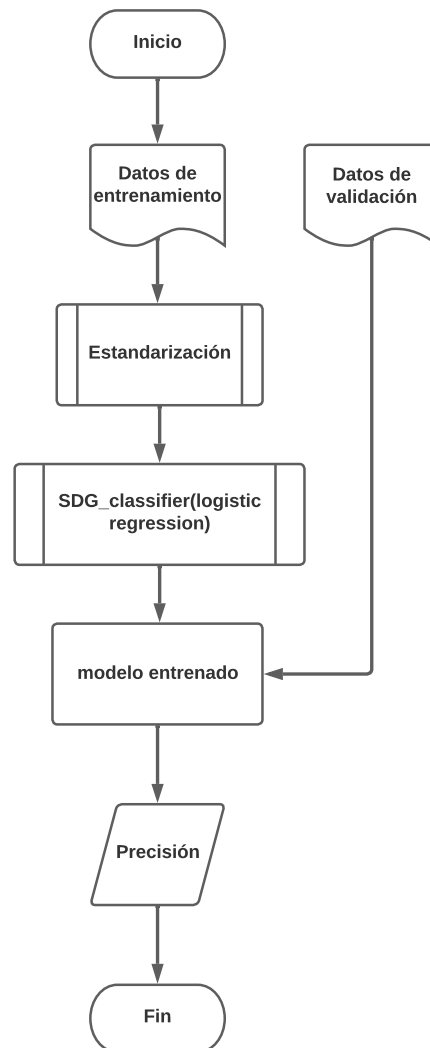


Figura 4.5: SDG Classifier (logistic-regression)

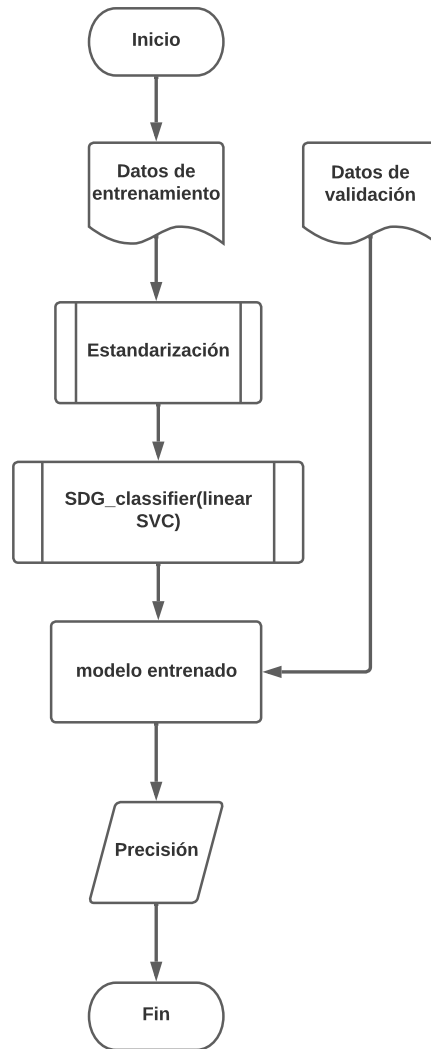


Figura 4.6: SDG Classifier (linear-SVC)

Aunque se obtuvieron buenos resultados en ambos modelos, fue pertinente realizar más pruebas en las cuales se usaran diferentes combinaciones de hiperparámetros con el fin de intentar mejorar el rendimiento del modelo. Por tal motivo, se hizo uso de una herramienta que ofrece [scikit-learn](#) llamada [search grid cv](#), la cual es una herramienta que entrena un modelo usando las diferentes combinaciones entre los

hiperparámetros definidos previamente y entrega como resultado el modelo que obtuvo el mejor rendimiento en su precisión (figura 4.7). En este caso particular el clasificador que obtuvo el mejor puntaje fue el logistic regressor con una precisión del 90.18 % usando un alpha de 0.001, un máximo de iteraciones de 4000 y un penalty igual a L1.

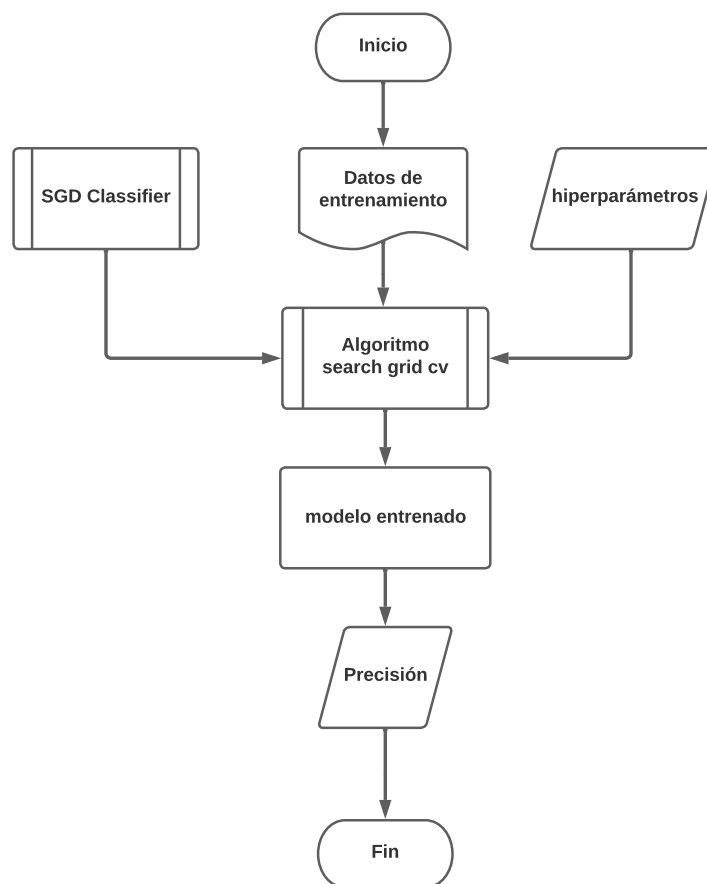


Figura 4.7: SGD Classifier usando search grid

Todos los modelos entrenados fueron guardados usando la librería [Pickle](#) de Python que permite almacenar objetos en formato .pkl, esto se realizó con el fin de

usar los modelos entrenados en otras aplicaciones. Cabe señalar que, el modelo entrenado usado en esta investigación lleva como nombre `solar_iama_classifier_model` y corresponde al clasificador que obtuvo la mejor precisión (un 90.18%). por otro lado recordemos que, este modelo predice un grupo al que corresponde una serie de datos entregados al modelo que corresponden a temperatura, irradiación y potencia del módulo fotovoltaico.

El desarrollo de esta sección se realizó exclusivamente con fines académicos y de investigación. Los procedimientos realizados se encuentran de forma detallada en el repositorio de [GitLab](#) en el cual se explica paso a paso cómo se seleccionaron y entrenaron los modelos usando Scikit-learn (*entrenamiento del modelo de aprendizaje: repositorio en línea*).

4.3. Resultados

Cada uno de los entrenamientos realizados obtuvo una precisión diferente según el clasificador y sus hiperparámetros. Para observar de una forma más clara se presenta la siguiente tabla:

Clasificador	Precisión
Logistic regressor	89.70 %
Linear support vector machine	89.43 %
Logistic regressor con search grid	90.18 %

Tabla 4.1: Modelos entrenados

4.4. Conclusiones

El desarrollo experimental del entrenamiento de modelos de machine learning como paso fundamental para conocer el comportamiento de un modelo de clasificación entrenado, requiere el desarrollo de pruebas donde se someta el modelo a

entrenamientos con diferentes clasificadores e hiperparámetros.

Como parte del estudio, se desarrollaron pruebas con los clasificadores lineal support vector classifier y logistic regressor usando diferentes valores para algunos de sus hiperparámetros. El método para la estimación de la mejor combinación de hiperparámetros propuestos fue el algoritmo search grid.

Con el desarrollo experimental y como uno de los principales aportes de los resultados obtenidos, se encuentra la mejora de la precisión al cambiar los valores de los hiperparámetros del modelo. Cabe resaltar que, los valores obtenidos de cada entrenamiento pueden verse afectados por diferentes factores, además, según se sigan probando más algoritmos y combinaciones de hiperparámetros los resultados obtenidos podrían mejorar considerablemente.

Referencias

- [1] Jürgen Schedlbauer, Georgios Raptis, and Bernd Ludwig. Medical informatics labor market analysis using web crawling, web scraping, and text mining. *International Journal of Medical Informatics*, 150:104453, 2021. 54
- [2] Mazin Abed, Karrar Hameed Mohammed, Garcia-Zapirain Abdulkareem, Mostafa Begonya, A Salama, Mashael S Maashi, Alaa S Al-Waisy, Mohammed Ahmed, Ammar Awad Subhi, Le Mutlag, et al. A comprehensive investigation of machine learning feature extraction and classification methods for automated diagnosis of covid-19 based on x-ray images. *Computers, Materials, & Continua*, pages 3289–3310, 2021. 54
- [3] Mohamed A Kassem, Khalid M Hosny, Robertas Damaševičius, and Mohamed Meselhy Eltoukhy. Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review. *Diagnostics*, 11(8):1390, 2021. 54
- [4] Xiaoyu Luo. Efficient english text classification using selected machine learning techniques. *Alexandria Engineering Journal*, 60(3):3401–3409, 2021. 55
- [5] Ivan Cvitić, Dragan Peraković, Marko Periša, and Brij Gupta. Ensemble machine learning approach for classification of iot devices in smart home. *International Journal of Machine Learning and Cybernetics*, 12(11):3179–3202, 2021. 55

- [6] U. Muksin, E. Riana, A. Rudyanto, K. Bauer, A.V.H. Simanjuntak, and M. Weber. Neural network-based classification of rock properties and seismic vulnerability. *Global Journal of Environmental Science and Management*, 9(1):15 – 30, 2023. Cited by: 0. 55
- [7] Toly Chen and Yu-Cheng Wang. Hybrid big data analytics and industry 4.0 approach to projecting cycle time ranges. *The International Journal of Advanced Manufacturing Technology*, 120(1):279–295, 2022. 55
- [8] Aref Eskandari, Jafar Milimonfared, and Mohammadreza Aghaei. Fault detection and classification for photovoltaic systems based on hierarchical classification and machine learning technique. *IEEE Transactions on Industrial Electronics*, 68(12):12750–12759, 2020. 55
- [9] D Manno, G Cipriani, G Ciulla, V Di Dio, S Guarino, and V Lo Brano. Deep learning strategies for automatic fault diagnosis in photovoltaic systems by thermographic images. *Energy Conversion and Management*, 241:114315, 2021. 55
- [10] Spyros Theocharides, George Makrides, George E Georghiou, and Andreas Kyprianou. Machine learning algorithms for photovoltaic system power output prediction. In *2018 IEEE International Energy Conference (ENERGYCON)*, pages 1–6. IEEE, 2018. 55
- [11] Jiaming Li, John K Ward, Jingnan Tong, Lyle Collins, and Glenn Platt. Machine learning for solar irradiance forecasting of photovoltaic system. *Renewable energy*, 90:542–553, 2016. 55

Capítulo 5

IMPLEMENTACIÓN DE UNA APLICACIÓN WEB

5.1. Funcionamiento general

Una vez entrenado el modelo de aprendizaje se creó una interfaz capaz de usar el modelo entrenado para entregar al usuario un resultado de 5 módulos fotovoltaicos que mejor se adapten a las condiciones entregadas, en este caso: temperatura, irradiación y potencia del módulo que se desea. Para ello, la API debe ser capaz de realizar las consultas a la base de datos, por lo cual el funcionamiento general de la aplicación web consta de un frontend y un backend que se encargan de todo el trabajo (figura 5.1).

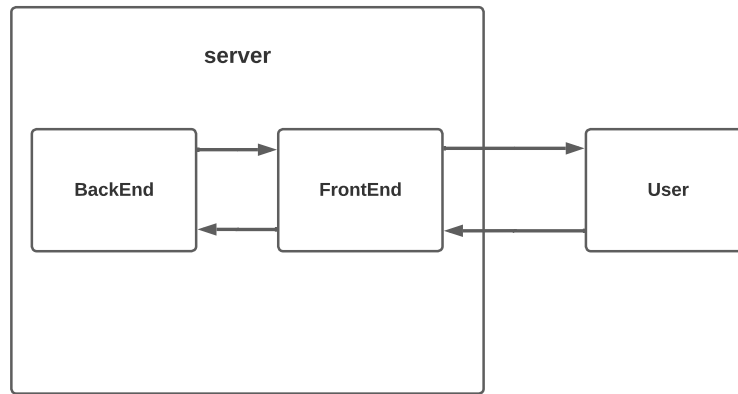


Figura 5.1: Estructura general

5.2. BackEnd

El punto inicial fue el desarrollo de la base de datos, para ello se usaron los 5 grupos antes creados. No obstante, antes de realizar la creación de la base de datos, cada uno de los dataframes tuvo una modificación por motivos de simplicidad en las consultas a realizar en la base de datos. Se usó un algoritmo de [PCA](#) (análisis de características principales) para realizar una reducción de características tal que la temperatura, la irradiación y la potencia se redujeran a una sola, este nuevo valor se agregó a su respectivo dataframe; posteriormente se creó la base de datos usando la librería [SQLAlchemy](#) de Python y el motor de bases de datos [MariaDB](#). La comunicación a la base de datos se realiza mediante una API escrita con el framework [FastAPI](#) que gestiona el tráfico de consulta SQL correspondiente a la petición, es decir, la API recibe valores de irradiación, temperatura y potencia, los cuales son entregados al modelo entrenado que predice un grupo. También, se les realiza una reducción de características para obtener un solo valor y luego realiza la consulta SQL buscando valores que estén entre más o menos 20 % del valor absoluto obtenido

de la reducción. Cabe resaltar que, la consulta además de filtrar por este valor, organiza los datos por generación y devuelve los primeros 5 módulos del grupo que predijo el modelo, garantizando así que de los módulos que se encuentren en el rango establecido sean organizados de mayor a menor según haya sido su generación bajo condiciones similares de irradiación, temperatura y potencia del módulo. Una vez realizado estos procesos, la API devuelve la información general en condiciones STC de los módulos (figura 5.2).

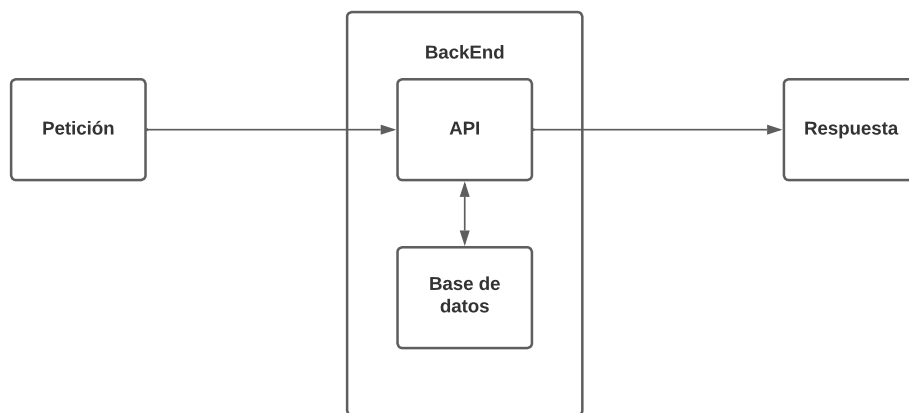


Figura 5.2: BackEnd con FastAPI y MariaDB

5.3. FrontEnd

La interfaz gráfica se diseñó con [Next.js](#), que ofrece ventajas sobre la renderización de la página web. Por lo tanto, mejora el tiempo de respuesta, que resulta en una mejor experiencia de usuario. El diseño está basado en un concepto minimalista el cual solo incluye lo necesario, su funcionamiento está basado en un formulario en el que los usuarios pueden digitar los datos y un botón que permite ejecutar la consulta a la base de datos mediante la API (figura 5.3). Como respuesta se presentan los resultados justo debajo del formulario presentando el logo de los fabricantes de

los módulos que hayan sido seleccionados (figura 5.4). Por otro lado, al dar click sobre alguno de los módulos de interés, se despliega la información general del módulo fotovoltaico bajo condiciones STC (figura 5.5).



Lorem ipsum dolor sit amet consectetur adipisicing elit. Voluptatem distinctio possimus labore voluptates. Pariatur quia sapiente adipisci maxime eos possimus, corrupti ad reprehenderit, architecto nemo aliquam rerum libero! Aut eveniet numquam aspernatur explicabo similique laudantium, quaerat qui impedit optio beatae sint, unde quibusdam magnam modi in eligendi odio sit distinctio itaque? Voluptatem enim nihil perspiciatis autem ducimus ab atque aspernatur laboriosam, adipisci, voluptatibus, maxime tenetur debitis?.

Lorem ipsum dolor sit amet consectetur adipisicing elit. Voluptatem distinctio possimus labore voluptates. Pariatur quia sapiente adipisci maxime eos possimus, corrupti ad reprehenderit, architecto nemo aliquam rerum libero! Aut eveniet numquam aspernatur explicabo similique laudantium, quaerat qui impedit optio beatae sint, unde quibusdam magnam modi in eligendi odio sit distinctio itaque? Voluptatem enim nihil perspiciatis autem ducimus ab atque aspernatur laboriosam, adipisci, voluptatibus, maxime tenetur debitis?.

Completa el siguiente formulario para estimar las mejores opciones:

Potencia del módulo	W
Temperatura	°C
Irradiación por mes	kWh/m²

BUSCAR

Figura 5.3: Vista principal de la aplicación web

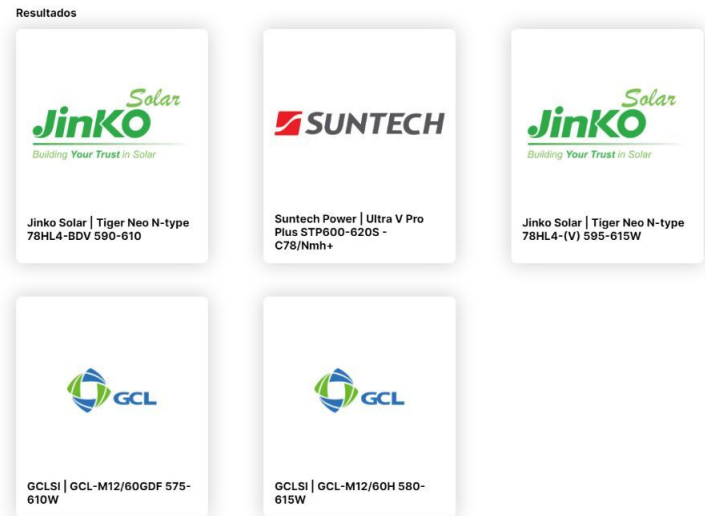


Figura 5.4: Resultados de la petición

Figure 5.5 displays a table with specifications for Jinko Solar Tiger Neo N-type 78HL4-BDV 590-610. The table is overlaid on a background image of solar panels.

Especificaciones de Jinko Solar Tiger Neo N-type 78HL4-BDV 590-610	
Fabricante:	Jinko Solar
Modelo:	JKM610N-78HL4-BDV
Tipo:	Bifacial
Potencia máxima:	610 W
Voltaje máximo:	45,6 V
Corriente en corto circuito:	14,03 A
Corriente máxima:	13,38 A
Eficiencia:	21,82 %
Tolerancia:	+ 3 %
Rango de temperatura:	-40~85 °C
Coefficiente de temperatura de potencia máxima:	-0,3 %/°C
Coefficiente de temperatura de voltaje en circuito abierto:	-0,25 %/°C
Coefficiente de temperatura de corriente en corto circuito:	0,046 %/°C

Figura 5.5: Información STC del módulo FV

5.4. Despliegue de la aplicación web

Como parte fundamental en el despliegue de la aplicación se hizo uso de los contenedores de [Docker](#), los cuales nos permiten ejecutar las aplicaciones en ambientes

aislados, esto permite evitar problemas con las versiones, incompatibilidad de paquetes, problemas con el sistema operativo y con el hardware, además ofrece una fácil migración entre sistemas. Se usaron contenedores diferentes para el backend y frontend y se utilizó como servicio de cloud azure debido a la gran compatibilidad con Docker y su facilidad y rapidez de despliegue, entre otras cosas.

El desarrollo de esta sección se realizó exclusivamente con fines académicos y de investigación. Los procedimientos realizados se encuentra de forma detallada en el repositorio de [GitLab](#) en el cual se explica paso a paso cómo se creó la pagina web (*aplicación web repositorio en línea*).

5.5. Conclusiones

Este trabajo de investigación culmina con la implementación de una aplicación web que utiliza un modelo inteligente para recomendar paneles que se adaptan a las condiciones de temperatura e irradiación propias de una zona, en este caso la departamento del Magdalena, Colombia.

La aplicación web propuesta e implementada en el desarrollo de esta investigación debe concebirse como un sistema capaz de entregar una respuesta adecuada según los datos de entrada y también resaltar que lo aplicado en esta investigación puede ser extrapolado a diferentes zonas de Colombia.