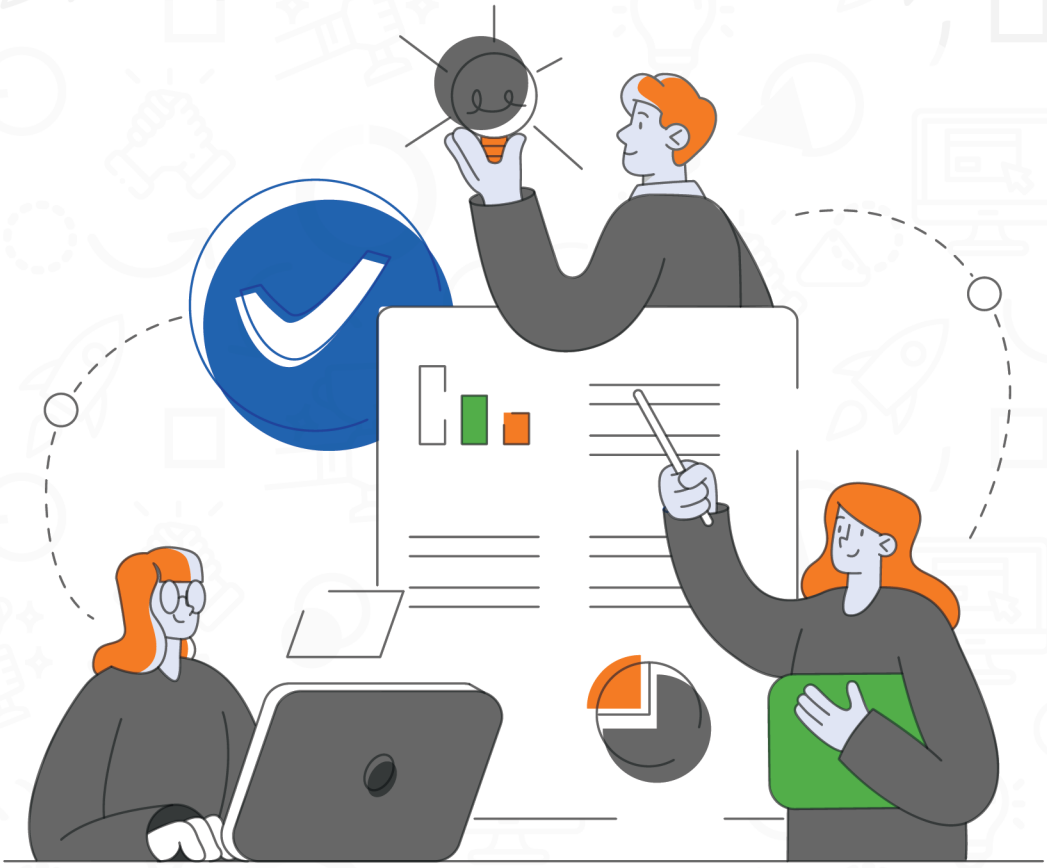
MANUAL PARA EL MANEJO DEL SISTEMA DE MONITOREO DE LA SEGURIDAD ALIMENTARIA Y NUTRICIONAL (SM-SAN)



Financiado por Unión Europea Protección Civil y Ayuda Humanitaria

CONSORCIO DE ORGANIZACIONES HUMANITARIAS

tech IS

Spin-Off UNIVERSIDAD DE GRANADA

ACCIÓN CONTRA EL HAMBRE

CRÉDITOS

Proyecto *"Maximizar el impacto de la asistencia humanitaria en el Corredor Seco Centroamericano mejorando la generación y difusión de datos de Seguridad Alimentaria y Nutricional"* - ECHO/-AM/BUD/2020/91013, cuyo objetivo es de contribuir a mejorar la respuesta humanitaria a la inseguridad alimentaria de los actores nacionales y regionales del Corredor Seco Centroamericano.

Financiado por: Dirección de Protección Civil y Ayuda Humanitaria de la Comisión Europea (ECHO)

Dirección:
Fundación Acción Internacional contra el Hambre
Consorcio de Organizaciones Humanitarias

Autoría:
MSc. Fidae El Morei
MSc. Jorge Hernández Marín
Dr. Francisco Javier Abarca Alvarez

Diseño y diagramación:
Hiliana Nuñez

Las ideas y argumentaciones expresadas en este documento son las de los autores, y no reflejan necesariamente el punto de vista de Acción contra el Hambre ni del Consorcio de Organizaciones Humanitarias ni del donante.
Todos los derechos reservados. Se autoriza la reproducción y difusión del material contenido en este producto para fines educativos u otros fines no comerciales sin previa autorización escrita de los titulares de los derechos de autor, siempre que se especifique claramente la fuente.

Julio, 2021

ÍNDICE

1. Introducción a la plataforma del SM - SAN ——— 13

1.1. Objetivos 14

1.2. Funcionamiento 15

1.2.1. Mapas 15

1.3. Datos de encuestas SAN 20

1.4. Espacio de reportes 20

2. Fuentes de datos ——— 22

2.1. Datos socioeconómicos 23

2.2. Datos agroclimáticos 23

2.3. Datos de campo SAN 24

3. Cálculo de índices agroclimáticos ——— 25

3.1. Instalación de QGIS 26

a. Instalador autónomo de QGIS (Recomendada) 26

3.2. Cálculo de índices con MODIS 28

3.2.1. Descarga de imágenes desde EARTHDATA 29

3.2.2. Cálculo del índice NDVI 34

3.2.3. Cálculo del índice EVI 38

3.2.4. Cálculo del índice ARVI 40

3.2.5. Cálculo del índice SIPI 42

3.3. Cálculo del índice SPI 43

3.4. Cálculo de estadísticas zonales 46

4. Preparación de datos para el modelo predictivo — 49

4.1. Datos socioeconómicos 50

4.2. Datos agroclimáticos 51

4.3. Datos de Seguridad Alimentaria y nutricional 54

4.4. Montaje de la tabla final 57

5. Introducción a la ciencia de datos — 59

5.1. Objetivos y aplicaciones 60

5.2. Modelos predictivos 60

5.3. Tipos de modelos predictivos: regresión / clasificación 61

5.3.1. Algoritmos de clasificación 61

5.3.2. Algoritmos de regresión 62

6. Manejo del software Orange — 63

6.1. Instalación de Orange 64

6.2. Primeros pasos 65

6.3. Ejemplo de aplicación 67

6.3.1. Bloque de herramientas de datos 67

6.3.2. Bloque de visualización 69

6.3.3. Bloque de modelos 71

6.3.4. Bloque de validación 72

7. Preparación del modelo predictivo — 74

7.1. Introducción de datos 76

7.2. Generación de modelo predictivo 77

7.3. Introducción del conjunto de predicción 79

8. Introducción a Power BI — **85**

8.1. ¿Qué es y para qué sirve Power BI? 86

8.2. Arquitectura de la aplicación 86

8.3. Conexiones de datos para Power BI 87

9. Instalación de Power BI Desktop — **88**

10. Primeros pasos con la interfaz — **91**

11. Estructura interna de la plataforma — **94**

12. Relación entre las visualizaciones y los conjuntos de datos — **99**

13. Actualización de datos de la plataforma — **101**

14. Publicación de la plataforma — **106**

ÍNDICE DE ILUSTRACIONES

Ilustración 1. Interfaz del espacio de Mapas 15

Ilustración 2. Vista de la variable agroclimática NDVI en el espacio de Mapas 16

Ilustración 3. Vista de la variable agroclimática SPI en el espacio de Mapas 17

Ilustración 4. Vista de la Categoría SAN real en el espacio de Mapas 18

Ilustración 5. Vista de la Incidencia del Hambre en el Hogar (HHS) en el espacio de Mapas 19

Ilustración 6. Vista de las Zonas de Preocupación predichas en el espacio de Mapas 20
Ilustración 7. Vista del cuadro de mandos de las encuestas SAN en el espacio Datos de encuestas SAN 21
Ilustración 8. Vista del Espacio de Reportes 21
Ilustración 9. Vista de la web de QGIS 26
Ilustración 10. Opciones de descarga de QGIS para diferentes sistemas operativos 27
Ilustración 11. Opciones de descarga de QGIS 27
Ilustración 12.Vista inicial de la plataforma Earthdata 29
Ilustración 13. Ubicación de la caja de búsqueda en la plataforma Earthdata 30
Ilustración 14. Selección de productos para la descarga de imágenes desde Earthdata 30
Ilustración 15. Selección de imagen para la descarga 31
Ilustración 16. Vista previa de las descargas 32
Ilustración 17. Vista de los enlaces de descarga 33
Ilustración 18. Vista inicial de carga de la imagen en QGIS 33
Ilustración 19. Vista de la imagen descargada de la plataforma Earthdata 34
Ilustración 20. Ubicación de la herramienta “Calculadora ráster” 35
Ilustración 21. Procesamiento de la imagen NDVI en bruto 36
Ilustración 22. Cambio del tipo de renderizado a Pseudocolor monobanda 37
Ilustración 23. Selección de las opciones de representación de la imagen NDVI procesada 37
Ilustración 24. Imagen procesada con los valores del NDVI 38
Ilustración 25. Procesamiento de la imagen EVI en bruto 39

Ilustración 26. Imagen procesada con los valores del EVI 39
Ilustración 27. Cálculo del ARVI mediante la calculadora ráster 40
Ilustración 28. Valores del ARVI representados en 4 intervalos, representando masas de agua y zonas urbanas, vegetación y masas arbóreas 41
Ilustración 29. Imagen comparativa de los valores del ARVI (izquierda) y una imagen satelital con colores naturales (derecha) 41
Ilustración 30. Cálculo del SIPI mediante la calculadora ráster 42
Ilustración 31. Parámetros a introducir en el apartado de “Variable” en la página de Climate Engine 43
Ilustración 32. Parámetros a introducir en el apartado de “Time Period” en la página de Climate Engine 44
Ilustración 33: Paso de descarga en el repositorio de Climate Engine 45
Ilustración 34: Capa ráster con los valores de SPI para el ámbito de CA4 45
Ilustración 35. Capa vectorial que contiene la ubicación de las parcelas de grano básico 46
Ilustración 36. Localización de la función Estadísticas de zona 47
Ilustración 37. Panel de Estadísticas de zona 47
Ilustración 38. Panel de Estadísticas de zona 48
Ilustración 39. Tabla de atributos 48
Ilustración 40. Tabla 51
Ilustración 41. Exportar capa 51
Ilustración 42. Guardado de capa 52
Ilustración 43. Excel 53
Ilustración 44. Campos de tabla dinámica de Excel 53
Ilustración 45. Tabla dinámica de Excel 54

Ilustración 46. Tabla de Nivel de seguridad alimentaria por municipio 54
Ilustración 47. Generación de la categoría SAN 55
Ilustración 48. Tabla de recuento de categorías por municipio 56
Ilustración 49. Generación de la columna de zona de preocupación 56
Ilustración 50. Recuento de las zonas de preocupación por municipio 57
Ilustración 51. Montaje de la tabla final 58
Ilustración 52. Gráficos de proceso de clasificación 61
Ilustración 53. Gráfico de regresión lineal 62
Ilustración 54. Nuevo documento en Orange 65
Ilustración 55. Cinco bloques de herramientas 66
Ilustración 56. Tablero de Orange 66
Ilustración 57. Importación de un fichero csv 67
Ilustración 58. Exploración de datos cargados 67
Ilustración 59. Eliminar de conexión 68
Ilustración 60. Seleccionar columnas 68
Ilustración 61. Panel de selección de columnas 69
Ilustración 62. Gráficos 70
Ilustración 63. Diagrama de barras 70
Ilustración 64.Algoritmo Random Forest 71
Ilustración 65.Matriz de confusión 72
Ilustración 66.Métodos de evaluación 73
Ilustración 67. Burbujas necesarias para la introducción y la exploración de datos 76
Ilustración 68. Tabla de datos para el entrenamiento 76
Ilustración 69. Introducción de la burbuja de preprocesamiento 77

Ilustración 70. Selección de parámetros para el preprocesamiento 77
Ilustración 71. Esquema con la introducción de los modelos de predicción y la burbuja de validación 78
Ilustración 72. Resultados del modelo a partir del conjunto de entrenamiento 78
Ilustración 73. Esquema de introducción, exploración y selección de columnas para los datos de predicción 79
Ilustración 74. Esquema general de la generación del modelo predictivo 80
Ilustración 75. Resultados arrojados por los modelos predictivos 80
Ilustración 76. Introducción de la burbuja de distribuciones para el estudio de las probabilidades resultantes 81
Ilustración 77. Probabilidad de pertenencia a las zonas de preocupación altas para el modelo Random Forest 81
Ilustración 78. Probabilidad de pertenencia a las zonas de preocupación altas para el modelo de Redes Neuronales 82
Ilustración 79. Guardado de las predicciones 82
Ilustración 80. Esquema de datos incluyendo las predicciones realizadas 83
Ilustración 81. Distribución de los municipios en función de cuántas desviaciones típicas se tengan en cuenta 84
Ilustración 82. Web de descarga de Power Bi 89
Ilustración 83. Opciones de descarga de Power Bi 90
Ilustración 84. Espacios del programa Power Bi 92
Ilustración 85. Mapa de Seguridad alimentaria y nutricional en Power Bi 95
Ilustración 86. Transformación del tipo de datos de una columna 96
Ilustración 87. Relación entre tablas 97

Ilustración 88. Editar relación entre tablas 98
Ilustración 89. Configuración de fuentes de datos 103
Ilustración 90. Cambiar fuente de datos 104
Ilustración 91. Actualizar datos 105
Ilustración 92. Publicar panel 107
Ilustración 93. Selección del espacio de trabajo 107
Ilustración 94. Publicar panel 108
Ilustración 95. Publicar panel en la web 108
Ilustración 96. Generación del enlace del informe 109

ACRÓNIMOS

EVI	**Índice de vegetación mejorada**
FEWS NET	**Famine Early Warning Systems Network**
HHS	**Incidencia del hambre en el hogar**
ID	**Identificación**
IEA	**Índice de estrategias de afrontamiento**
NDVI	**Índice de vegetación de diferencia normalizada**
PCA	**Puntaje de consumo de alimentos**
PGA	**Porcentaje de gasto en alimentos**
SAN	**Seguridad Alimentaria y Nutricional**
SPI	**Índice estandarizado de precipitación**

INTRODUCCIÓN

El presente **Manual para el Manejo del Sistema de Monitoreo de la Seguridad Alimentaria y Nutricional** se elabora en el marco del proyecto "Maximizar el impacto de la asistencia humanitaria en el Corredor Seco Centroamericano mejorando la generación y difusión de datos de Seguridad Alimentaria y Nutricional", ECHO/-AM/BUD/2020/91013, financiado por la Dirección de Protección Civil y Ayuda Humanitaria de la Comisión Europea (ECHO) e implementado por Acción contra el Hambre y el Consorcio de Organizaciones Humanitarias: OXFAM, TROCAIRE, We World GVC, COOPI.

El Manual forma parte de los instrumentos de trabajo generados a partir del **Curso Manejo del Sistema de Monitoreo de la Seguridad Alimentaria y Nutricional -SM-SAN**, bajo el Programa de creación de capacidad regional y local para evaluaciones estandarizadas de seguridad alimentaria y nutricional a través del cual se espera poder efectuar un seguimiento de los riesgos conexos para las acciones previstas en el Corredor Seco Centroamericano de acuerdo con las normas internacionales.

El SM-SAN es una plataforma regional para la generación análisis y difusión de datos relacionado al estado de la inseguridad alimentaria en Centroamérica. La información alojada permite anticiparnos a una situación de crisis alimentaria y mejorar la respuesta humanitaria efectiva y focalizada de los actores nacionales y regionales. Con esta acción se espera generar capacidades en los equipos técnicos del consorcio de organizaciones humanitarias para recopilar, procesar y analizar datos fiables y de buena calidad sobre la SAN al tiempo que, se genera evidencia de información que plantee el estado actual y la evolución de la SAN en la región centroamericana.

El Manual ha sido elaborado por GIS4TECH Spin Off – Universidad de Granada, es un instrumento orientador que plantea el paso a paso para la gestión y el análisis de información de seguridad alimentaria y nutricional. Está dividido en catorce secciones, a partir de los cuales se hace un recorrido sobre el funcionamiento de la plataforma, principales indicadores, preparación de datos, modelo predictivo, software para el análisis de datos, arquitectura de la plataforma, visualización y publicación de información.

1

INTRODUCCIÓN A LA PLATAFORMA DEL SM-SAN

1. Introducción a la plataforma del SM-SAN

1.1 Objetivos

El objetivo principal del Sistema de Monitoreo de la Seguridad Alimentaria y Nutricional (de ahora en delante denominado Plataforma), es la gestión, visualización y el seguimiento de la información generada durante el transcurso del Proyecto ECHO/-AM/BUD/2020/91013, desarrollado entre los meses de julio de 2020 y junio de 2021.

La información generada en el seno del proyecto mencionado se puede clasificar en tres categorías, que pasarán a ser desglosadas posteriormente:

- **Datos socioeconómicos:**

  Consiste en información que caracteriza a los diferentes municipios de los países de CA4, haciendo alusión a datos como la población, la densidad, la mortalidad infantil, entre otros.

- **Datos agroclimáticos:**

  Estos datos se extraen de proveedores de imágenes satelitales, y sirven para evaluar el estado de los cultivos y determinar el impacto de la variabilidad meteorológica sobre los mismos.

- **Datos SAN:**

  Este paquete comprende la información levantada durante las campañas de toma de datos en campo realizadas por el Consorcio de ONG durante la duración del Proyecto.

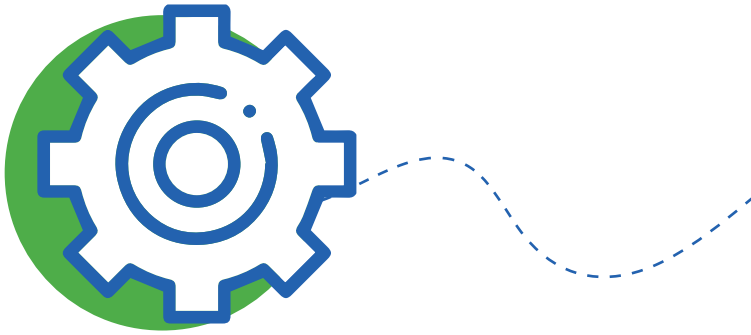
En cuanto a los objetivos particulares de la Plataforma, se pueden incluir los siguientes:

- Permitir la visualización de los indicadores parciales derivados de las tomas de datos en campo, así como de los propios resultados de las encuestas.
- Seguimiento de datos agroclimáticos, incluyendo la visualización mensual de los índices calculados y su evolución temporal.
- Almacenamiento, generación y descarga de informes relacionados con el ámbito de la Seguridad Alimentaria y Nutricional.

1.2 Funcionamiento

La Plataforma está estructurada en diferentes unidades:

- **Mapas**
  - **Variables agroclimáticas**
    - NDVI
    - SPI
  - **SAN Predicha**
    - Zonas de preocupación
  - **SAN Real**
    - Categoría SAN
    - HHS
    - IEA
    - PCA
    - PGA
- **Datos de encuestas SAN**
- **Espacio de reportes**



1.2.1. Mapas

Se trata del conjunto de pestañas principal, en el que se muestran los datos contenidos en la Plataforma a través de gráficos y mapas. Esta parte, a su vez, se desglosa en tres subcategorías.

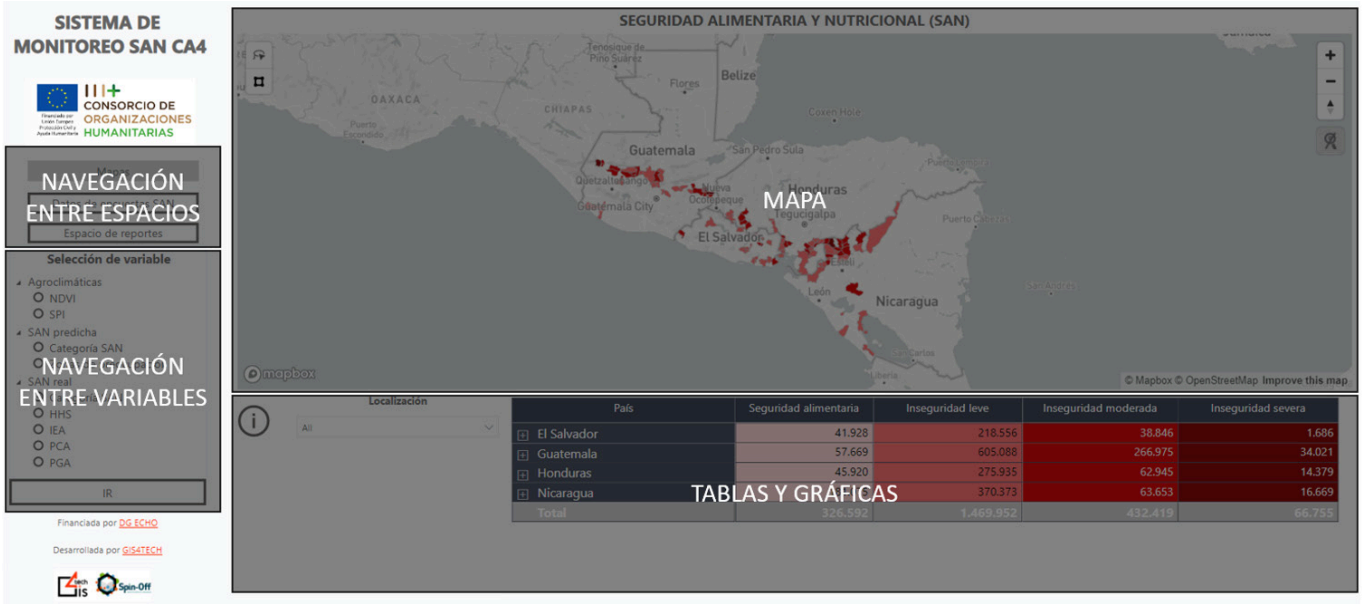


Ilustración 1. Interfaz del espacio de Mapas.

Variables agroclimáticas

Se muestran los índices agroclimáticos Normalized Difference Vegetation Index (NDVI por sus siglas en inglés) y Standardized Precipitation Index (SPI), que son los indicadores habitualmente empleados para la evaluación de la severidad de la sequía en una región dada.

- **NDVI**

  La información aparece reflejada en un mapa en el que se puede ver la anomalía del NDVI a lo largo del tiempo. Es decir, se toma el NDVI calculado en un mes determinado y se compara con el mismo mes de toda la serie histórica de la cual se disponen datos. De esta manera, se puede saber cuántas veces se desvía el valor calculado de lo habitual para la época del año estudiada. La fórmula seguida para realizar el cálculo es la siguiente:

$$Anomalía\ NDVI_{mes} = \frac{NDVI_{mes} - Media\ NDVI_{serie\ histórica}}{Desviación\ típica_{serie\ histórica}}$$

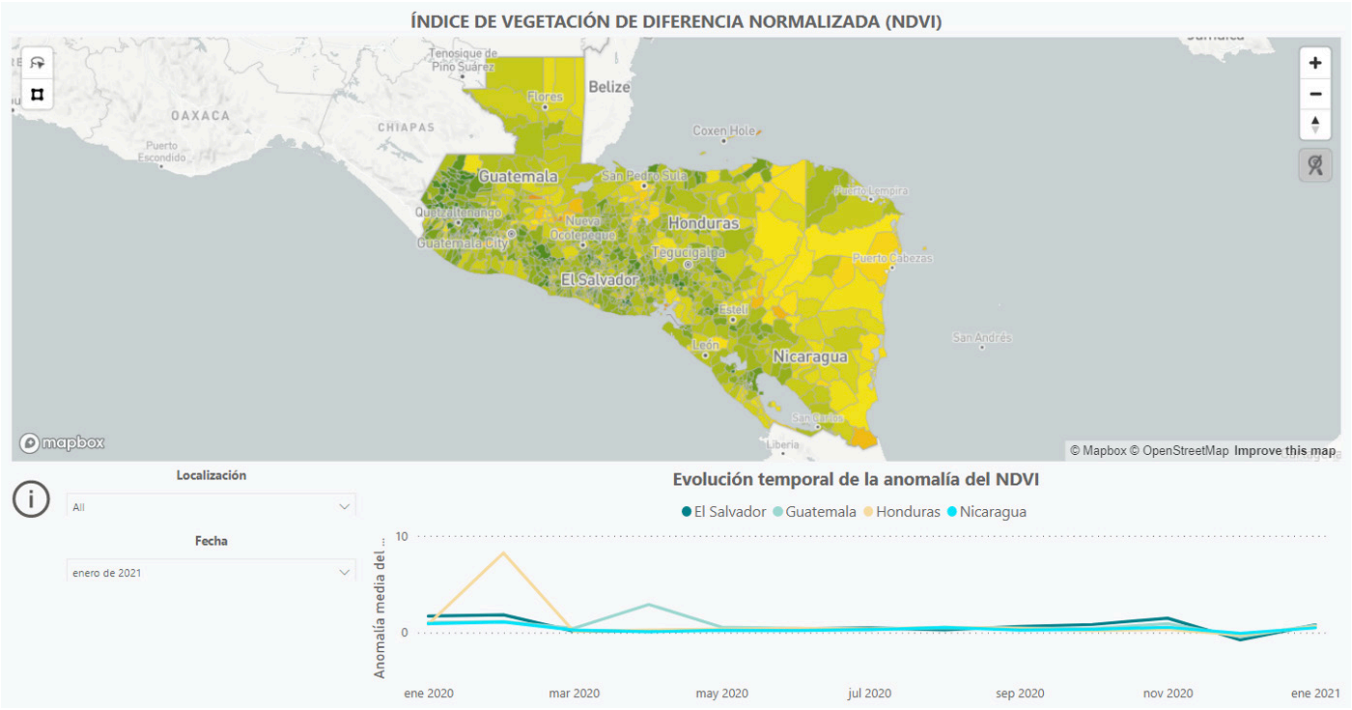


Ilustración 2. Vista de la variable agroclimática NDVI en el espacio de Mapas.

- **SPI**

El cálculo realizado para obtener el índice SPI es muy similar al del NDVI, con la particularidad de que no se realiza únicamente para un mes. Esto es, se puede comparar la media de precipitaciones de los últimos 3, 6, 9 ó 12 meses con la serie histórica. De esta manera, lo que se está obteniendo es un indicador de cuánto se diferencian las precipitaciones en un periodo determinado con el mismo periodo de otros años.

$$SPI_{mes} = \frac{Precipitaciones_{mes} - Media_{serie\ histórica}}{Desviación\ típica_{serie\ histórica}}$$

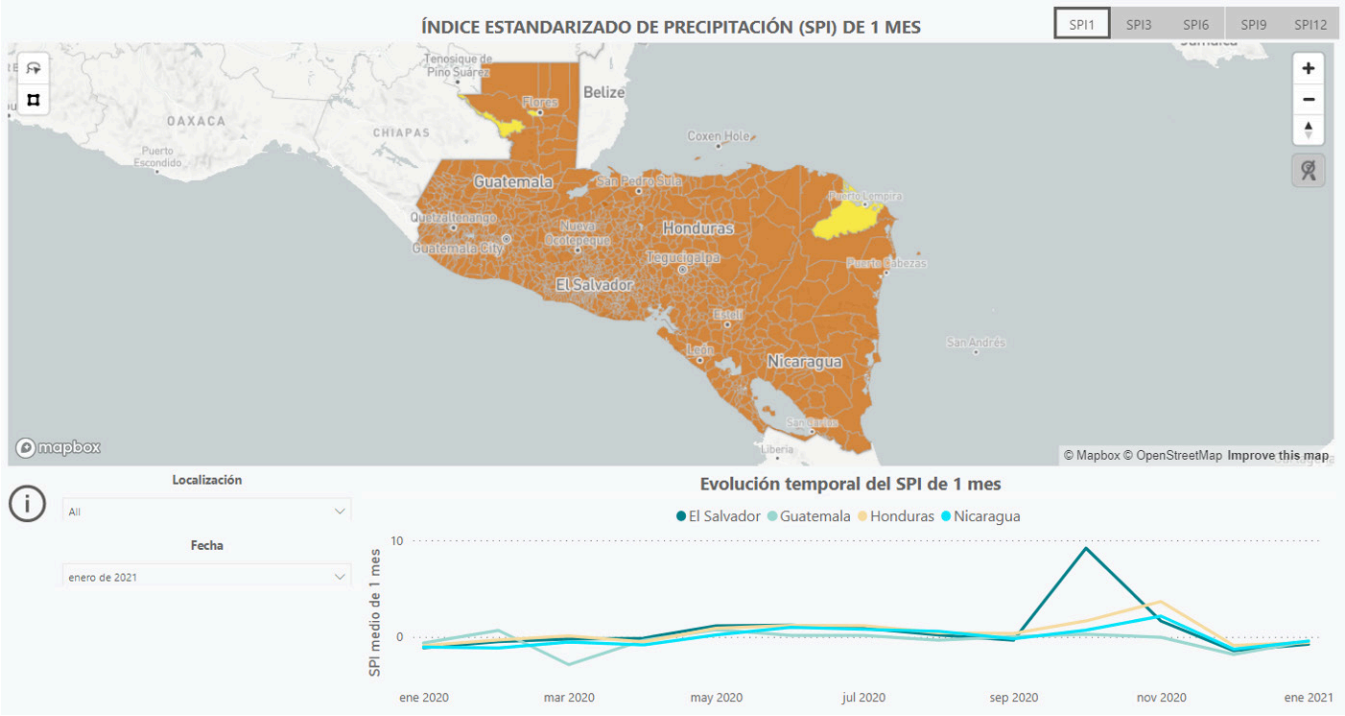


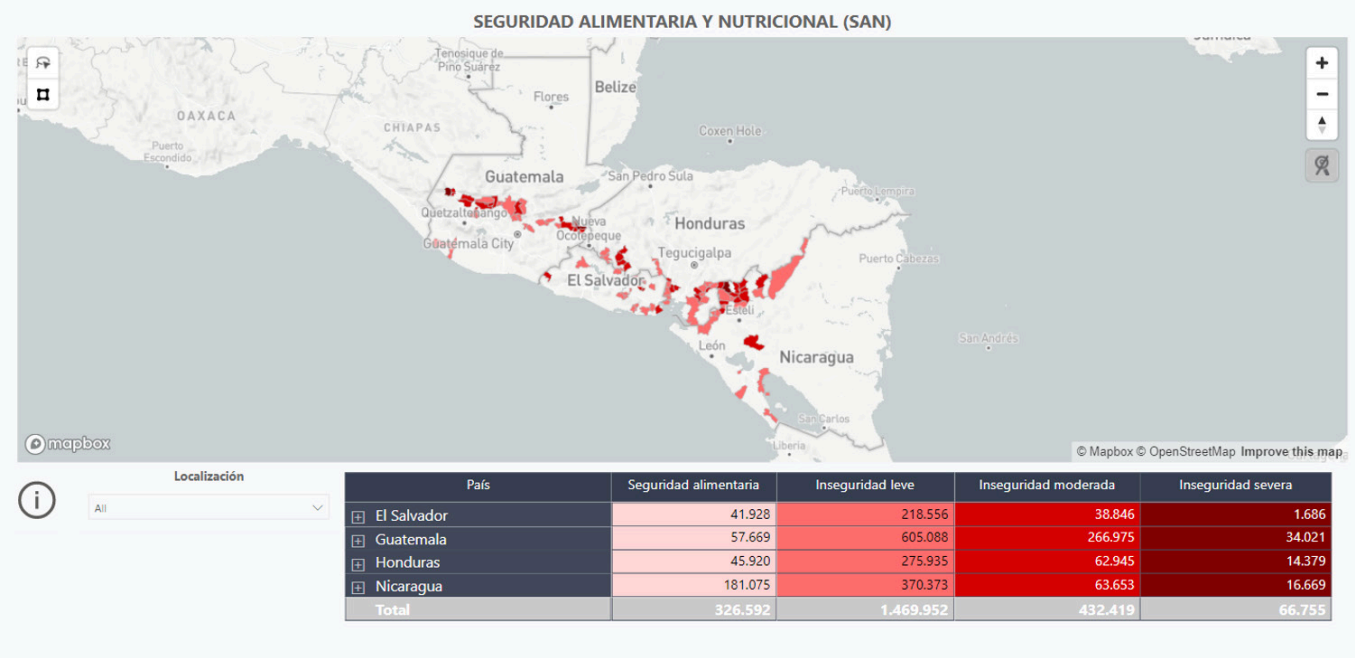
Ilustración 3. Vista de la variable agroclimática SPI en el espacio de Mapas.

SAN real

Esta categoría contiene la información de las encuestas tomadas en campo a lo largo del tiempo, incluyendo los índices parciales (HHS, IEA, PCA y PGA) y la categoría SAN resultante.

- **Categoría SAN**

  Muestra los resultados de la categoría de Seguridad Alimentaria y Nutricional de acuerdo con los datos tomados en campo. Para la determinación de la categoría a la que pertenece un municipio, se ha tomado como criterio la regla del 20%. Esta regla consiste en que, si un territorio tiene más de un 20% de encuestas que se encuentran en Inseguridad Alimentaria Severa, se le asignará dicha categoría. En caso contrario, se deberá comprobar cuántas encuestas pertenecen al siguiente nivel, es decir, a la Inseguridad Alimentaria Moderada. Si supera el 20% de encuestas, se le asignará esta última categoría, y así de manera sucesiva.



País	Seguridad alimentaria	Inseguridad leve	Inseguridad moderada	Inseguridad severa
El Salvador	41.928	218.556	38.846	1.686
Guatemala	57.669	605.088	266.975	34.021
Honduras	45.920	275.935	62.945	14.379
Nicaragua	181.075	370.373	63.653	16.669
Total	326.592	1.469.952	432.419	66.755

Ilustración 4. Vista de la Categoría SAN real en el espacio de Mapas.

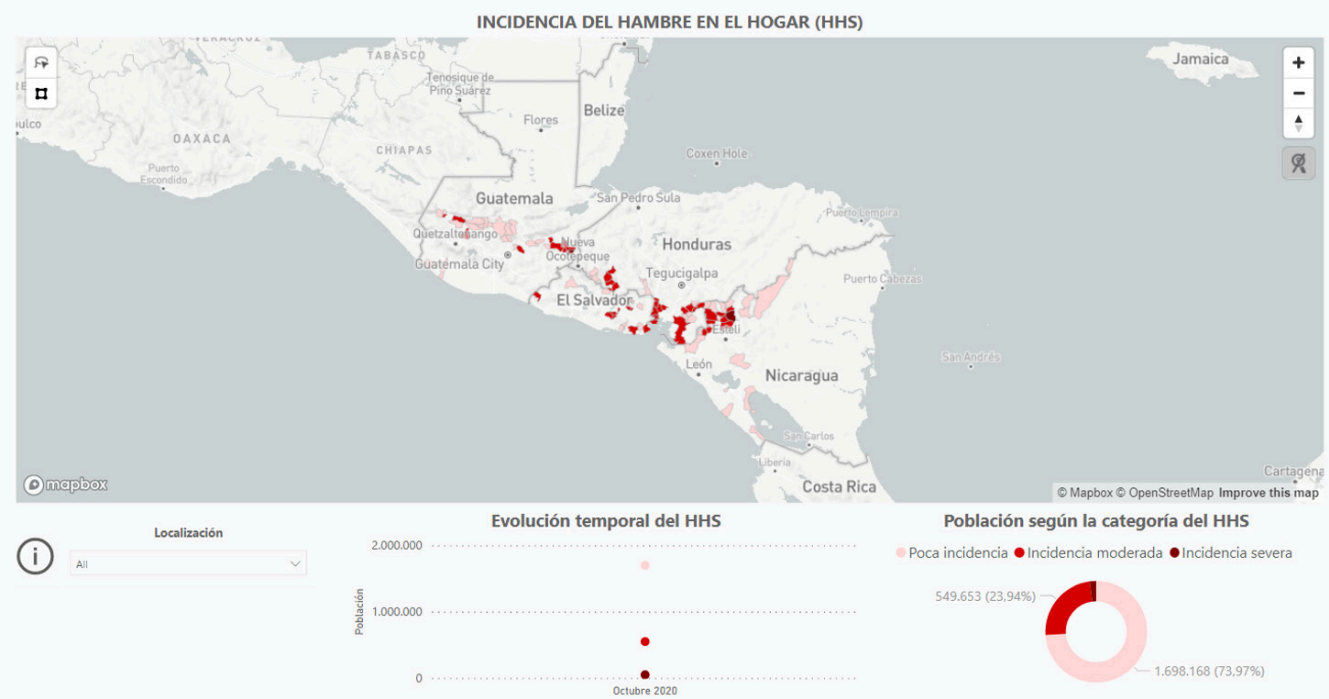


Ilustración 5. Vista de la Incidencia del Hambre en el Hogar (HHS) en el espacio de Mapas.

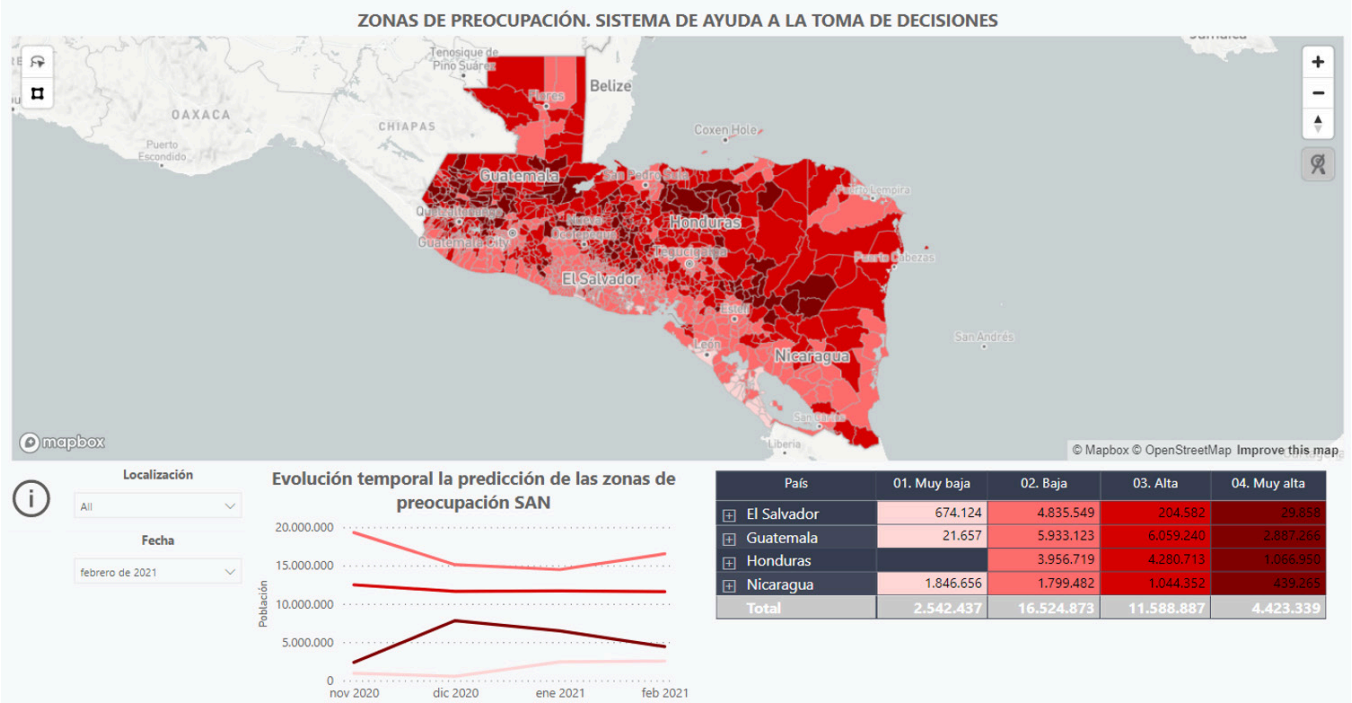
- **Índices parciales**

  Como índices parciales se muestran el Puntaje de Consumo de Alimentos (PCA), la Proporción de Gasto en Alimentos (PGA), el Índice de Estrategias de Afrontamiento (IEA) y la Incidencia del Hambre en el Hogar (HHS). De manera general, se muestra la categoría de los municipios.

SAN predicha

En esta categoría se muestran las zonas de preocupación obtenidas a partir de la metodología propuesta por GIS4Tech. De manera similar a los casos anteriores, la información aparece en pantalla mostrando un mapa en el que se pueden ver los municipios con la categoría a la que se estima que pertenecen de acuerdo con la metodología creada. De acuerdo con dicha metodología, un municipio puede pertenecer a 4 categorías según el nivel de preocupación: muy baja, baja, alta o muy alta.

El gráfico inferior muestra la evolución temporal de la población perteneciente a cada una de las categorías, y la tabla que se sitúa a la derecha recoge el desglose territorial de la población perteneciente a cada uno de esos niveles.



País	01. Muy baja	02. Baja	03. Alta	04. Muy alta
El Salvador	674.124	4.835.549	204.582	29.858
Guatemala	21.657	5.933.123	6.059.240	2.887.266
Honduras		3.956.719	4.280.713	1.066.950
Nicaragua	1.846.656	1.799.482	1.044.352	439.265
Total	2.542.437	16.524.873	11.588.887	4.423.339

Ilustración 6. Vista de las Zonas de Preocupación predichas en el espacio de Mapas.

1.3 Datos de encuestas SAN

Este cuadro de mando recoge la información de los indicadores parciales de acuerdo con los resultados de las encuestas. Se puede realizar un desglose a nivel territorial, de modo que se permite visualizar los datos por país, departamento y municipio.

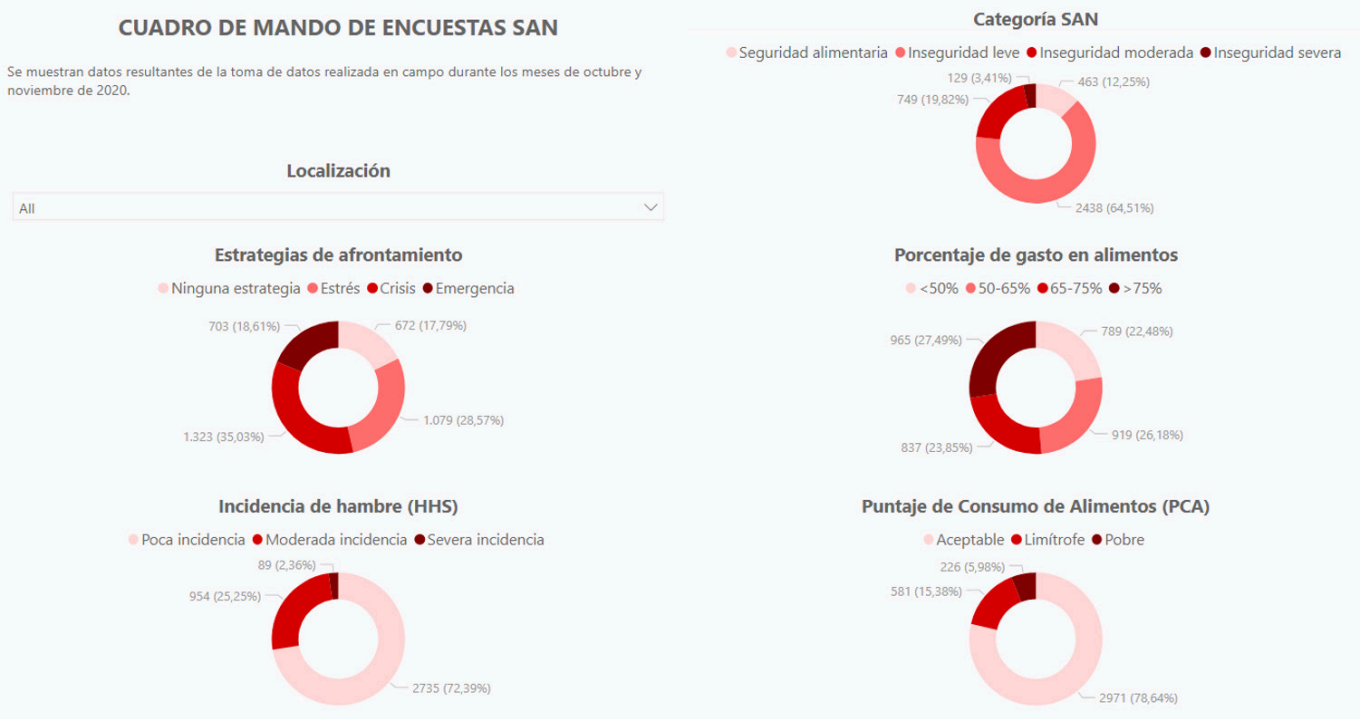


Ilustración 7. Vista del cuadro de mandos de las encuestas SAN en el espacio Datos de encuestas SAN.

A diferencia de los casos anteriores, en los que para los indicadores parciales y la categoría SAN se realizaba una extrapolación basada en los resultados de las encuestas, en este caso se muestra exactamente el número de encuestas de acuerdo con los indicadores obtenidos.

1.4 Espacio de reportes

En este apartado se pueden incluir reportes de especial relevancia para la temática abordada en la Plataforma, así como los enlaces a los mismos para que puedan ser descargados por el público en general.

Ilustración 8. Vista del Espacio de Reportes.

2

FUENTES DE DATOS

2. Fuentes de datos

2.1 Datos socioeconómicos

La mayoría de los datos han sido extraídos de repositorios públicos como los respectivos organismos oficiales de estadística de cada país, de fuentes como **Our World in Data, Humanitarian Data Exchange.**

2.2 Datos agroclimáticos

A continuación, se muestra el listado de variables agroclimáticas utilizadas para la caracterización y la evaluación del estado de los cultivos.

Índice	Definición	Fuente satelital	Resolución (metros)
SPI	**Standardized Precipitation Index o Índice Estandarizado de Precipitación.** Indicador utilizado para la caracterización de sequías meteorológicas. Mide el número de desviaciones típicas que se aleja la precipitación de un periodo observado con respecto al resto de la serie histórica	CHIRPS	5.000
NDVI	**Normalized Difference Vegetation Index o Índice Normalizado de Diferencia de Vegetación.** Índice empleado para estimar la cantidad, calidad y desarrollo de la vegetación, asistir en la predicción de zonas vulnerables a incendios, entre otros usos.	Earthdata	250
EVI	**Enhanced Vegetation Index o Índice de Vegetación Mejorado.** Utilizado para la monitorización del estado de la vegetación en zonas con altos índices de biomasa	Earthdata	250
ARVI	**Atmospherically Resistant Vegetation Index o Índice de Vegetación Atmosféricamente Resistente.** Este índice se emplea en zonas con alto contenido de aerosoles atmosféricos, y corrige el efecto de dispersión que produce este fenómeno en la reflectancia de la luz.	Earthdata	250
SIPI	**Structure Insensitive Pigment Index o Índice de Insensibilidad a la Estructura del Pigmento.** Sirve para la evaluación del estado de los cultivos en zonas con una cobertura arbórea variable.	Earthdata	250

2.3 Datos de campo SAN

Los datos derivados de las encuestas realizadas en campo introducidos en la plataforma son los siguientes:

Índice	Posibles valores
Puntaje de Consumo de Alimentos (PCA)	Aceptable/Limítrofe/Pobre
Índice de Estrategias de Afrontamiento (IEA)	Ninguna estrategia/Estrategias de estrés/Estrategias de crisis/Estrategias de emergencia
Incidencia del Hambre en el Hogar (HHS)	Poca incidencia/Incidencia moderada/Incidencia severa
Proporción de Gasto en Alimentos (PGA)	<50% / 50 – 65% / 65 – 75% / >75%
Categoría SAN	Seguridad Alimentaria / Inseguridad Alimentaria Leve / Inseguridad Alimentaria Moderada / Inseguridad Alimentaria Severa

3

CÁLCULO DE ÍNDICES AGROCLIMÁTICOS

3. Predicción de las zonas de preocupación

Los datos necesarios para la alimentación de la plataforma consisten en un conjunto de información agroclimática y otro de tipo socioeconómico. Por lo que respecta a los datos de tipo agroclimático, en los siguientes apartados se procederá a mostrar cómo obtener dicha información.

Para la generación de este tipo de información, se empleará el software QGIS, que permite realizar operaciones y analizar datos con componente geoespacial.

3.1 Instalación de QGIS

Para realizar la instalación de QGIS, debemos acceder a la página oficial **http://www.qgis.org/es/site/** .

Ilustración 9. Vista de la web de QGIS

Una vez aquí, seleccionamos “Descargar ahora”. Nos aparecerán las siguientes opciones:

Ilustración 10. Opciones de descarga de QGIS para diferentes sistemas operativos.

Podemos optar por dos vías diferentes para realizar la instalación, mediante el instalador de QGIS, o mediante el instalador de OSGeo:

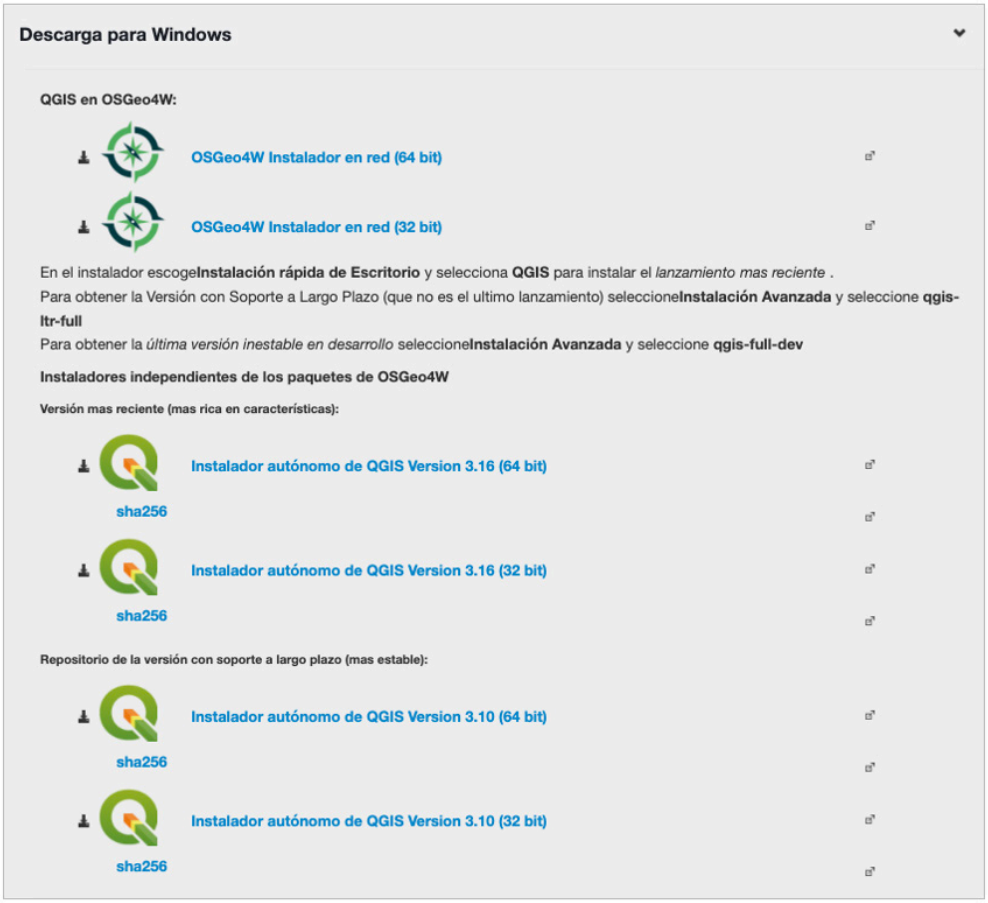
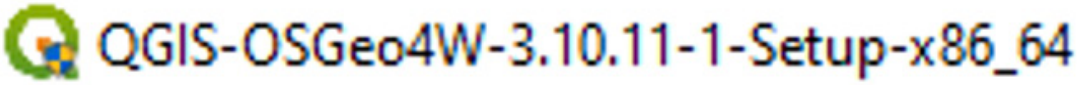


Ilustración 11. Opciones de descarga de QGIS.

a. Instalador autónomo de QGIS (recomendada)

Podríamos elegir la última versión de QGIS (en este caso, la versión 3.16), según nuestro sistema operativo y las características de éste.

Para el desarrollo del presente curso, recomendamos la instalación de la versión 3.10.



Una vez descargado el instalador, se instala siguiendo los pasos habituales a la hora de instalar cualquier software.

En el caso de disponer de una versión de QGIS instalada previamente en el ordenador, resulta necesario desinstalarla por completo antes de proceder a instalar la nueva versión deseada.

3.2 Cálculo de índices con MODIS

MODIS (Moderate Resolution Imaging Spectroradiometer) es un instrumento que se encuentra a bordo de los satélites Terra y Aqua, ambos pertenecientes a la NASA. El satélite Terra, que es el que nos ocupa, orbita alrededor de la Tierra de norte a sur a través de la línea del Ecuador cada 1 ó 2 días.

El objetivo de este instrumento es mejorar la comprensión sobre las dinámicas globales a nivel climático tanto en zonas terrestres como en los océanos. Se emplea de manera habitual para la monitorización de grandes masas de vegetación debido a su resolución espacial (desde 250 metros).

3.2.1. Descarga de imágenes desde EARTHDATA

Para la obtención de imágenes de MODIS, se recurrirá al sitio web **Earthdata**, un repositorio en el que se pueden encontrar imágenes satelitales de múltiples tipos, resoluciones, proveedores y áreas temáticas.

Para poder descargar las imágenes que necesitamos para seguir los siguientes pasos descritos en este manual, es necesario registrarse en la plataforma de Earthdata. Para ello, debemos ir en primer lugar al apartado "FIND DATA", que nos dirigirá al visor en el que podremos visualizar las imágenes satelitales previamente a su descarga.

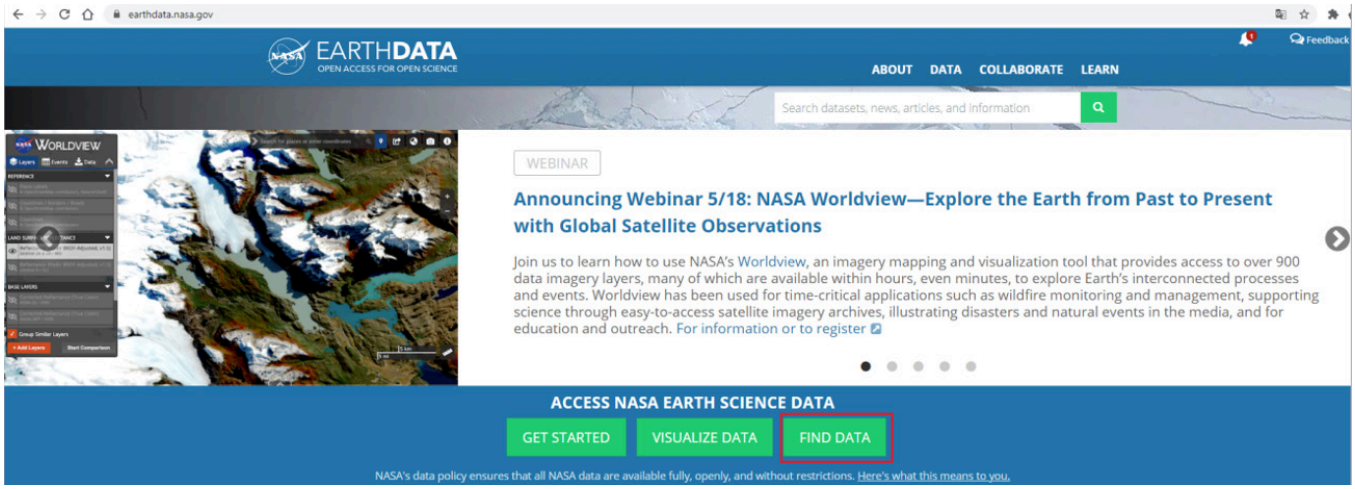


Ilustración 12.Vista inicial de la plataforma Earthdata.

Una vez en el visor, nos dirigiremos a la parte superior derecha, en la que aparece la opción "Earthdata Login". A continuación, la página web nos dirigirá a un formulario de registro en el que cada uno deberá introducir su información personal y sus credenciales de acceso a la plataforma Earthdata.

Una vez que hemos finalizado el proceso de registro, nos dirigimos a la caja de búsqueda ubicada en la parte superior izquierda de la pantalla. En esta caja, buscaremos el producto "MOD13Q1", correspondiente a las imágenes de 250 metros de resolución del satélite Terra.

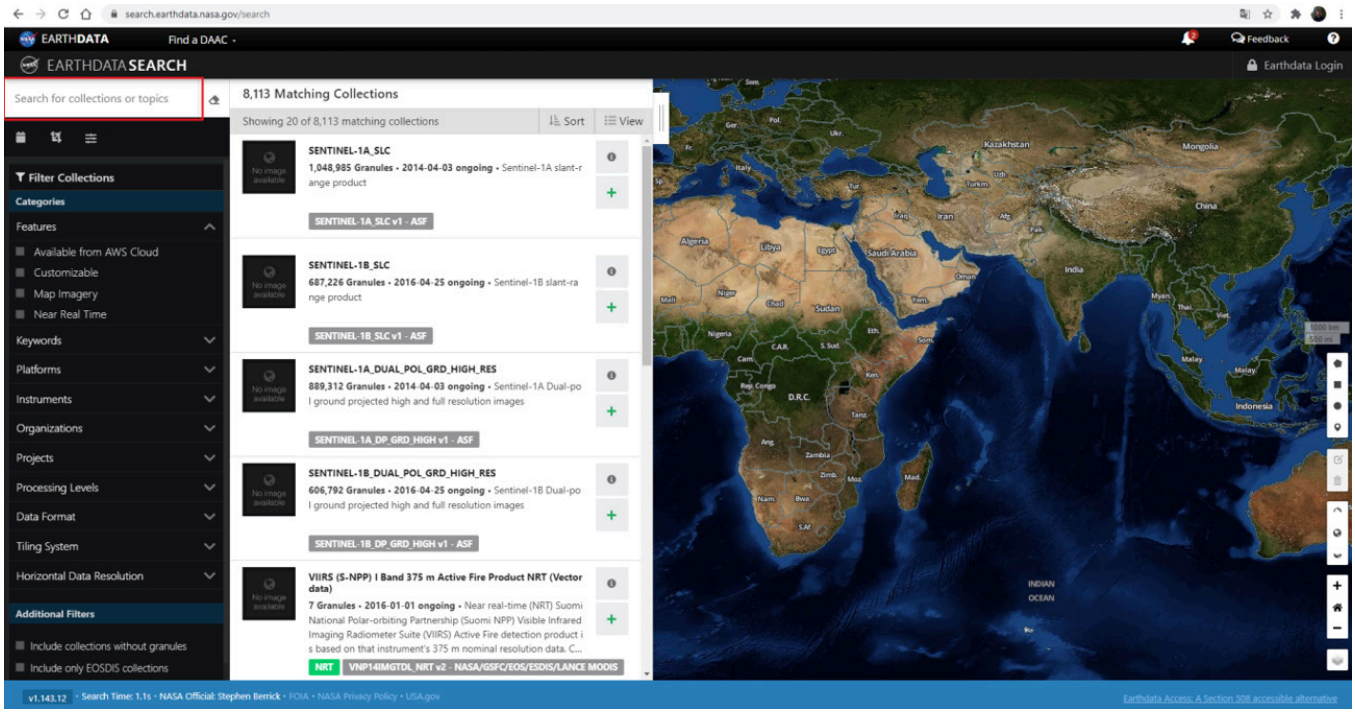


Ilustración 13. Ubicación de la caja de búsqueda en la plataforma Earthdata.

La plataforma nos devolverá una serie de productos que coincidirán con nuestros criterios de búsqueda, pero para acotar aún más, debemos introducir el ámbito geográfico del cual queremos descargar las imágenes. Para ello, nos dirigimos al icono señalado en la siguiente ilustración, y colocamos el puntero en cualquier país de CA4. La búsqueda se reducirá a dos opciones, y en nuestro caso elegiremos la denominada "MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006".

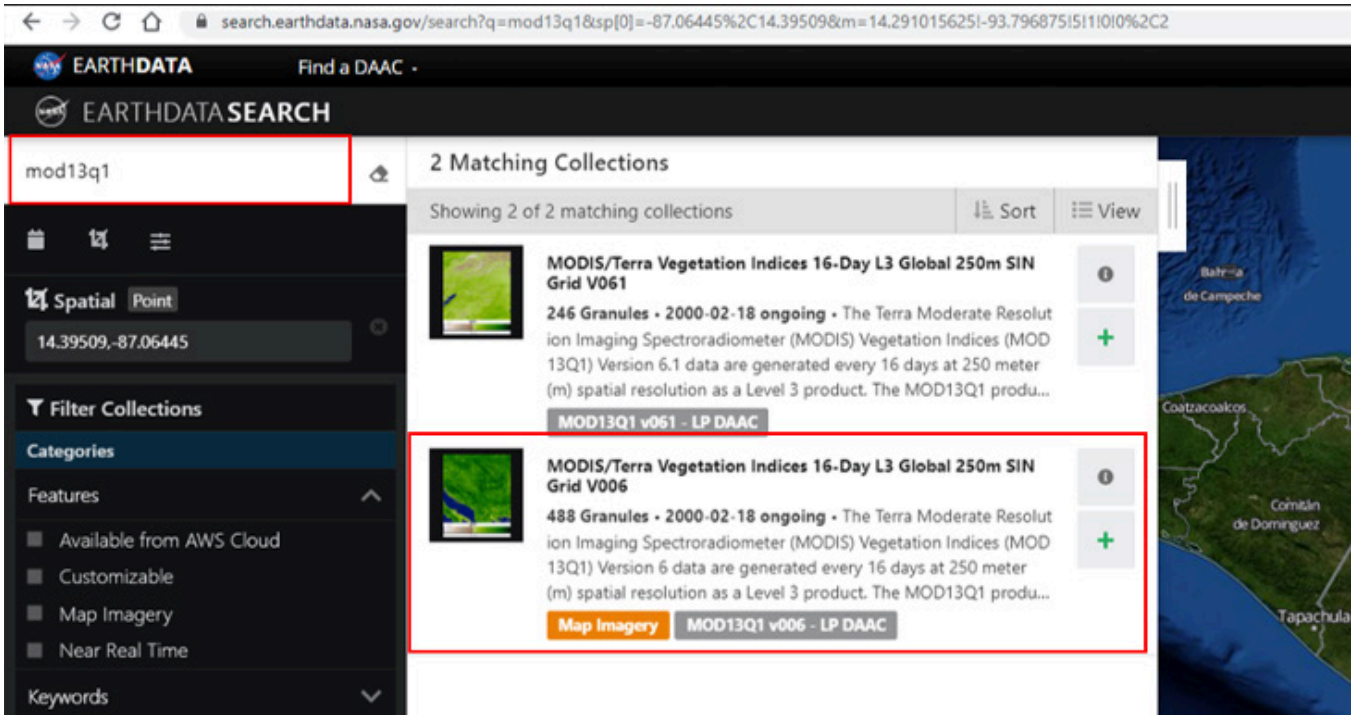


Ilustración 14. Selección de productos para la descarga de imágenes desde Earthdata.

Seleccionamos la última imagen disponible, que es la situada en la parte superior izquierda del panel de imágenes. Hacemos clic sobre el icono “+” para añadir dicha imagen al listado de descargas, y a continuación nos dirigimos al botón “Download”, ubicado en la parte inferior del panel de descargas.

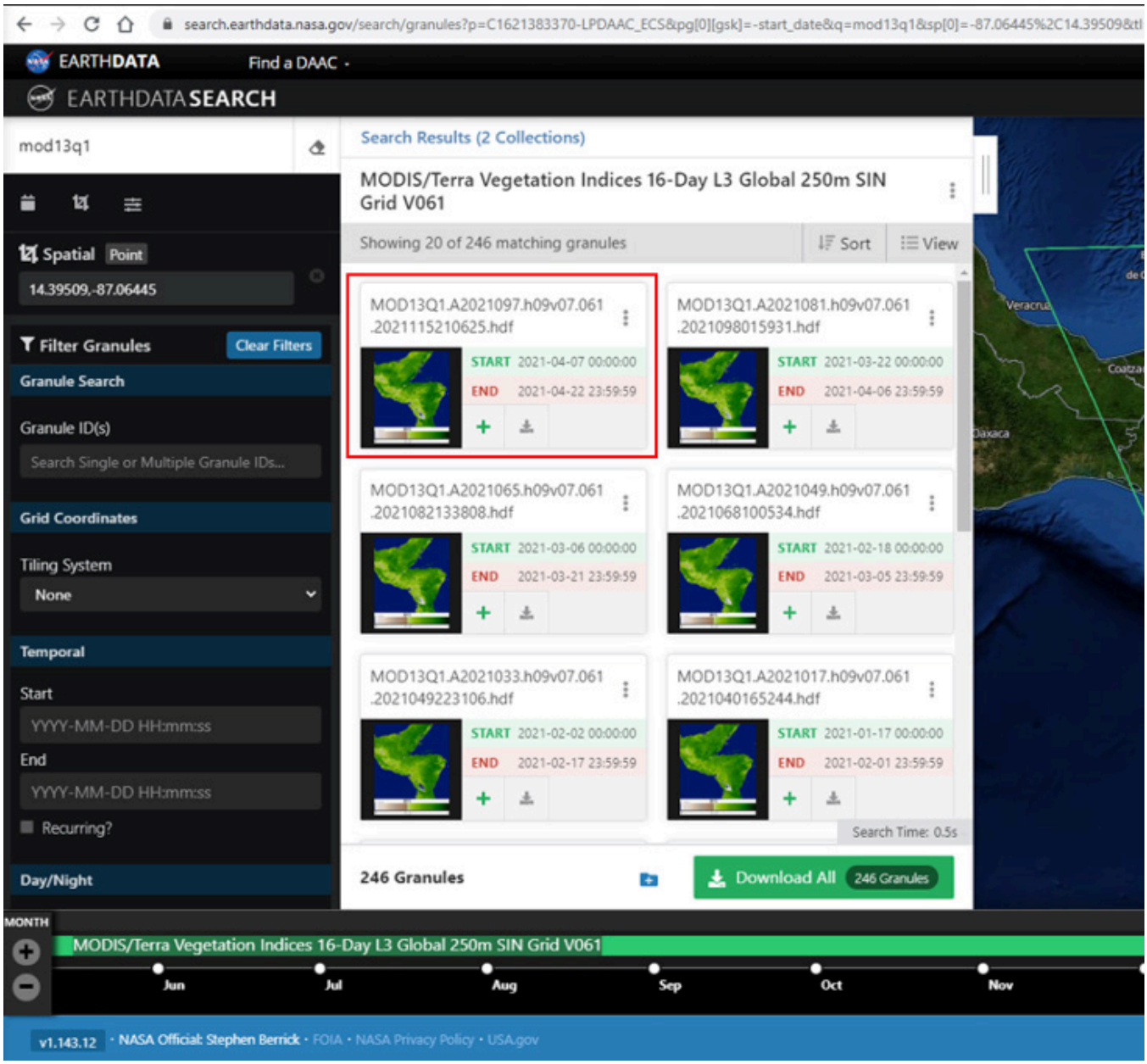


Ilustración 15. Selección de imagen para la descarga.

En la siguiente pantalla se mostrará un resumen de los datos de la descarga, incluyendo el número de imágenes seleccionadas, la cantidad de colecciones o el tamaño de la propia imagen. Pulsamos el botón “Download Data” ubicado en la parte inferior del panel, y la página nos dirigirá a una vista en la que se encuentra el enlace de la imagen seleccionada.

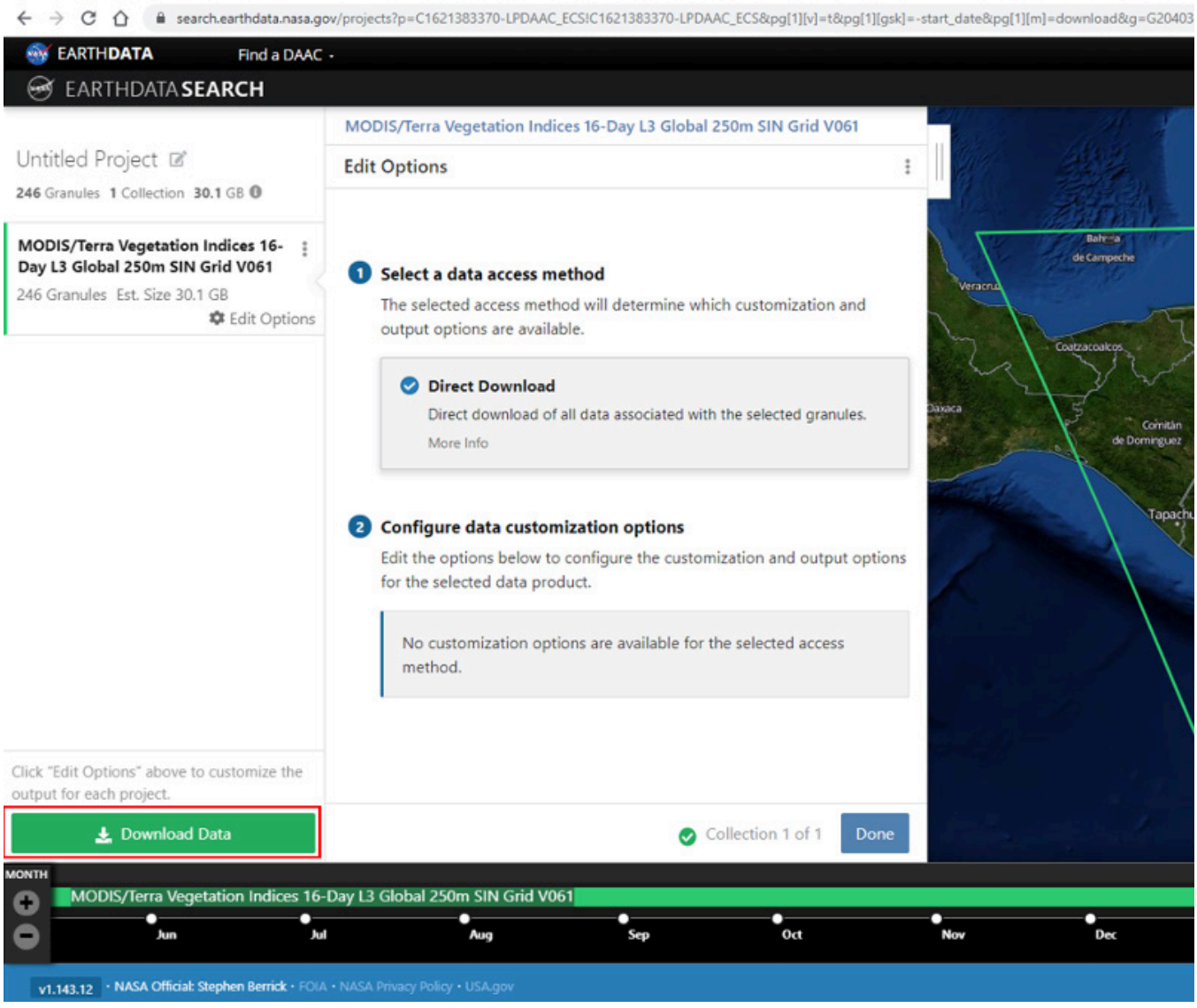


Ilustración 16. Vista previa de las descargas.

Si el proceso de registro se realizó con éxito en fases previas de este manual, a continuación, únicamente debemos hacer clic sobre el enlace mostrado en pantalla para descargar la imagen en la ruta que deseemos.

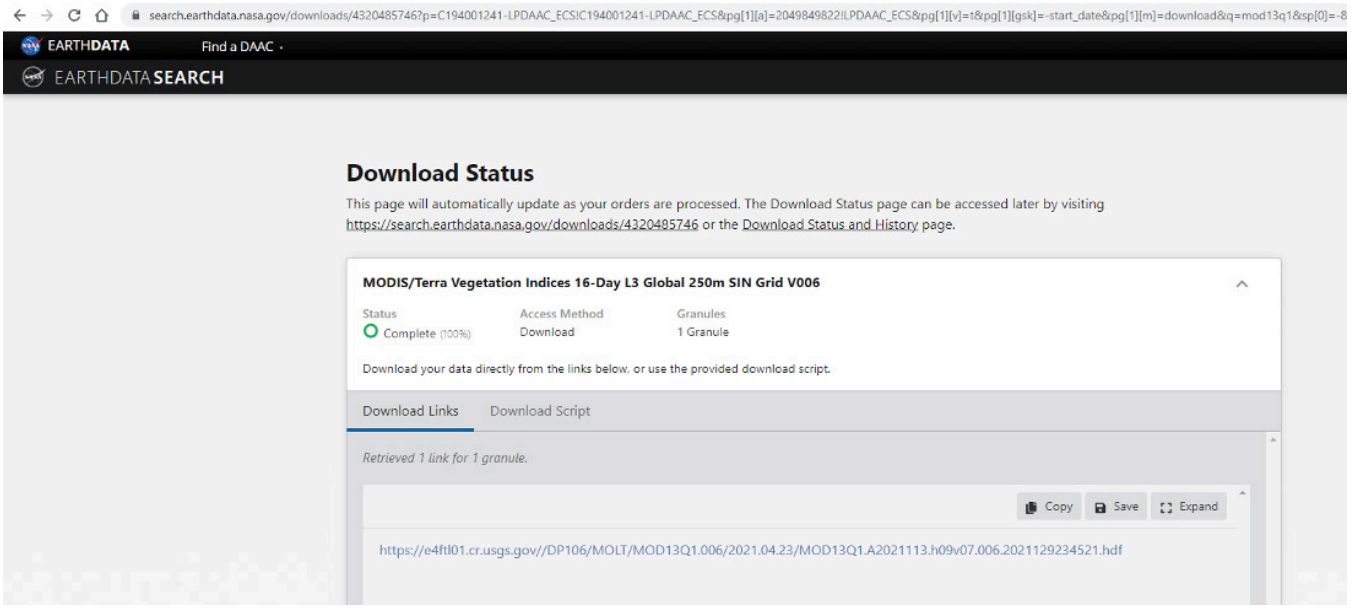


Ilustración 17. Vista de los enlaces de descarga.

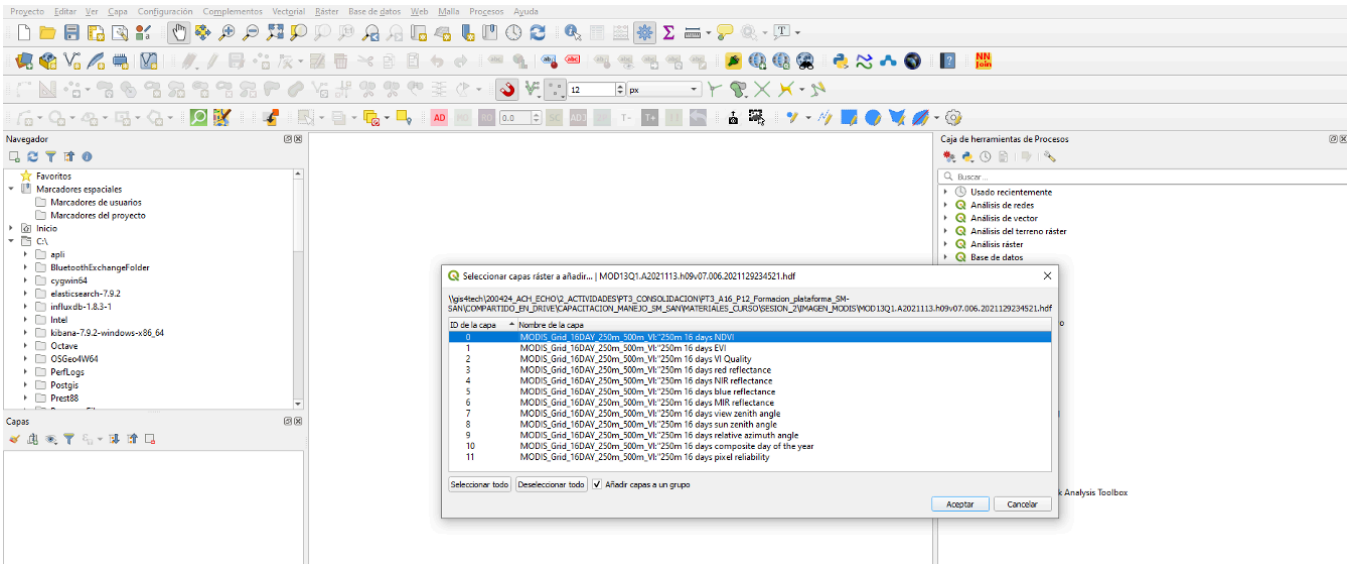


Ilustración 18. Vista inicial de carga de la imagen en QGIS.

Para hacernos una idea del contenido de la información descargada, abrimos el software QGIS y arrastramos la imagen descargada al visor. Seleccionamos la primera opción, que es la que hace referencia al índice NDVI, y vemos que se representan unos valores de -2e+07 a 9.99e+07. En apartados posteriores de este curso se explicará cómo procesar esta imagen para que los valores resultantes sean interpretados de manera más sencilla.

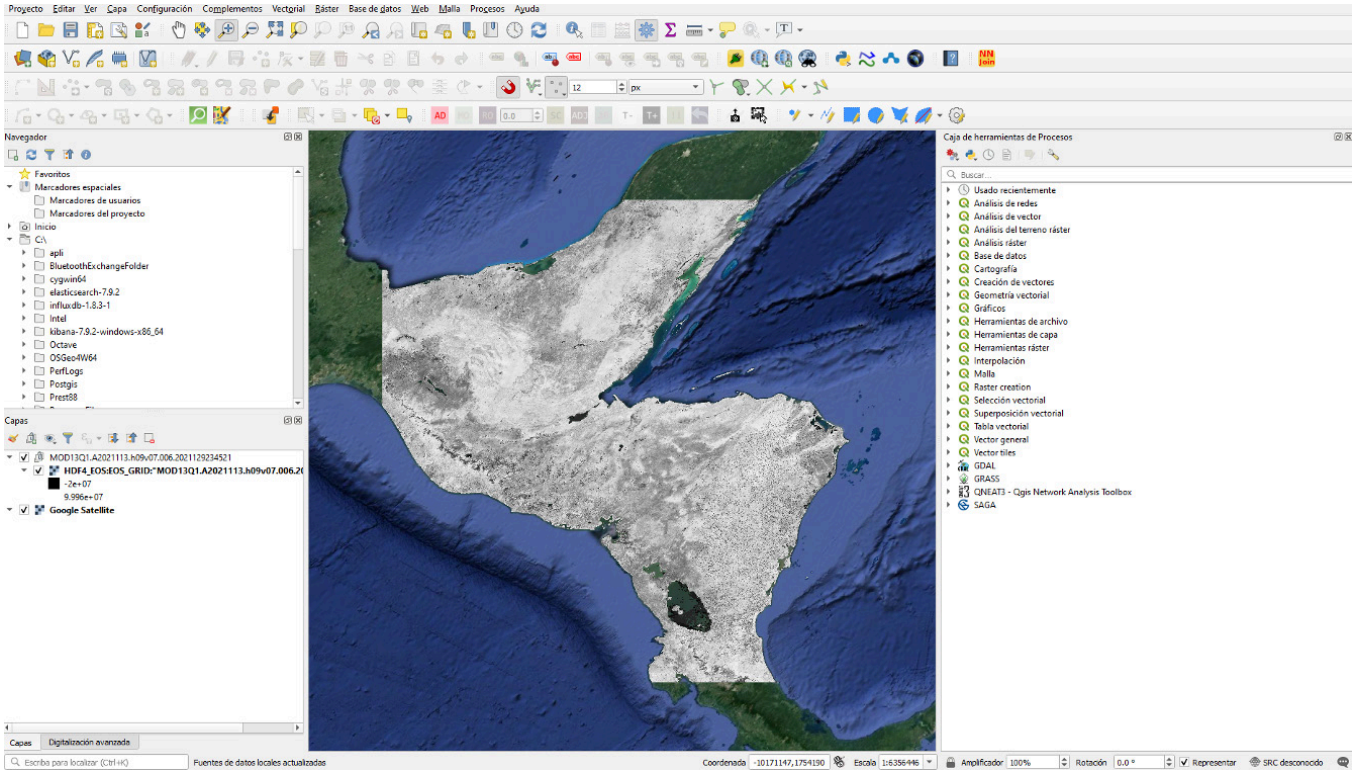


Ilustración 19. Vista de la imagen descargada de la plataforma Earthdata.

3.2.2. Cálculo del índice NDVI

En el ámbito del estudio de la teledetección, el índice más comúnmente empleado es el denominado NDVI (Normalized Difference Vegetation Index). Puede señalarse como umbral crítico para cubiertas vegetales un valor de NDVI de 0.1 y para la vegetación densa de 0.5. Cualquier valor negativo implica ausencia de vegetación por lo que debieran reclasificarse a cero.

En el caso particular de MODIS, el índice NDVI ya viene dado por el proveedor de imágenes, por lo que no se precisa realizar cálculos utilizando diferentes bandas de la imagen satelital. De este modo, el único procesamiento requerido para obtener este índice consiste en multiplicar los valores de la imagen descargada por un factor de escala.

Una vez que ya tenemos la imagen, nos dirigimos a la pestaña “Ráster” de la barra de herramientas, y en el menú contextual seleccionamos la opción “Calculadora ráster…”.

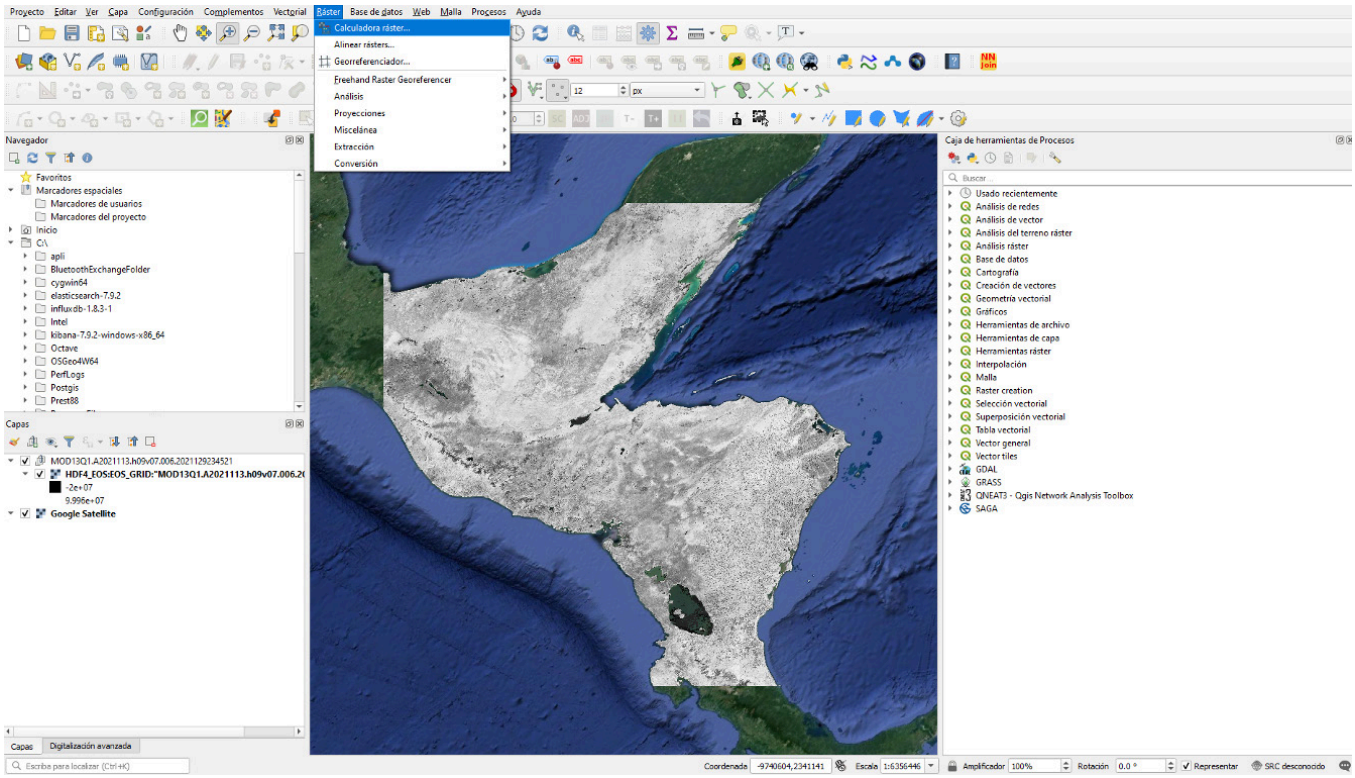
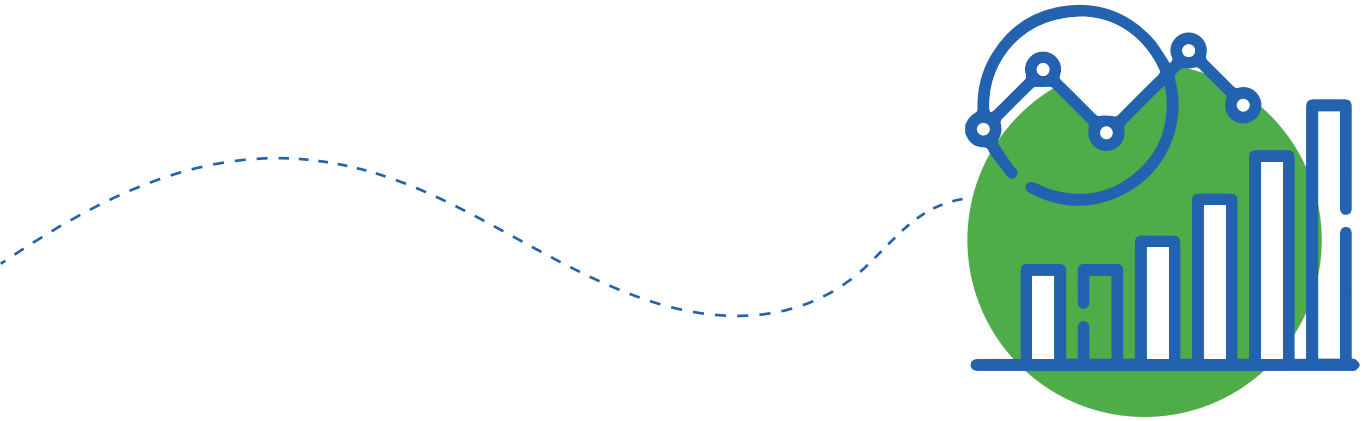


Ilustración 20. Ubicación de la herramienta "Calculadora ráster".

Como se adelantó previamente, los datos del NDVI ya vienen dados directamente en una de las bandas de la imagen descargada, aunque se deben multiplicar por un factor de escala para facilitar su interpretabilidad. Por ello, en la Calculadora ráster introduciremos la capa que queremos procesar ("NDVI_bruto"), y la multiplicamos por 0,00000001.

En la parte superior de la ventana introducimos la ruta de la capa de salida y nos aseguramos de que el formato elegido para la imagen resultante es GeoTIFF.



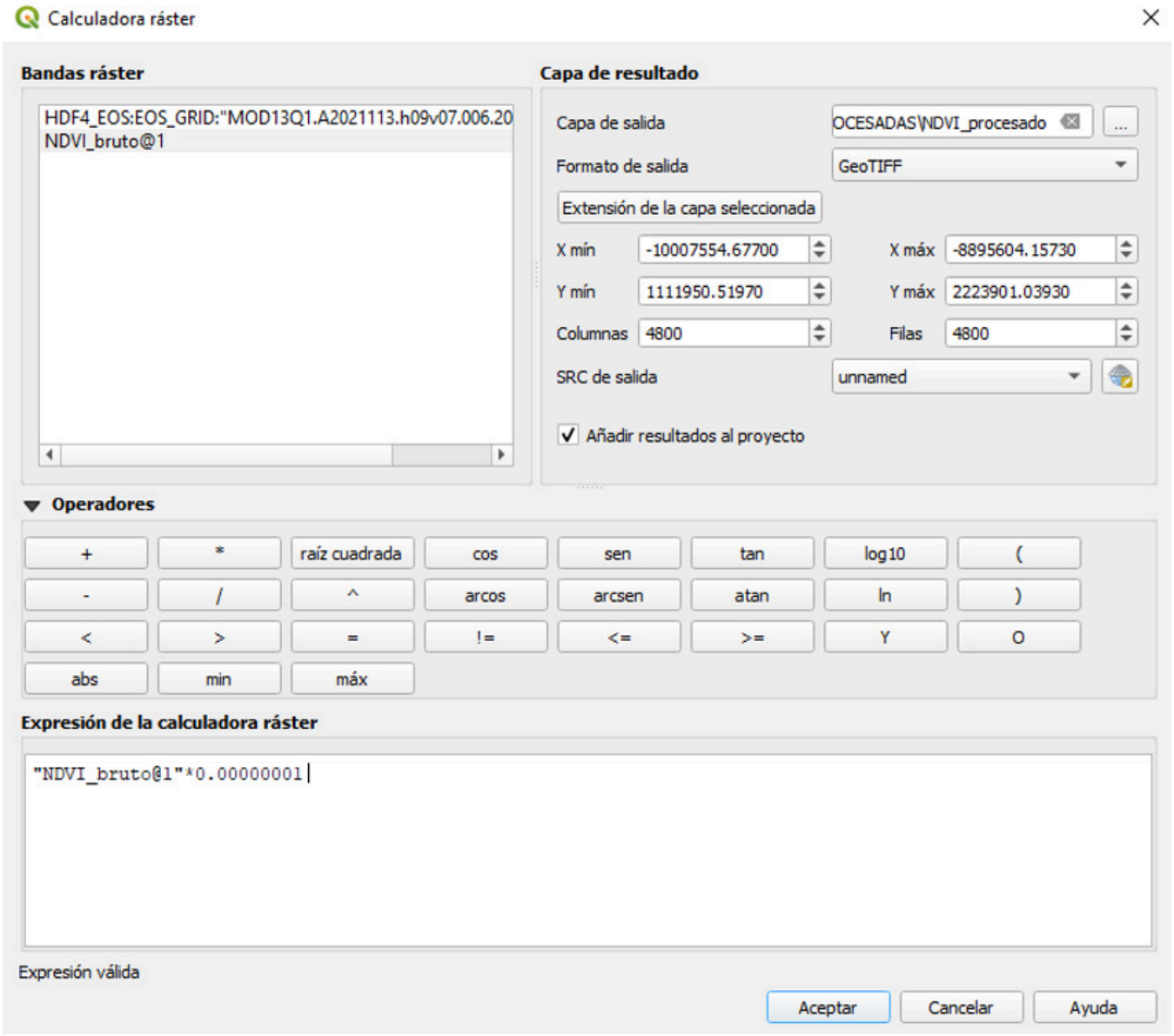


Ilustración 21. Procesamiento de la imagen NDVI en bruto.

De este modo, obtendremos una imagen denominada “NDVI_procesado”, en el que el rango de valores se moverá entre -0,2 y 0,999, lo cual facilita significativamente la interpretabilidad de los datos.

A efectos de representación, y para facilitar la comprensión del proceso realizado hasta ahora, nos disponemos a cambiar la paleta de colores de la imagen procesada.

Una vez en el menú de Propiedades de la imagen, en el panel lateral izquierdo, nos dirigimos a la pestaña de Simbología, donde cambiaremos el Tipo de renderizador de Gris monobanda a Pseudocolor monobanda.

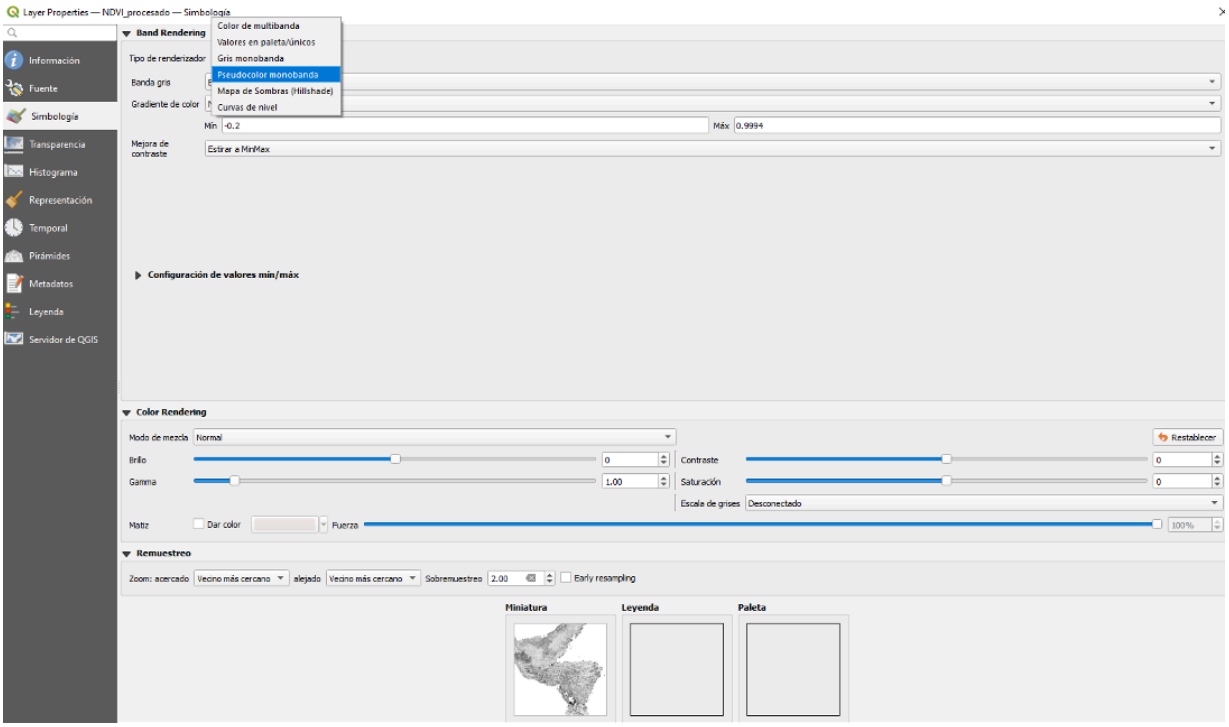


Ilustración 22. Cambio del tipo de renderizado a Pseudocolor monobanda.

El tipo de representación es una cuestión que permite cierta flexibilidad, y que se debe adaptar a las necesidades del analista. Para el presente caso, se propone dejar la paleta de colores por defecto, y en la opción “Modo”, seleccionar “Intervalo igual”. Podemos modificar los valores de corte a nivel de representación e introducir los siguientes: -0,2; 0; 0,33; 0,66 y 0,999.

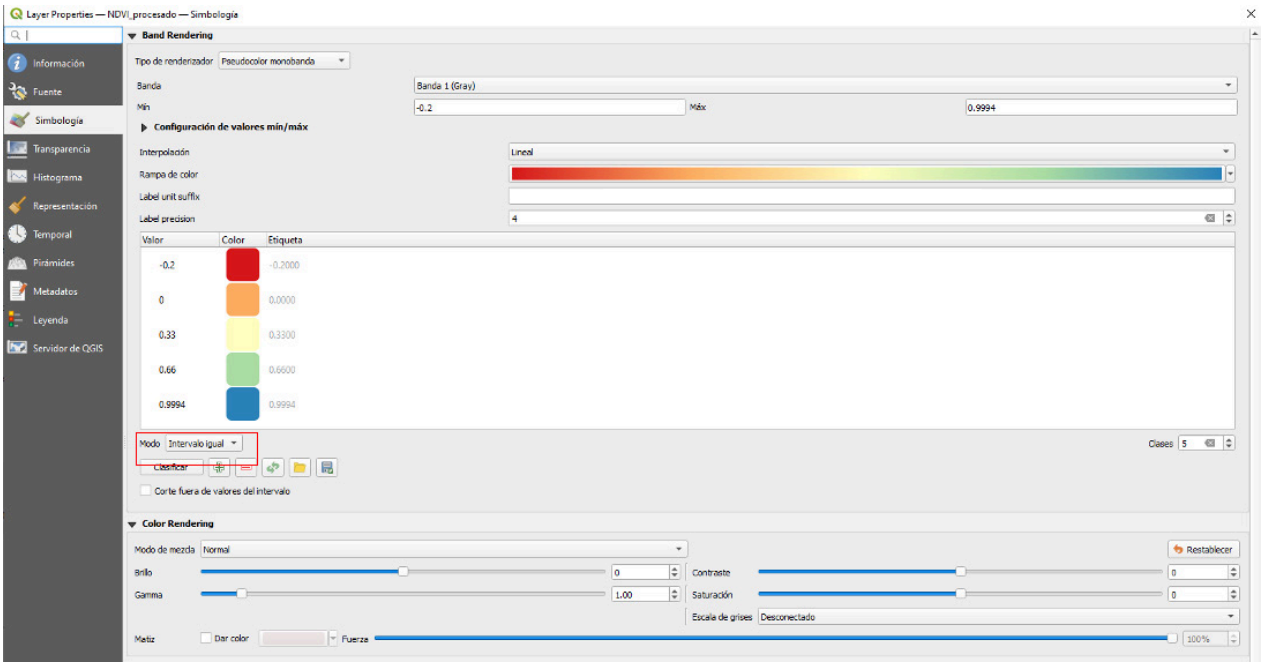


Ilustración 23. Selección de las opciones de representación de la imagen NDVI procesada.

La imagen resultante es la que se muestra a continuación. Se puede ver a simple vista que los valores más bajos, que tienen colores rojizos y anaranjados, se corresponden en gran parte con masas de agua, aunque si nos acercamos a las zonas urbanas, veremos que también coincide que los valores del NDVI son bastante bajos.

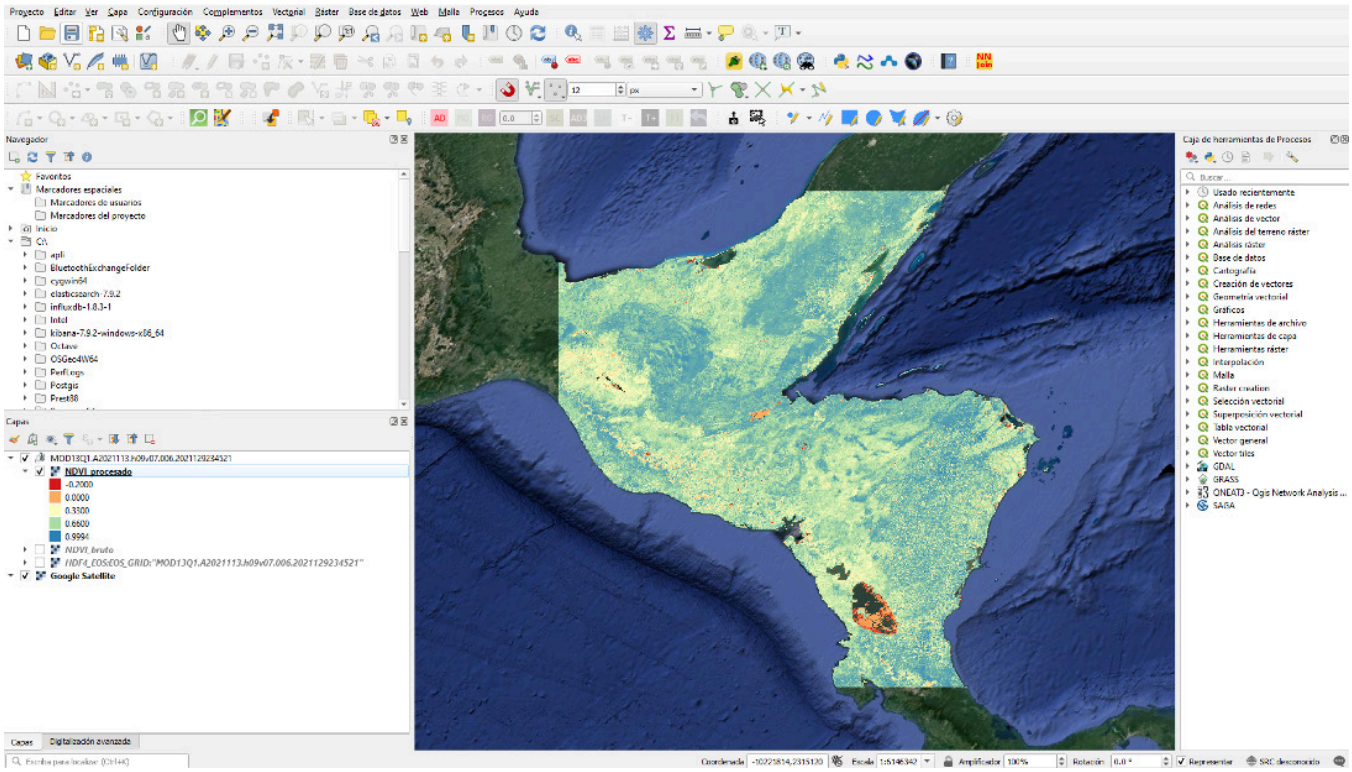


Ilustración 24. Imagen procesada con los valores del NDVI.

3.2.3. Cálculo del índice EVI

Otro de los productos que genera el sensor MODIS es el Enhanced Vegetation Index o EVI. A diferencia del NDVI, el EVI es menos sensible a la influencia del suelo y de efectos atmosféricos que el NDVI debido a que incorpora la banda relativa al espectro azul. De esta manera, el EVI es sensible al incremento de la densidad de la cobertura vegetal, especialmente en zonas donde se produce una saturación del NDVI. El EVI tiene un rango de valores que se mueve entre -1 y 1, y se suele considerar que la vegetación sana se encuentra entre 0,2 y 0,8.

Para la obtención de este índice a partir del sensor MODIS, no se requiere realizar ningún cálculo. Al igual que en el caso del NDVI, la única operación que se debe realizar es la multiplicación de la imagen provista por un factor de escala.

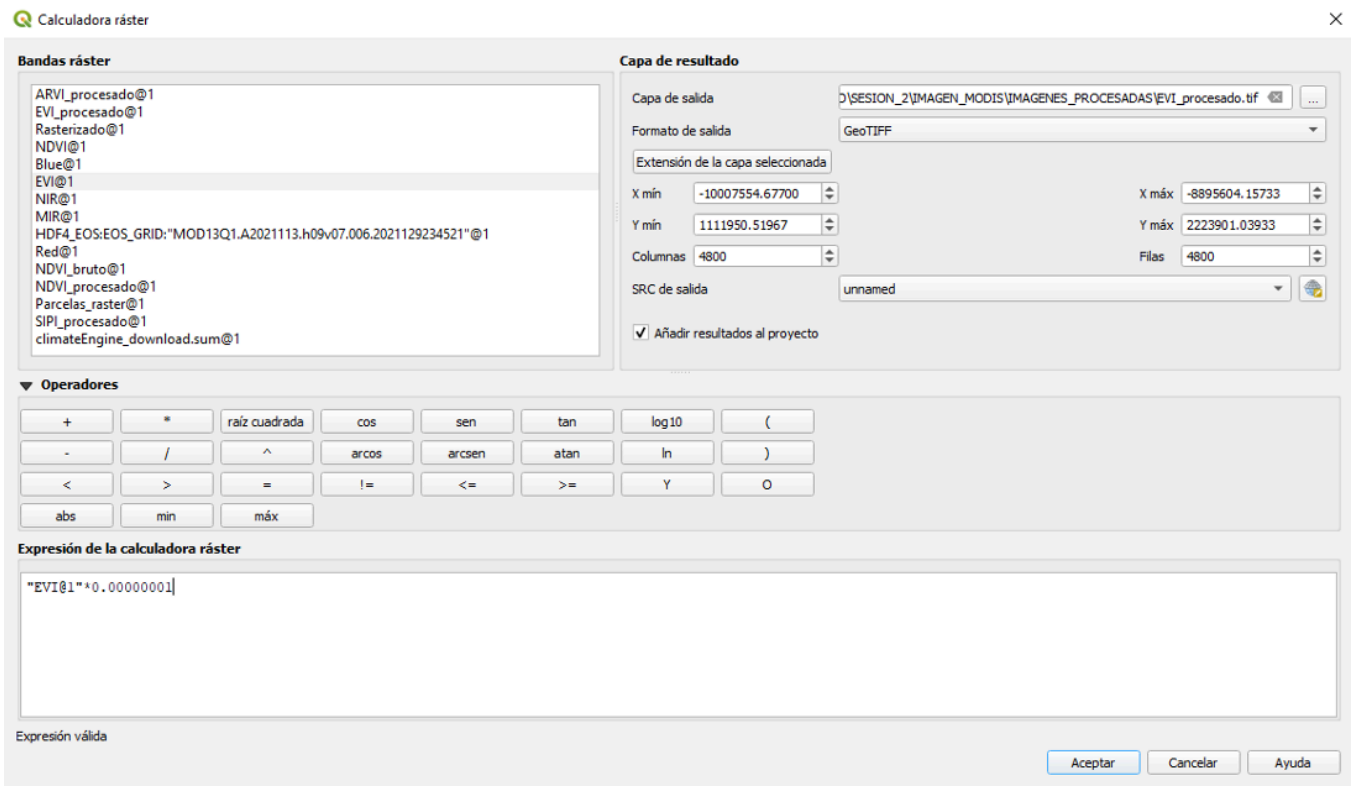


Ilustración 25. Procesamiento de la imagen EVI en bruto.

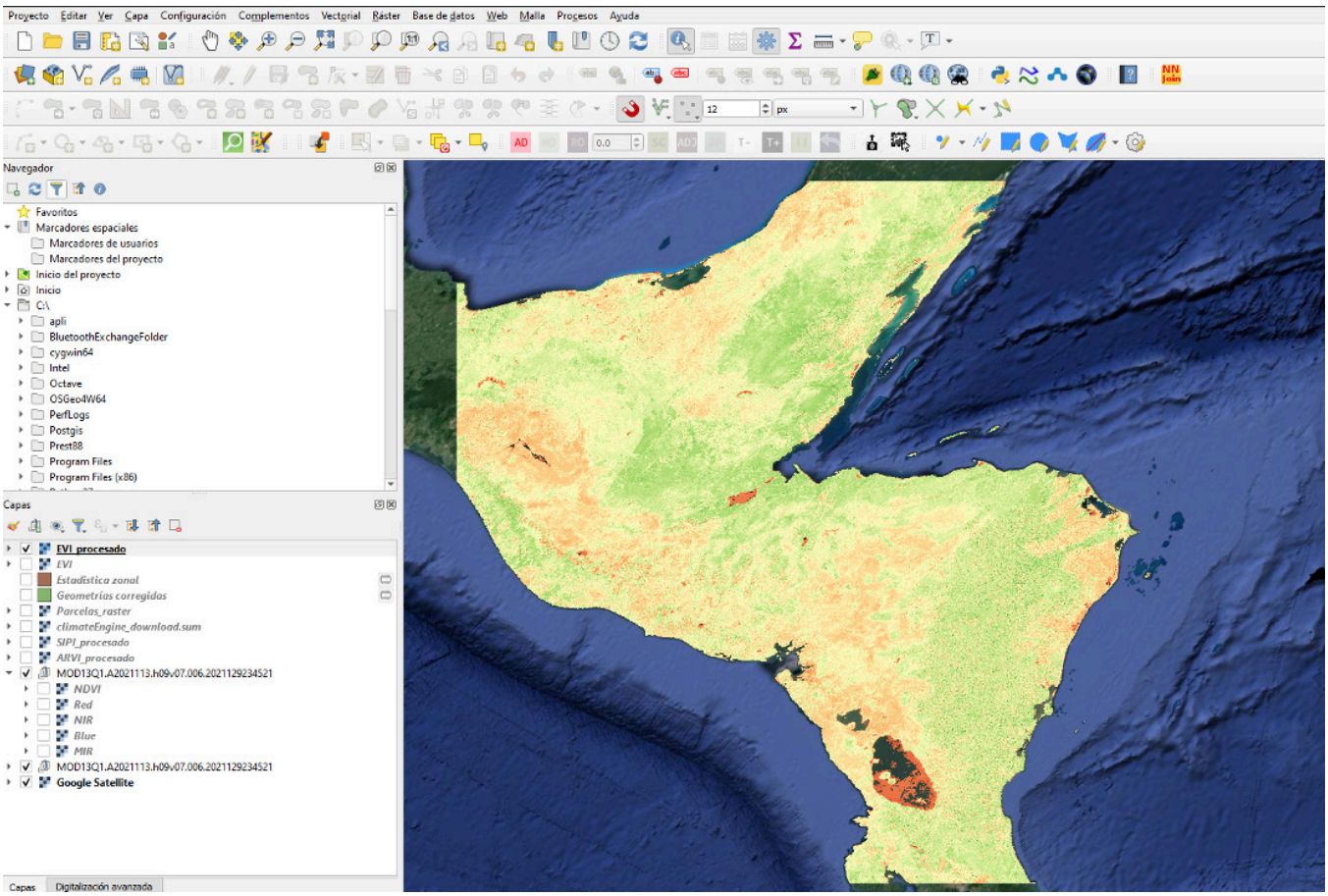


Ilustración 26. Imagen procesada con los valores del EVI.

3.2.4. Cálculo del índice ARVI

A diferencia de los índices anteriores, el índice ARVI sí requiere realizar una operación entre las diferentes bandas de la imagen original, que se corresponde con la siguiente expresión:

$$\frac{NIR - (2 * R - B)}{NIR + (2 * R - B)}$$

Para la realización de este cálculo con QGIS, deberemos dirigirnos de nuevo a la calculadora ráster e introducir la fórmula de acuerdo con la expresión anterior. De nuevo, seleccionamos una ruta para el guardado de esta imagen y nos aseguramos de que el formato elegido es GeoTIFF. Introducimos la fórmula del índice ARVI en la consola, y hacemos clic en Aceptar.

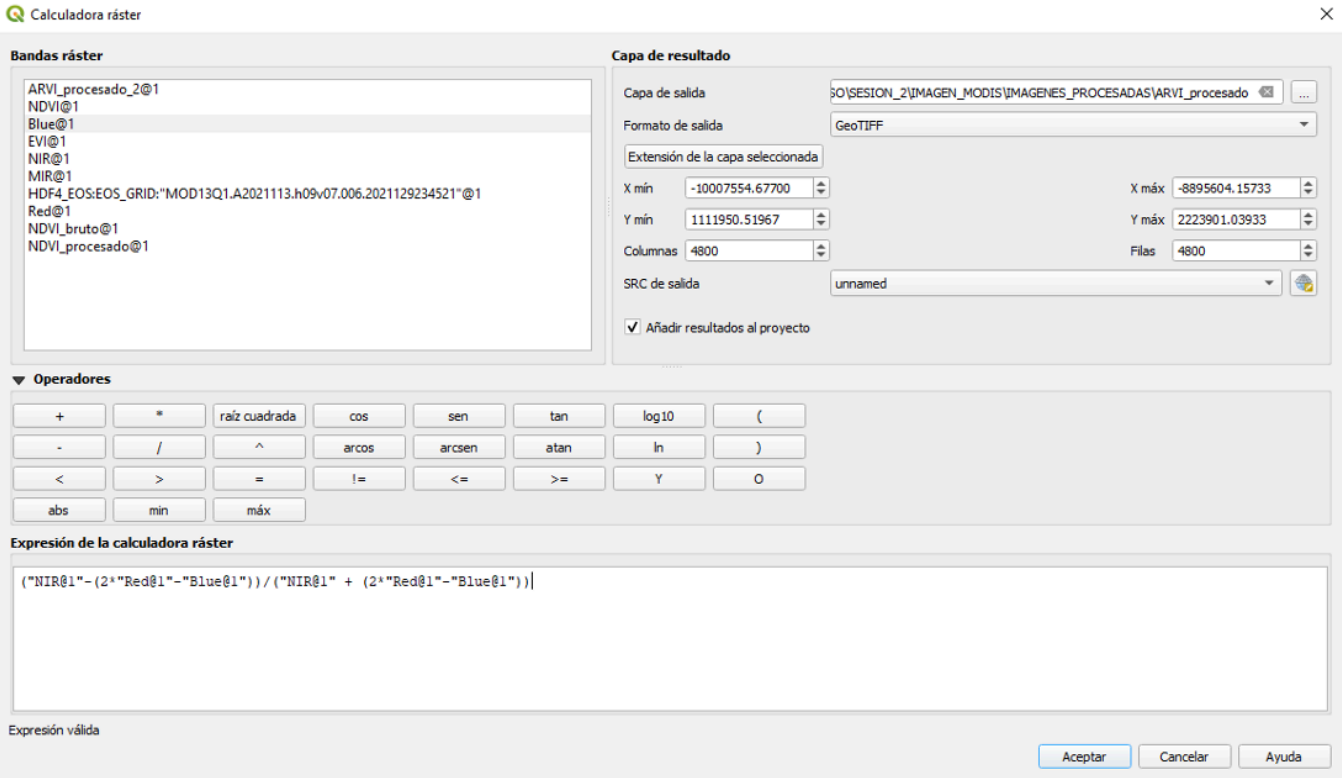


Ilustración 27. Cálculo del ARVI mediante la calculadora ráster.

De acuerdo con la literatura, el ARVI muestra la vegetación para unos índices de entre 0,2 y 0,8. Los valores por debajo de 0,2 suelen coincidir con masas de agua o zonas sin vegetación, mientras que los valores superiores a 0,8 tienden a mostrar masas arbóreas y zonas de vegetación de alta densidad.

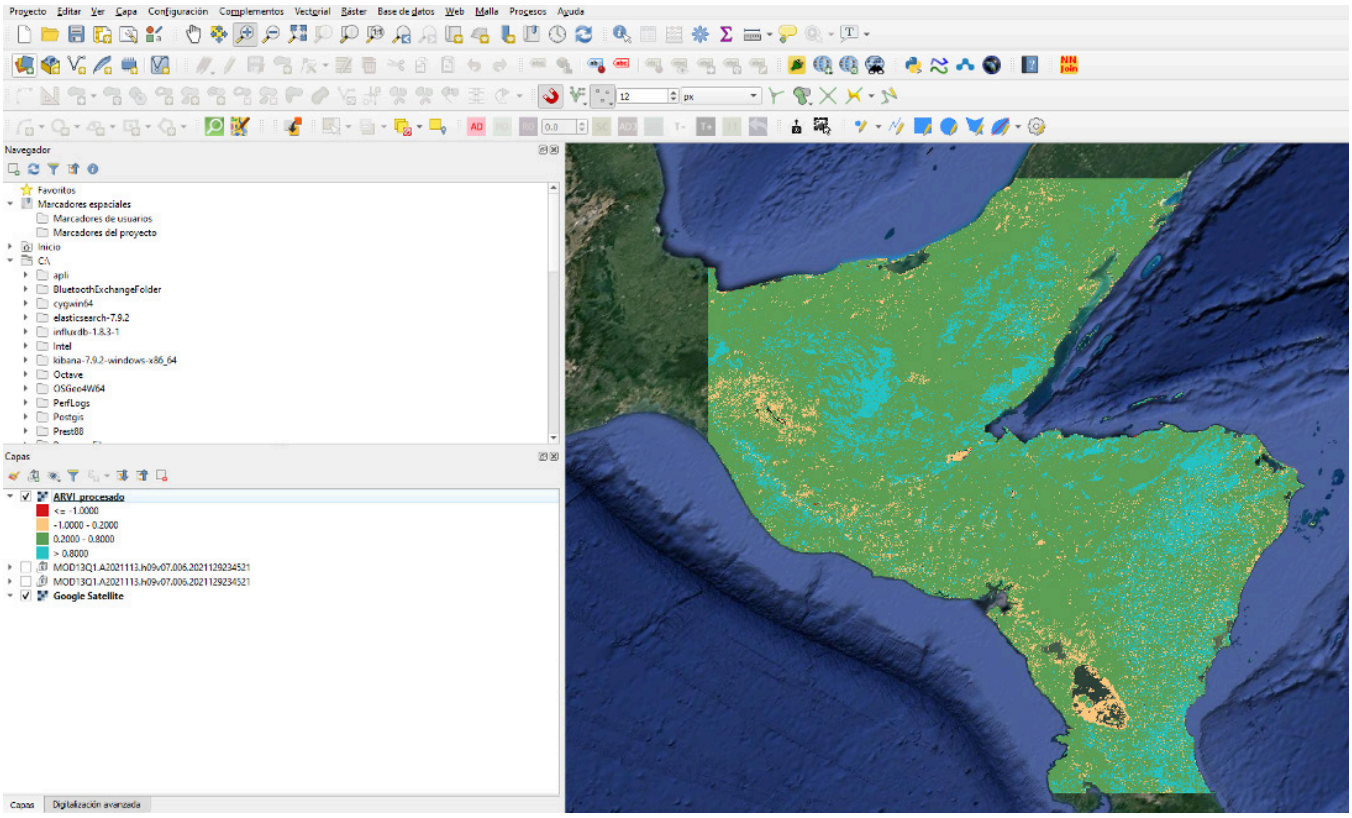


Ilustración 28. Valores del ARVI representados en 4 intervalos, representando masas de agua y zonas urbanas, vegetación y masas arbóreas

A modo de ejemplo, en la siguiente imagen se puede ver cómo, siguiendo este criterio, podemos identificar de manera aproximada dónde están las zonas urbanas y algunas masas de agua urbanas como los lagos o los humedales en contraste con las zonas donde se pueden apreciar las explotaciones agrícolas.

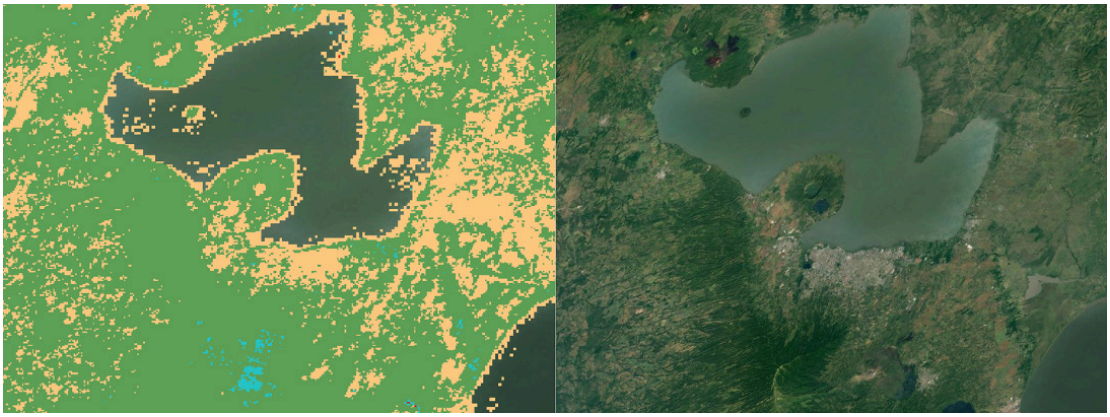


Ilustración 29. Imagen comparativa de los valores del ARVI (izquierda) y una imagen satelital con colores naturales (derecha)

3.2.5. Cálculo del índice SIPI

El índice SIPI o Structure Insensitive Pigment Index es un indicador asociado a la ratio entre carotenoides y la clorofila, y permite la mitigación de los efectos de las propiedades de la estructura foliar de la vegetación sobre la reflectancia, dado que tiene en consideración el espectro cercano al infrarrojo. Los valores habituales de este índice se mueven entre 0 y 2, y según la literatura, la vegetación saludable se encuentra entre 0,8 y 1,8.

De manera similar al caso anterior, el SIPI requiere realizar una operación con las bandas contenidas en la imagen original, utilizando en este caso la siguiente expresión:

$$\frac{NIR - Blue}{NIR + Red}$$

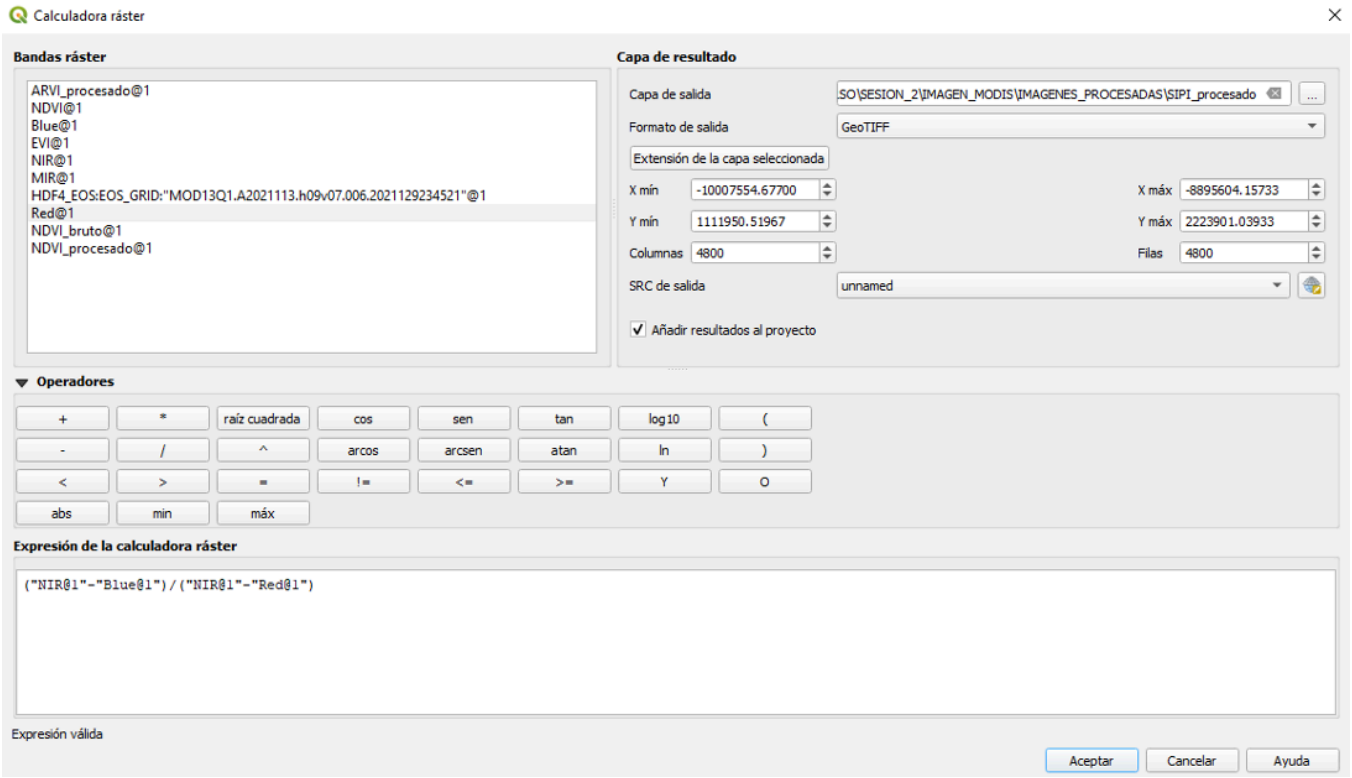


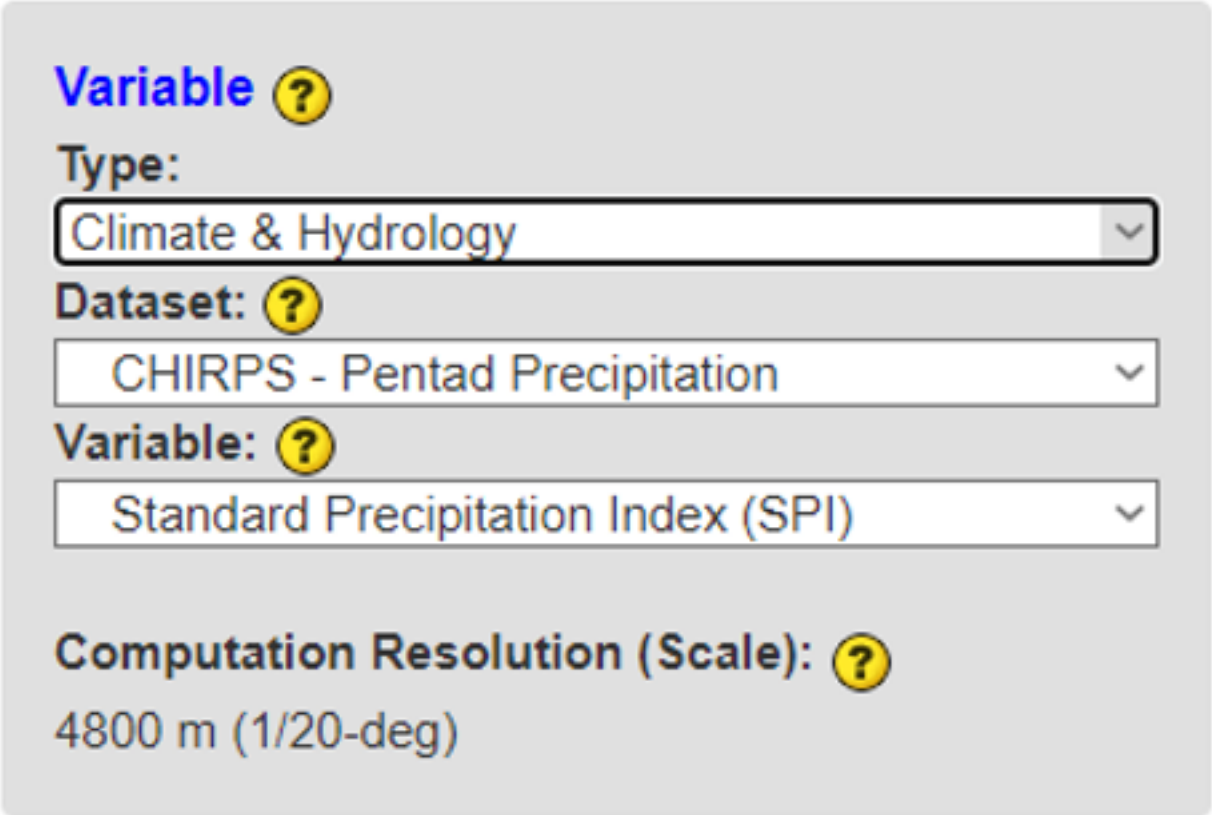
Ilustración 30. Cálculo del SIPI mediante la calculadora ráster.

3.3 Cálculo del índice SPI

El **SPI o Standardized Precipitation Index** se calcula a partir de registros de precipitación, obteniendo una serie de probabilidades en base a datos de precipitación acumulada. La principal ventaja de este índice es que facilita el análisis de las sequías en distintas escalas temporales (puede variar desde un solo mes hasta varios años).

Para este ejemplo, vamos a emplear el rango de 1 mes para evaluar si la situación de sequía meteorológica. En este caso lo vamos a realizar para toda la extensión de CA4 (Guatemala, Honduras, El Salvador y Nicaragua).

Vamos a obtener la información procedente de la plataforma **"Climate Engine"** del FEWS NET (Famine Early Warning Systems Network). Este repositorio nos permite descargar multitud de variables climáticas de numerosas fuentes. Lo primero que debemos de hacer es establecer qué variable queremos descargar, en este caso datos referentes al índice SPI en base a datos del producto CHIRPS. Para ello, nos dirigimos al apartado de "Variable" (menú de la parte izquierda) y establecemos dichos parámetros:



Illustración 31. Parámetros a introducir en el apartado de "Variable" en la página de Climate Engine.

En el apartado de procesado dejamos la opción “Standardized Precipitation Index” que viene por defecto. Continuamos con los parámetros temporales, donde en la pestaña “Aggregation Period” establecemos la opción de 1 mes. Esto quiere decir que la capa que vamos a obtener hace referencia al índice SPI con un periodo de tiempo del último mes. Se pueden establecer otro tipo de intervalos temporales según el objetivo de estudio. La pestaña de datos históricos la dejamos por defecto, que abarque el periodo completo de 1981-2021.

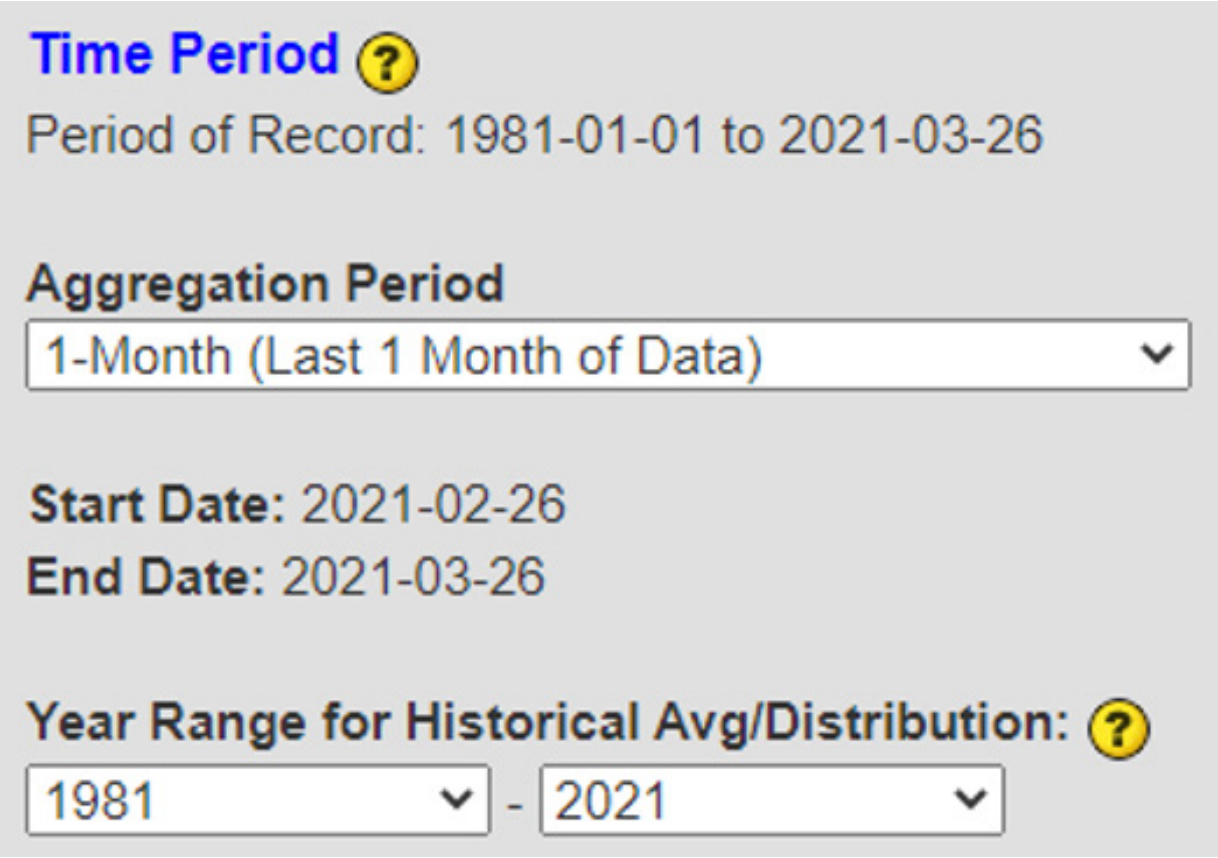


Ilustración 32. Parámetros a introducir en el apartado de “Time Period” en la página de Climate Engine.

Por último, nos dirigimos a la pestaña de descarga, establecemos una zona de interés, y pulsamos “Download Map Layer”, por lo que se descargará un .rar con el SPI obtenido para el periodo seleccionado. Por último, importamos el resultado a QGIS y simbolizamos la capa con una paleta de colores que queramos.

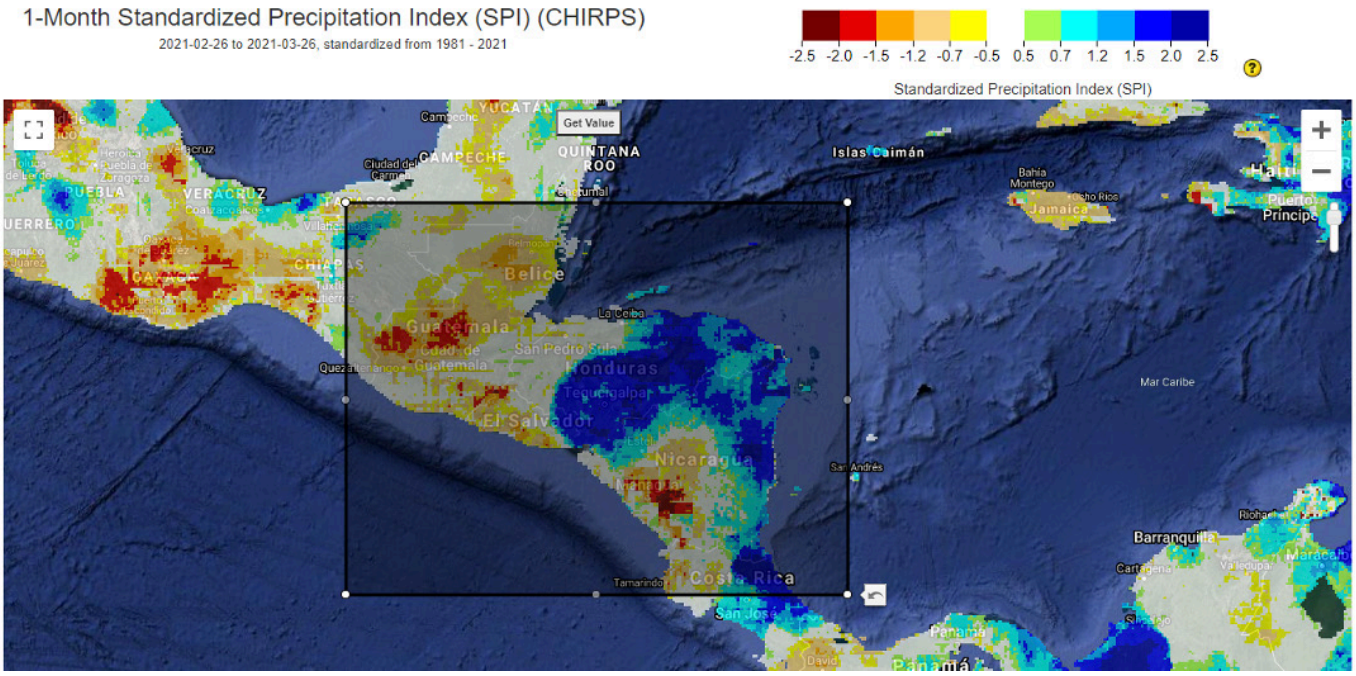


Ilustración 33: Paso de descarga en el repositorio de Climate Engine.

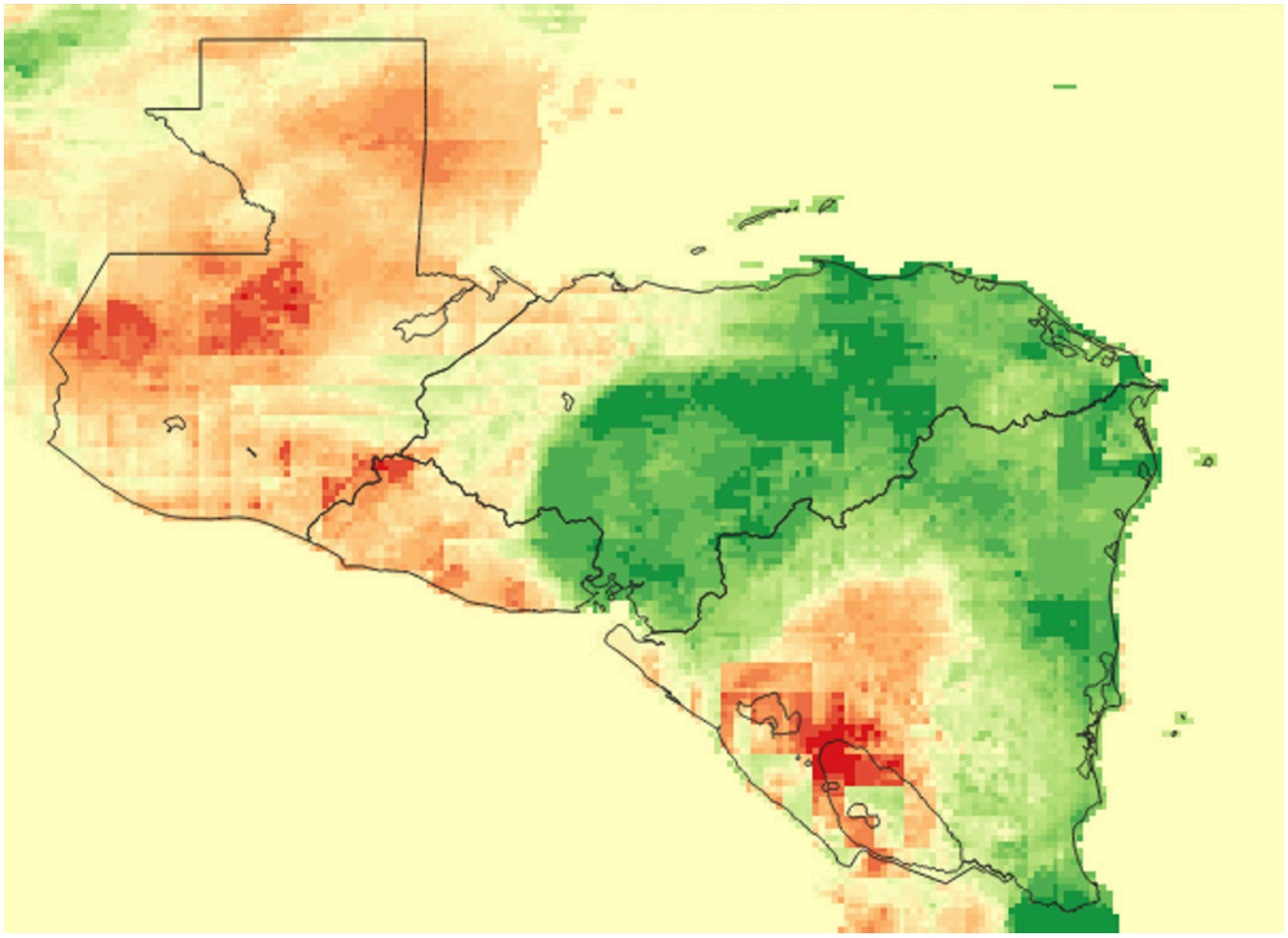


Ilustración 34: Capa ráster con los valores de SPI para el ámbito de CA4.

Se puede observar que las zonas más afectadas por la sequía en los parámetros que hemos considerado son Guatemala y El Salvador de manera general, el noroeste de Honduras y la región central de Nicaragua. Por otra parte, se pueden apreciar zonas con un volumen de precipitaciones superior a la media en el centro y este de Honduras, así como en el este de Nicaragua.

3.4 Cálculo de estadísticas zonales

Una vez que hemos obtenido las diferentes capas de información que nos servirán para la evaluación del estado de los cultivos, procedemos al cálculo de las estadísticas zonales. Este paso requiere, además de los índices calculados hasta el momento, disponer de una capa de información que nos permita identificar dónde se encuentran los cultivos de grano básico. La forma de obtener la capa de información de las parcelas de grano básico queda fuera del alcance de este curso, por lo que se proveerá para poder ejecutar los pasos del presente manual. A modo de ejemplo, se muestra la información contenida en dicha capa en la siguiente imagen:

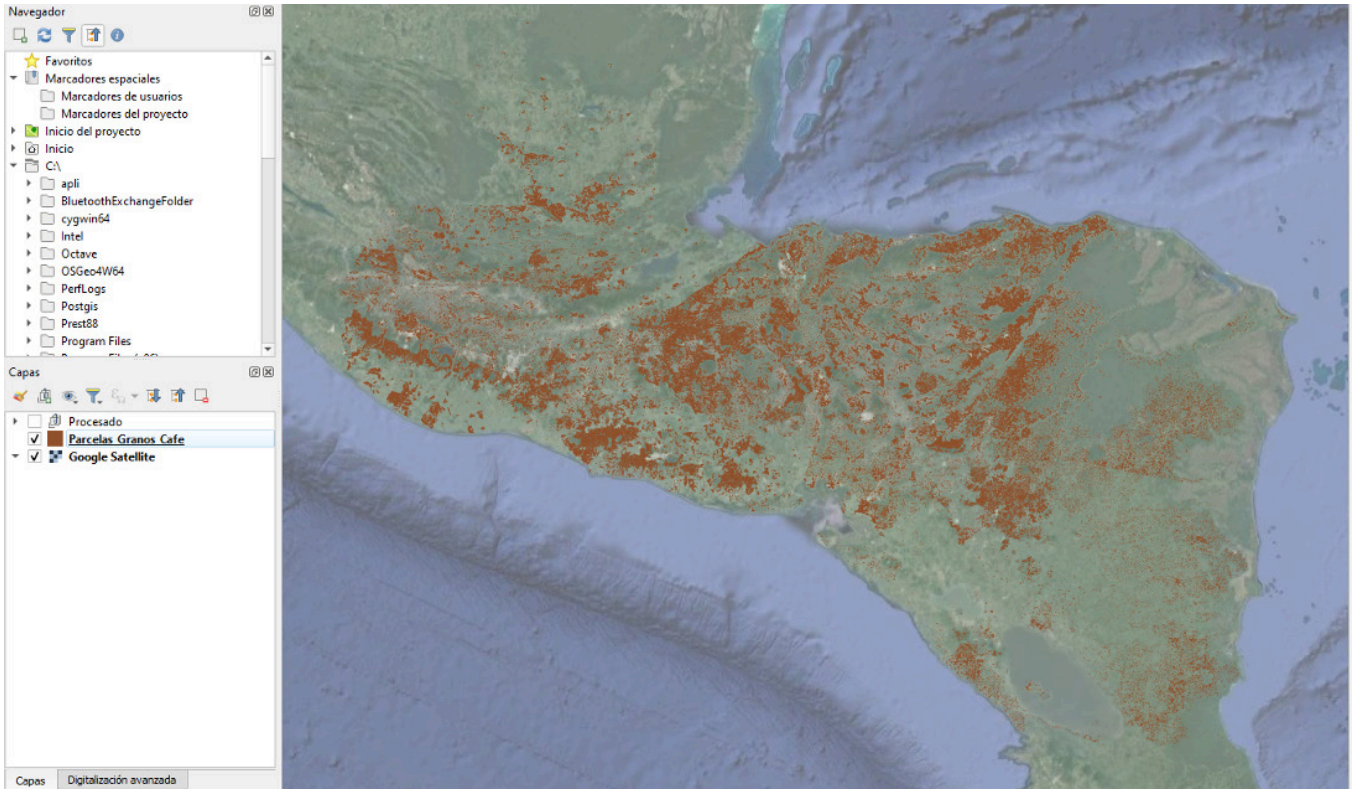


Ilustración 35. Capa vectorial que contiene la ubicación de las parcelas de grano básico.

La función que utilizaremos para calcular estas estadísticas es una que viene por defecto en QGIS, y no requiere de ninguna instalación adicional. Se trata de la función Estadísticas de zona, que se puede encontrar en la Caja de herramientas.

Illustración 36. Localización de la función Estadísticas de zona.

Como capa de entrada en el primer parámetro introducimos la capa que contiene la geometría vectorial de las parcelas de grano básico y café. La capa ráster introducida se corresponderá con el índice que queramos calcular en un momento dado, y como prefijo de la columna de salida escribiremos NDVI_, ya que es el índice que vamos a computar para este ejemplo.

Adicionalmente, se deben configurar dos elementos:

- Por un lado, se tiene que modificar el comportamiento del algoritmo seleccionado con respecto a las geometrías inválidas. Para ello, nos dirigimos a la opción mostrada en la ilustración inferior, y en la siguiente pantalla cambiamos el valor de parámetro Invalid feature filtering de Use Default a Skip (Ignore) Features with Invalid Geometries.

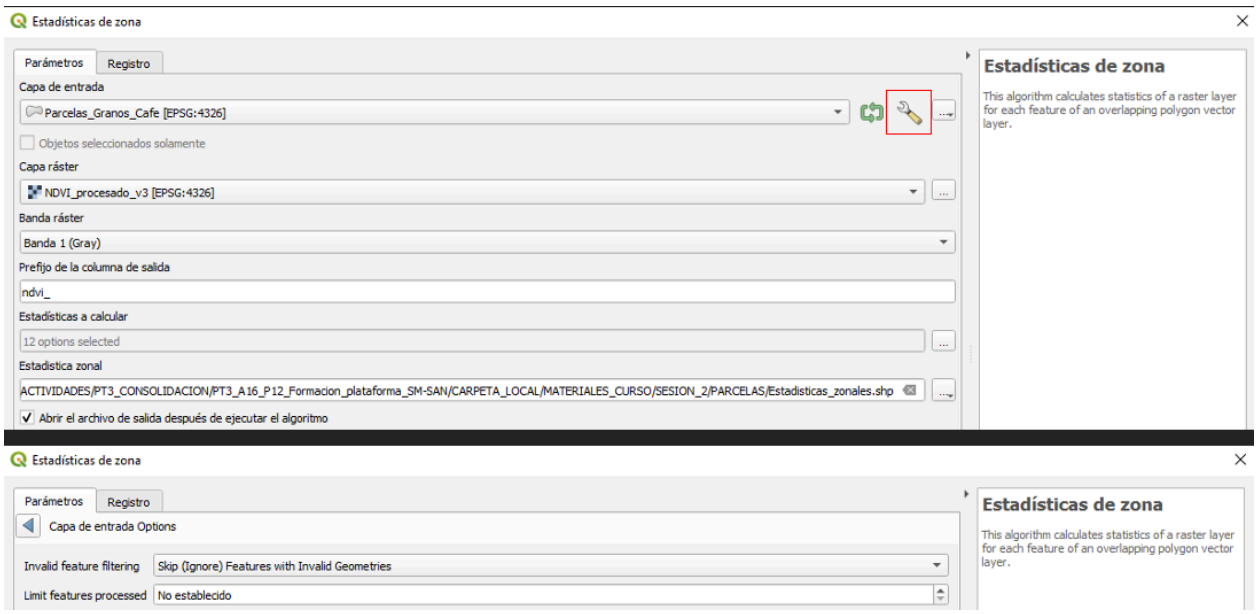


Ilustración 37. Panel de Estadísticas de zona.

A continuación, en el parámetro Estadísticas a calcular, seleccionamos la media para caracterizar el comportamiento del índice en la zona de estudio. A efectos prácticos, se puede utilizar cualquier otro estadístico para este fin, ya que cuanta más información tengamos disponible, mejor podremos determinar la influencia de las condiciones climáticas sobre el estado de los cultivos.

Ilustración 38. Panel de Estadísticas de zona.

Lanzamos el algoritmo y obtenemos la misma capa de parcelas de cultivo, pero con tantas columnas como estadísticos hayamos calculado. De este modo, el resultado es un conjunto de parcelas identificadas con la columna ID, donde cada valor se corresponde con un municipio de CA4, el tipo de cultivo, y el valor del estadístico obtenido para el raster del índice NDVI.

Esta operación se vuelve a realizar para cada uno de los índices que estamos calculando, aunque en lugar de utilizar como base la capa de polígonos denominada Parcelas_Granos_Cafe, emplearemos la que acabamos de crear con nombre Estadísticas_NDVI. De esta manera, podemos ir agregando la información de manera secuencial, y los índices se irán colocando como nuevas columnas:

	ID	Cultivo	ndvi_mean	arvi_mean	evi_mean	sipi_mean
1	1	CafÃ©	0.634002886746...	0.572218091173...	0.439555674408...	1.149842948677...
2	1	Granos BÃ¡sicos	0.504797850579...	0.375592846497...	0.310025068397...	1.253603644824...
3	2	CafÃ©	0.357210224339...	0.440964603269...	0.272985703355...	1.149856981092...
4	2	Granos BÃ¡sicos	0.19422892199313	0.282420247416...	0.156258697752...	0.702616309293...
5	3	CafÃ©	0.748967863018...	0.681002482786...	0.493842329113...	1.071912185138...
6	3	Granos BÃ¡sicos	0.505842858901...	0.331650104663...	0.320506493186...	1.292180545918...
7	4	CafÃ©	0.503082576964...	0.613403749267...	0.372405727911...	1.069335642101...
8	4	Granos BÃ¡sicos	0.468933333953...	0.563206483920...	0.291299988826...	0.874950448671...
9	5	CafÃ©	0.54036122438859	0.384570707909...	0.308106123640...	1.259743443557...

Ilustración 39. Tabla de atributos.

PREPARACIÓN DE DATOS PARA EL MODELO PREDICTIVO

4

4. Preparación de datos para el modelo predictivo

La alimentación de la Plataforma, y, por lo tanto, del modelo predictivo, se basa en la estructuración de manera ordenada de la información socioeconómica y agroclimática recopilada hasta el momento. La estructuración de estos datos se hará mediante el software Microsoft Excel.

Cabe destacar los siguientes puntos:

- El indicador de referencia que tomamos es el ID, que se trata de un número que identifica de manera única a cada uno de los municipios de CA4.
- Se debe realizar un conjunto de operaciones de ordenación y de transformación que se aplicarán sobre los diferentes grupos de datos generados u obtenidos.

Para describir cómo se deben ordenar los datos para posteriormente usarlos en el modelo predictivo, se desglosará este apartado en varios subapartados, que se corresponderán con los conjuntos de datos empleados para dicho modelo.

4.1 Datos socioeconómicos

La mayoría de los datos socioeconómicos han sido extraídos de repositorios públicos como los respectivos organismos oficiales de estadística de cada país, de fuentes como **Our World in Data, Humanitarian Data Exchange.** Otra porción de estos datos ha sido elaborada mediante la realización de geoprocesos, como son el cálculo de la altitud o la pendiente media. Dado que la obtención de estos datos queda fuera del alcance de este curso, se proporcionará la tabla con los datos socioeconómicos necesarios.

PAIS	DPTO	MUN	ID	Población	Superficie (km2)	Densidad_Población (hab/km2)
EL SALVADOR	AHUACHAPAN	AHUACHAPAN	1	110 511	251.09	440.12
EL SALVADOR	AHUACHAPAN	APANECA	2	8 383	40.74	205.79
EL SALVADOR	AHUACHAPAN	ATIQUIZAYA	3	33 587	62.99	533.19
EL SALVADOR	AHUACHAPAN	CONCEPCION DE ATACO	4	12 786	68.31	187.18
EL SALVADOR	AHUACHAPAN	EL REFUGIO	5	8 171	11.33	721.41
EL SALVADOR	AHUACHAPAN	GUAYMANGO	6	19 037	78.73	241.81
EL SALVADOR	AHUACHAPAN	JUJUTLA	7	28 539	179.80	158.06
EL SALVADOR	AHUACHAPAN	SAN FRANCISCO MENENDEZ	8	42 607	266.69	159.76
EL SALVADOR	AHUACHAPAN	SAN LORENZO	9	9 194	42.73	215.17
EL SALVADOR	AHUACHAPAN	SAN PEDRO PUXTLA	10	7 773	25.07	310.04
EL SALVADOR	AHUACHAPAN	TACUBA	11	29 858	157.31	189.80
EL SALVADOR	AHUACHAPAN	TURIN	12	8 997	9.99	900.37
EL SALVADOR	CABAÑAS	CINQUERA	13	1467	31.95	45.92
EL SALVADOR	CABAÑAS	DOLORES	14	6 347	138.84	45.71
EL SALVADOR	CABAÑAS	GUACOTECTI	15	5 550	28.31	196.02
EL SALVADOR	CABAÑAS	ILOBASCO	16	61510	234.80	261.97
EL SALVADOR	CABAÑAS	JUTIAPA	17	6 584	86.30	76.29
EL SALVADOR	CABAÑAS	SAN ISIDRO	18	7 795	77.79	100.21
EL SALVADOR	CABAÑAS	SENSUNTEPEQUE	19	40 332	310.92	129.72

Ilustración 40. Tabla.

A modo de ejemplo, se puede ver cómo cada municipio viene identificado con la columna ID, y las columnas describen las características de dicho municipio.

4.2 Datos agroclimáticos

Los datos agroclimáticos obtenidos más atrás se deben transformar para poder adecuarlos a la forma en la que debemos estructurar la información para poder introducirla en el modelo predictivo. Para ello, partiremos de la última estadística zonal que obtuvimos, que contiene los diferentes índices que nos resultan de interés.

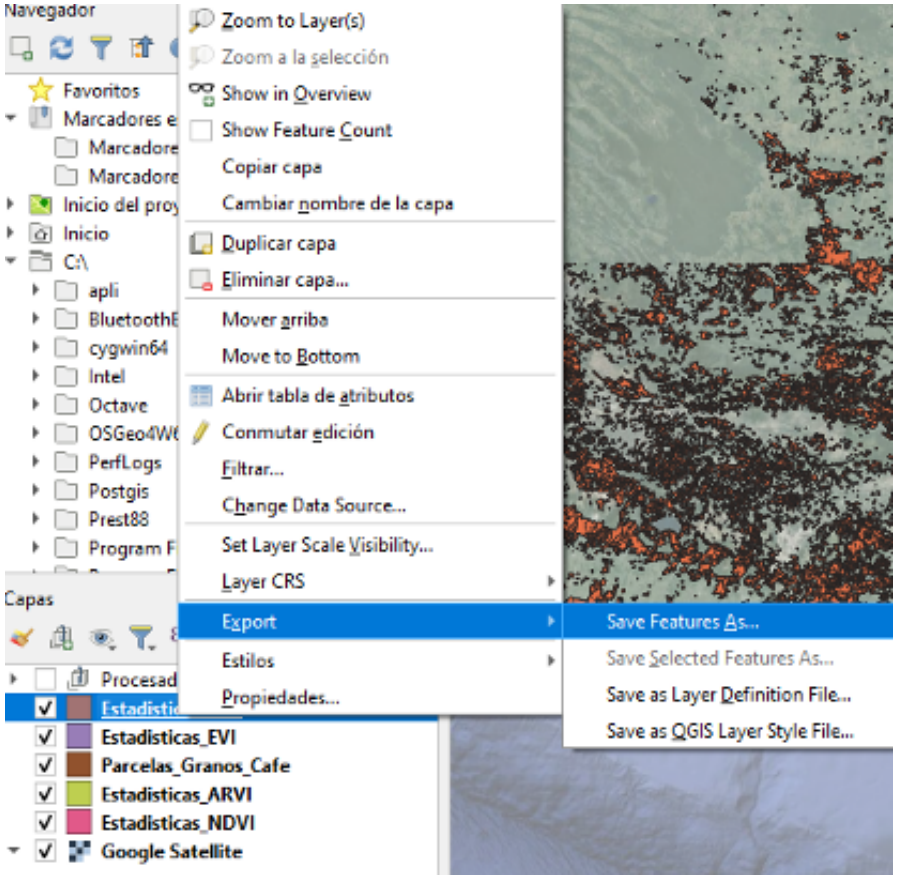


Ilustración 41. Exportar capa.

Para ello, nos dirigimos a la capa en la que se encuentran todos los índices calculados, pulsamos el botón derecho del ratón y en la opción Export, seleccionamos Save Features As…. En la opción de formato seleccionamos “Hoja de cálculo de MS Office Open XML (XLSX)”, y guardamos el archivo en nuestro lugar de elección.

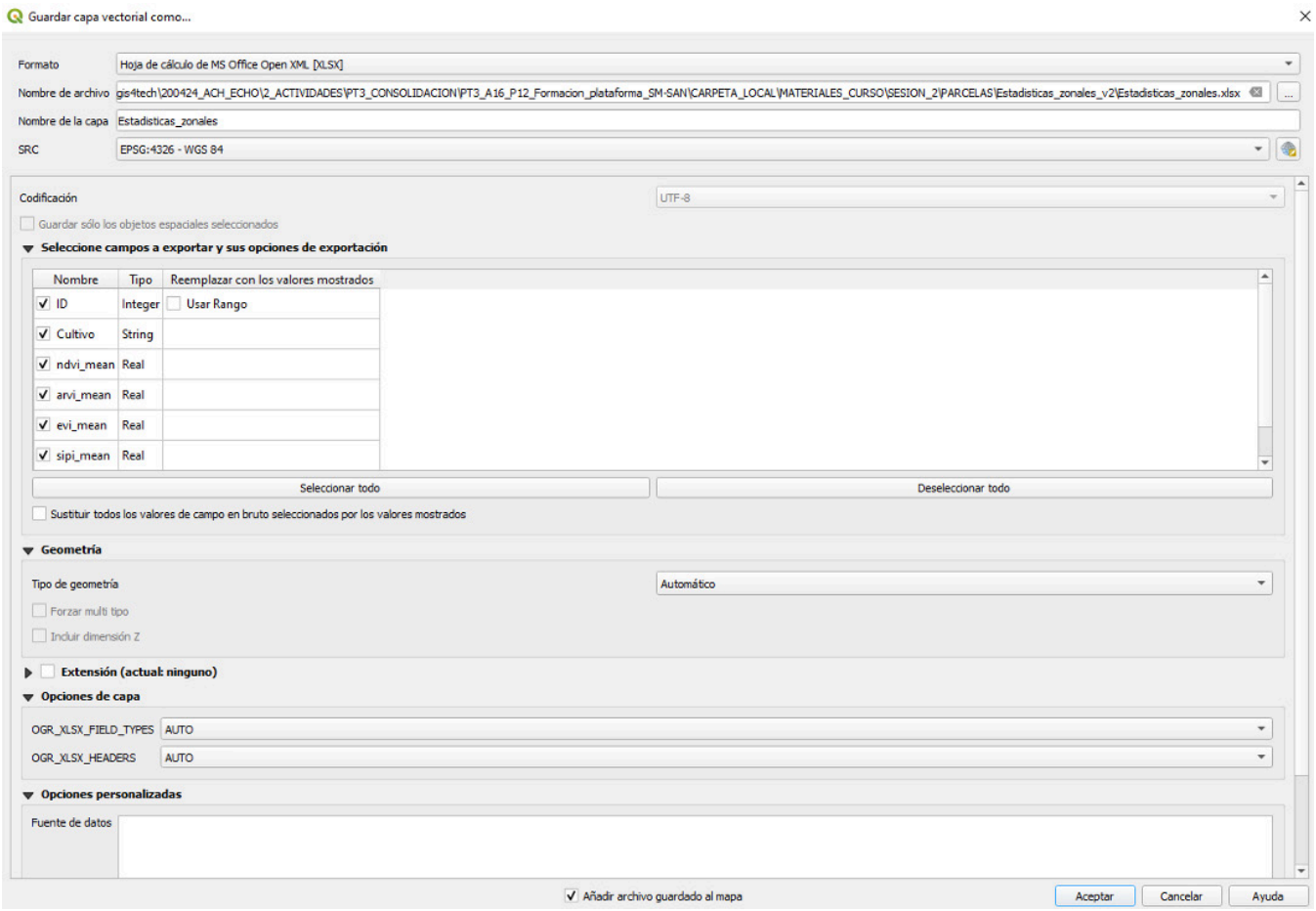
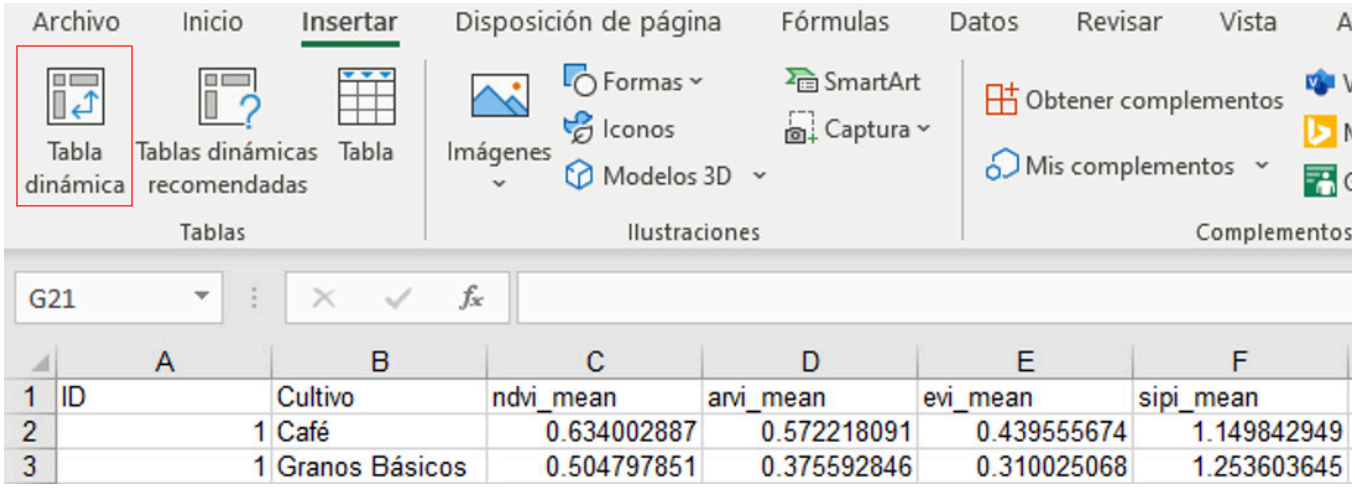


Ilustración 42. Guardado de capa.

Abrimos el fichero de Microsoft Excel donde tendremos las estadísticas zonales asociadas a cada tipo de cultivo y a cada municipio.

En la cinta de herramientas que se encuentra en la parte superior, nos dirigimos a la pestaña Insertar y seleccionamos Tabla dinámica. Aceptamos las opciones por defecto y nos aparecerá una nueva hoja de cálculo en la que aparecerá la tabla dinámica.



	A	B	C	D	E	F
1	ID	Cultivo	ndvi_mean	arvi_mean	evi_mean	sipi_mean
2	1	Café	0.634002887	0.572218091	0.439555674	1.149842949
3	1	Granos Básicos	0.504797851	0.375592846	0.310025068	1.253603645

Ilustración 43. Excel.

Los campos en la tabla dinámica se deben colocar tal y como aparecen en la siguiente imagen.

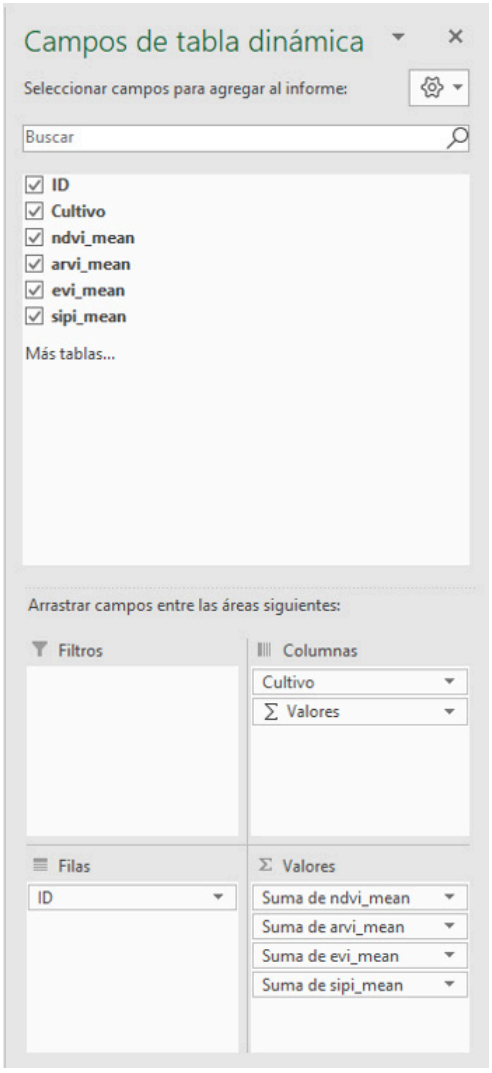
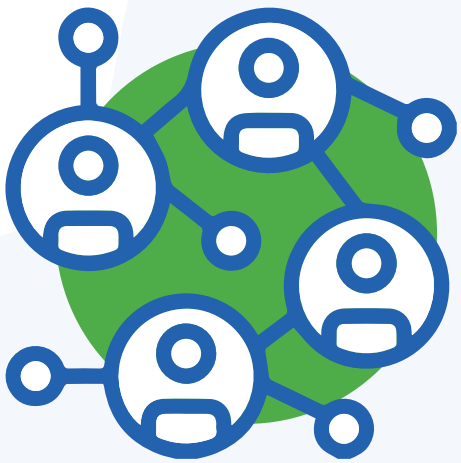


Ilustración 44. Campos de tabla dinámica de Excel

De esta forma, el resultado es una tabla en la que en cada fila tendremos el ID de los municipios de CA4, y en las columnas dispondremos de las estadísticas zonales de cada uno de los índices que hemos obtenido hasta el momento, desglosados por tipo de cultivo.

Se puede observar que al lado de los índices aparece el término “Suma”. Cabe destacar que en este caso cualquier método de agregación (suma, promedio, etc.) nos dará el mismo resultado, puesto que solo hay un dato por fila.



	Café				Granos Básicos			
Etiquetas de fila	Suma de ndvi_mean	Suma de arvi_mean	Suma de evi_mean	Suma de sipi_mean	Suma de ndvi_mean	Suma de arvi_mean	Suma de evi_mean	Suma de sipi_mean
1	0.634002887	0.572218091	0.439555674	1.149842949	0.504797851	0.375592846	0.310025068	1.253603645
2	0.357210224	0.440964603	0.272985703	1.149856981	0.194228922	0.282420247	0.156258698	0.702616309
3	0.748967863	0.681002483	0.493842329	1.071912185	0.505842859	0.331650105	0.320506493	1.292180546
4	0.503082577	0.613403749	0.372405728	1.069335642	0.468933334	0.563206484	0.291299989	0.874950449
5	0.540361224	0.384570708	0.308106124	1.259743444				
6	0.308019564	0.360934909	0.226702175	1.169580317	0.475240979	0.370482702	0.285673033	1.234337949
7	0.364211542	0.458280252	0.258122863	1.031832883	0.527621169	0.428994284	0.323953156	1.189146303
8	0.69220001	0.686148859	0.387620687	1.047537631	0.588709536	0.501954119	0.358942913	1.136428308
9					0.495326243	0.344595352	0.289730574	1.264211043
10	0.427536176	0.476160397	0.318528531	1.035752139	0.606785048	0.491318419	0.390538482	1.161036731
11	0.537972162	0.539952605	0.388710324	1.11672837	0.484672415	0.472614257	0.297343247	1.150011912
12	0.667179737	0.559714275	0.457247059	1.121011644	0.537976926	0.402605526	0.374012308	1.194193862
13					0.609776276	0.479565912	0.340618644	1.152987281
14					0.545767921	0.382354587	0.349372996	1.246389456
15					0.549837088	0.439673461	0.354880463	1.180454025
16	0.720699998	0.655190977	0.479181813	1.055008271	0.649754075	0.555896951	0.405056336	1.153107354
17					0.582610969	0.42672458	0.338442675	1.206206199
18					0.627058755	0.48455754	0.392739549	1.172892676
19					0.581730774	0.484350612	0.36309611	1.187770445
20	0.605291763	0.473789778	0.416037645	1.319690375	0.587448917	0.514029142	0.410421984	1.304723124

Ilustración 45. Tabla dinámica de Excel.

Copiamos los datos que nos resultan de interés en una nueva hoja de cálculo para poder realizar las manipulaciones que sean necesarias.

4.3 Datos de seguridad alimentaria y nutricional

Partimos de las encuestas tomadas en campo, específicamente de los resúmenes territoriales de dichas encuestas, que se proporcionan para la realización de este curso. Como se puede comprobar, en este caso se dispone de los diferentes municipios con su identificador único como se comentó anteriormente.

					Nivel de seguridad alimentaria			
País	Departamento	Municipio	ID	Número de encuestas	Seguridad alimentaria	Inseguridad alimentaria leve	Inseguridad alimentaria moderada	Inseguridad alimentaria severa
El Salvador	Ahuachapán	Tacuba	11	37	8.1%	59.5%	32.4%	0.0%
El Salvador	Cabanas	San Isidro	18	59	35.6%	62.7%	1.7%	0.0%
El Salvador	Cabanas	Victoria	21	49	22.4%	67.3%	10.2%	0.0%
El Salvador	Chalatenango	Nueva Concepcion	39	49	28.6%	71.4%	0.0%	0.0%
El Salvador	La Unión	Conchagua	118	49	6.1%	63.3%	28.6%	2.0%
El Salvador	La Unión	El Carmen	119	29	10.3%	79.3%	10.3%	0.0%
El Salvador	Morazan	Delicias de Concepcion	137	41	17.1%	78.0%	4.9%	0.0%
El Salvador	Morazan	Gualococti	140	42	21.4%	45.2%	33.3%	0.0%
El Salvador	Morazan	Guatajiagua	141	36	5.7%	54.3%	37.1%	2.9%
El Salvador	Morazan	Osicala	147	51	25.5%	66.7%	7.8%	0.0%

Ilustración 46. Tabla de Nivel de seguridad alimentaria por municipio.

A partir de esta información, se debe definir un criterio para decidir si un municipio pertenece a una categoría de seguridad alimentaria u otra. Para este curso, se tomará como criterio la proporción del 20% para cada municipio, es decir, se seguirá el siguiente flujo:

Si encuestas "Inseguridad alimentaria severa" > 20%:
Categoría = "Inseguridad alimentaria severa"

Si encuestas "Inseguridad alimentaria moderada" > 20%:
Categoría = "Inseguridad alimentaria moderada"

Si encuestas "Inseguridad alimentaria leve" > 20%:
Categoría = "Inseguridad alimentaria leve"

Cualquier otro caso:
Categoría = "Seguridad alimentaria"

Para definir este flujo en Microsoft Excel, introducimos la siguiente fórmula en la columna adyacente a los datos:

= SI (I3 > 0.2;$I$2;SI(H3 > 0.2; $H$2;SI(G3 > 0.2; $G$2;$F$2)))

De esta manera, nos quedaría lo siguiente:

					Nivel de seguridad alimentaria				
País	Departamento	Municipio	ID	Número de encuestas	Seguridad alimentaria	Inseguridad alimentaria leve	Inseguridad alimentaria moderada	Inseguridad alimentaria severa	Categoría SAN
El Salvador	Ahuachapán	Tacuba	11	37	8.1%	59.5%	32.4%	0.0%	=SI(I3>0.2;$I$2;SI(H3>0.2;$H$2;SI(G3>0.2;$G$2;$F$2)))
El Salvador	Cabanas	San Isidro	18	59	35.6%	62.7%	1.7%	0.0%	Inseguridad alimentaria leve
El Salvador	Cabanas	Victoria	21	49	22.4%	67.3%	10.2%	0.0%	Inseguridad alimentaria leve
El Salvador	Chalatenango	Nueva Concepcion	39	49	28.6%	71.4%	0.0%	0.0%	Inseguridad alimentaria leve
El Salvador	La Unión	Conchagua	118	49	6.1%	63.3%	28.6%	2.0%	Inseguridad alimentaria moderada
El Salvador	La Unión	El Carmen	119	29	10.3%	79.3%	10.3%	0.0%	Inseguridad alimentaria leve
El Salvador	Morazan	Delicias de Concepcion	137	41	17.1%	78.0%	4.9%	0.0%	Inseguridad alimentaria leve
El Salvador	Morazan	Gualococti	140	42	21.4%	45.2%	33.3%	0.0%	Inseguridad alimentaria moderada

Ilustración 47. Generación de la categoría SAN.

A continuación, generamos una nueva columna que contendrá la información relativa a las Zonas de preocupación. Si analizamos las diferentes categorías resultantes, podemos observar que existe un importante desequilibrio entre las posibles categorías. La consecuencia de este hecho es que el modelo predictivo no será capaz de predecir situaciones extremas (seguridad alimentaria o inseguridad severa), ya que existe muy poca información respecto a estos municipios.

	Seguridad alimentaria	Inseguridad alimentaria leve	Inseguridad alimentaria moderada	Inseguridad alimentaria severa
Números de municipios	0	58	39	2

Ilustración 48. Tabla de recuento de categorías por municipio.

Por lo tanto, se propone el concepto de Zonas de preocupación, que no se corresponde de manera directa con las categorías SAN según la metodología CARI, pero que la toma de base. Dividimos los municipios en dos subcategorías:

Si Categoría SAN es Severa o Moderada:
Zona de preocupación = Alta [1]

Si Categoría SAN es leve o segura:
Zona de preocupación = Baja [0]

La fórmula que debemos introducir en la columna adyacente en Microsoft Excel es la siguiente:

= SI (O (J3 = $I$2;J3 = $H$2); 1;0)

					Nivel de seguridad alimentaria					
País	Departamento	Municipio	ID	Número de encuestas	Seguridad alimentaria	Inseguridad alimentaria leve	Inseguridad alimentaria moderada	Inseguridad alimentaria severa	Categoría SAN	Zona de preocupación
El Salvador	Ahuachapán	Tacuba	11	37	8.1%	59.5%	32.4%	0.0%	Inseguridad alimentaria moderada	=SI(O(J3=$I$2;J3=$H$2);1;0)
El Salvador	Cabanas	San Isidro	18	59	35.6%	62.7%	1.7%	0.0%	Inseguridad alimentaria leve	0
El Salvador	Cabanas	Victoria	21	49	22.4%	67.3%	10.2%	0.0%	Inseguridad alimentaria leve	0
El Salvador	Chalatenango	Nueva Concepcion	39	49	28.6%	71.4%	0.0%	0.0%	Inseguridad alimentaria leve	0
El Salvador	La Unión	Conchagua	118	49	6.1%	63.3%	28.6%	2.0%	Inseguridad alimentaria moderada	1
El Salvador	La Unión	El Carmen	119	29	10.3%	79.3%	10.3%	0.0%	Inseguridad alimentaria leve	0
El Salvador	Morazan	Delicias de Concepcion	137	41	17.1%	78.0%	4.9%	0.0%	Inseguridad alimentaria leve	0
El Salvador	Morazan	Gualococti	140	42	21.4%	45.2%	33.3%	0.0%	Inseguridad alimentaria moderada	1

Ilustración 49. Generación de la columna de zona de preocupación.

La consecuencia de haber realizado esta transformación es que tendremos dos posibles clases que cuentan con un número similar de observaciones, lo cual va a facilitar las labores de predicción en fases posteriores del curso.

	Zonas de preocupación baja	Zonas de preocupación alta
Números de municipios	58	41

Ilustración 50. Recuento de las zonas de preocupación por municipio.

4.4 Montaje de la tabla final

Para la tabla final podemos partir de los datos socioeconómicos y mediante la función BUSCARV de Microsoft Excel, agregar los datos agroclimáticos. Para ello, debemos seguir los siguientes pasos:

- Abrimos los dos archivos (datos socioeconómicos y datos agroclimáticos).
- En el archivo de datos socioeconómicos, nos dirigimos a la última columna, y a continuación empleamos la siguiente fórmula:

```
= BUSCARV ($D2;' [02.Datos_agroclimaticos.xlsx] Datos agroclimáticos'!$A$1:$I$1044;2;FALSO)
```

Esta fórmula busca el ID de la fila en la que nos encontramos en el archivo de datos agroclimáticos, y nos devuelve el valor de la columna que deseemos para el ID correspondiente. En el primer caso, solicitaremos que nos devuelva el valor de la columna 2, que contiene el NDVI de las parcelas de café.

- Repetimos la operación anterior, modificando únicamente el número destacado en negrita, añadiendo 1 cada vez que cambiemos de columna

Este proceso se realiza del mismo modo para los datos SAN, seleccionando la columna de las Zonas de preocupación. El resultado se puede ver en la siguiente imagen:

PAIS	DPTO	MUN	ID	Vulnerabilidad.PorMediosDeVida(adim.)	Vulnerabilidad.PorCapacidadDeAdaptacion(adim.)	ndvi_mean_cafe	arvi_mean_cafe	evi_mean_cafe	sipi_mean_cafe	ndvi_mean_gb	arvi_mean_gb	evi_mean_gb	sipi_mean_gb	Zona preocupación
EL SALVADOR	AHUACHAPAN	AHUACHAPAN	1	2.00	5.00	0.634002887	0.572218091	0.439555674	1.149842949	0.504797851	0.375592846	0.310025068	1.253603645	#N/D
EL SALVADOR	AHUACHAPAN	APANECA	2	2.00	6.00	0.357210224	0.440964603	0.272985703	1.149856981	0.194228922	0.282420247	0.156258698	0.702616309	#N/D
EL SALVADOR	AHUACHAPAN	ATIQUIZAYA	3	2.00	6.00	0.748967863	0.681002483	0.493842329	1.071912185	0.505842859	0.331650105	0.320506493	1.292180546	#N/D
EL SALVADOR	AHUACHAPAN	CONCEPCION DE ATACO	4	2.00	5.00	0.503082577	0.613403749	0.372405728	1.069335642	0.468933334	0.563206484	0.291299989	0.874950449	#N/D
EL SALVADOR	AHUACHAPAN	EL REFUGIO	5	2.00	7.00	0.540361224	0.384570708	0.308106124	1.259743444	0	0	0	0	#N/D
EL SALVADOR	AHUACHAPAN	GUAYMANGO	6	2.00	5.00	0.308019564	0.360934909	0.226702175	1.169580317	0.475240979	0.370482702	0.285673033	1.234337949	#N/D
EL SALVADOR	AHUACHAPAN	JUJUTLA	7	2.89	1.00	0.364211542	0.458280252	0.258122863	1.031832883	0.527621169	0.428994284	0.323953156	1.189146303	#N/D
EL SALVADOR	AHUACHAPAN	SAN FRANCISCO MENENDEZ	8	4.29	1.00	0.69220001	0.686148859	0.387620687	1.047537631	0.588709536	0.501954119	0.358942913	1.136428308	#N/D
EL SALVADOR	AHUACHAPAN	SAN LORENZO	9	2.03	1.00	0	0	0	0	0.495326243	0.344595352	0.289730574	1.264211043	#N/D
EL SALVADOR	AHUACHAPAN	SAN PEDRO PUXTLA	10	2.00	6.00	0.427536176	0.476160397	0.318528531	1.035752139	0.606785048	0.491318419	0.390538482	1.161036731	#N/D
EL SALVADOR	AHUACHAPAN	TACUBA	11	2.18	3.00	0.537972162	0.539952605	0.388710324	1.11672837	0.484672415	0.472614257	0.297343247	1.150011912	1
EL SALVADOR	AHUACHAPAN	TURIN	12	2.00	7.00	0.667179737	0.559714275	0.457247059	1.121011644	0.537976926	0.402605526	0.374012308	1.194193862	#N/D
EL SALVADOR	CABANAS	CINQUERA	13	2.00	4.00	0	0	0	0	0.609776276	0.479565912	0.340618644	1.152987281	#N/D
EL SALVADOR	CABANAS	DOLORES	14	2.00	1.00	0	0	0	0	0.545767921	0.382354587	0.349372996	1.246389456	#N/D
EL SALVADOR	CABANAS	GUACOTECTI	15	2.00	6.00	0	0	0	0	0.549837088	0.439673461	0.354880463	1.180454025	#N/D
EL SALVADOR	CABANAS	ILOBASCO	16	2.00	3.00	0.720699998	0.655190977	0.479181813	1.055008271	0.649754075	0.555896951	0.405056336	1.153107354	#N/D
EL SALVADOR	CABANAS	JUTIAPA	17	2.00	1.00	0	0	0	0	0.582610969	0.42672458	0.338442675	1.206206199	#N/D
EL SALVADOR	CABANAS	SAN ISIDRO	18	2.00	4.00	0	0	0	0	0.627058755	0.48455754	0.392739549	1.172892676	0
EL SALVADOR	CABANAS	SENSUNTEPEQUE	19	2.00	1.00	0	0	0	0	0.581730774	0.484350612	0.36309611	1.187770445	#N/D
EL SALVADOR	CABANAS	TEJUTEPEQUE	20	2.00	3.00	0.605291763	0.473789778	0.416037645	1.319690375	0.587448917	0.514029142	0.410421984	1.304723124	#N/D
EL SALVADOR	CABANAS	VICTORIA	21	2.00	1.00	0	0	0	0	0.557044831	0.431707557	0.36123806	1.235234114	0
EL SALVADOR	CHALATENANGO	AGUA CALIENTE	22	2.00	1.00	0	0	0	0	0.600106372	0.507455608	0.416928749	1.15592941	#N/D
EL SALVADOR	CHALATENANGO	ARCATAO	23	2.00	6.00	0	0	0	0	0.425859006	0.254581386	0.35656236	1.329322443	#N/D
EL SALVADOR	CHALATENANGO	AZACUALPA	24	2.00	4.00	0	0	0	0	0.647	0.529902875	0.480124988	1.124681413	#N/D
EL SALVADOR	CHALATENANGO	CANCASQUE	25	2.00	3.00	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D	#N/D

Ilustración 51. Montaje de la tabla final.

En la imagen adjunta se pueden observar dos elementos:

- En las zonas de preocupación aparecen celdas que contienen un valor de "#N/D" o No Data. Esto se debe a que en esos municipios no se han realizado encuestas, por lo que no se dispone de ninguna clasificación.
- En los datos agroclimáticos aparecen celdas con dos valores diferentes al resto:
  - Cuando aparecen varias columnas con valores de 0 en una misma fila, significa que en ese municipio se han detectado cultivos de un tipo, pero no de otro.
  - Cuando en todas las columnas de una misma fila aparece el valor "#N/D", significa que en ese municipio no se han detectado cultivos ni de café ni de grano básico. Esto solo sucede en 11 municipios de los 1053 que hay en CA4, por lo que no es motivo de preocupación.

De este modo, podemos aplicar un filtro de columnas sobre los datos resultantes y quedarnos únicamente con las filas en las que sí tenemos valores de Zonas de preocupación.

Tomamos la información resultante de este filtrado y la guardamos en una nueva hoja de cálculo que denominaremos "Datos_entrenamiento.xls". Este conjunto de datos es el que le proporcionaremos al modelo predictivo para poder detectar cuáles son los patrones existentes entre las diferentes variables, lo cual nos permitirá realizar predicciones en fases posteriores del curso.

INTRODUCCIÓN A LA CIENCIA DE DATOS

5

5. Introducción a la ciencia de datos

5.1 Objetivos y aplicaciones

La ciencia de datos es una disciplina que emplea diversas tecnologías y métodos para analizar los datos recolectados por cualquier organización, para descubrir patrones y tendencias que optimicen la toma de decisiones estratégicas.

Los distintos modelos que se usan en ciencia de datos ya sean explícitos o implícitos, pueden clasificarse en función del objetivo del análisis que se esté realizando:

- **Análisis descriptivo:**

  Es comprender la realidad.

- **Análisis predictivo:**

  Es aquel que trata de responder a preguntas acerca de "qué va a pasar en el futuro", o, de manera más general, de predecir algo sobre datos no estudiados previamente.

- **Análisis prescriptivo:**

  Finalmente, el último escalón en la jerarquía de modelos es el de los utilizados para el análisis prescriptivo.

5.2 Modelos predictivos

Un modelo predictivo, como se ha explicado anteriormente, es aquel que trata de responder a preguntas acerca de qué va a ocurrir en el futuro. Se basa en el análisis de datos que han ocurrido en el pasado para poder predecir el futuro mediante el uso de técnicas que combinan las matemáticas e inteligencia artificial, a partir de variables específicas. La predicción implica y permite tomar mejores decisiones al conocer cómo va a funcionar cualquier ámbito del cual se desee obtener información.

El aprendizaje automático es una pieza fundamental de la ciencia de datos. El aprendizaje automático consta de una serie de algoritmos que permiten aprender con la experiencia. Es el cerebro de cualquier sistema de inteligencia artificial con el que se pretende aprender patrones de manera automática.

5.3 Tipos de modelos predictivos: resignación/clasificación

Existen dos tipos de modelos predictivos. A continuación, se van a analizar sus diferencias:

5.3.1. Algoritmos de clasificación

Los algoritmos de clasificación se usan cuando el resultado deseado es una etiqueta discreta, en otras palabras, son útiles cuando la respuesta al problema cae dentro de un conjunto finito de resultados posibles.

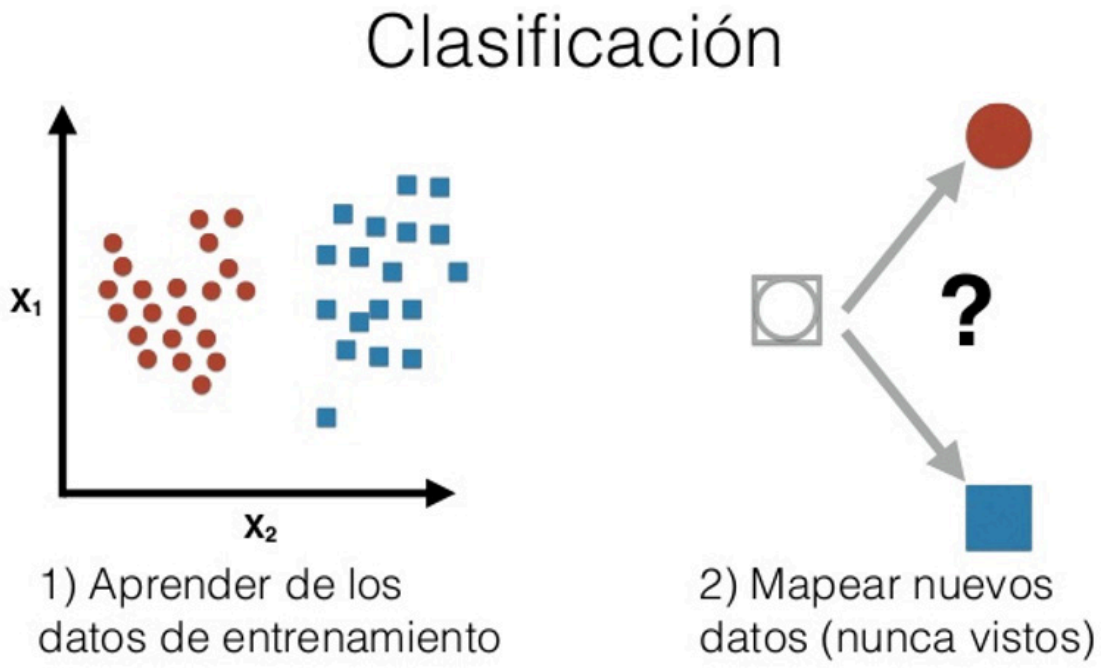
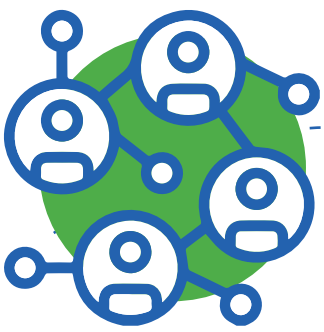


Ilustración 52. Gráficos de proceso de clasificación.

En el caso de que el modelo entrenado es para predecir cualquiera de las dos clases objetivos, verdadero o falso, por ejemplo, se le conoce como clasificación binaria. Algunos ejemplos de esto son: predecir si un alumno aprobará o no, predecir si un cliente comprará un producto nuevo o no.



5.3.2. Algoritmos de regresión

Por otro lado, la regresión es útil para predecir productos que son continuos, eso significa que la respuesta a su pregunta se presenta mediante una cantidad que puede determinarse de manera flexible en función de las entradas del modelo en lugar de limitarse a un conjunto de etiquetas.

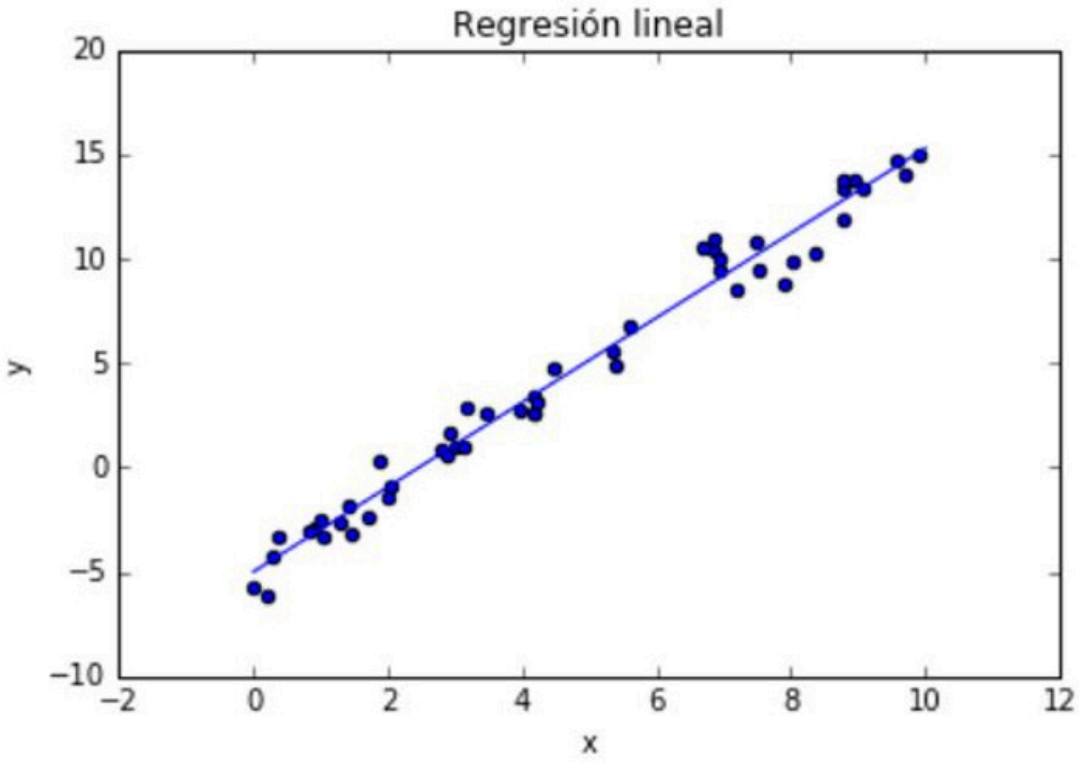


Ilustración 53. Gráfico de regresión lineal.

MANEJO DEL SOFTWARE ORANGE

6

6. Manejo del software Orange

Orange es una herramienta de software libre utilizada para la visualización y el análisis de datos, que facilita el procesamiento y el uso de técnicas de minería de datos para obtener información de alto valor añadido a partir de los datos provistos al software.

Dada su facilidad de uso y que se trata de una herramienta gratuita, utilizaremos Orange para explorar un conjunto de datos y generar un modelo predictivo muy sencillo, que nos facilitará la tarea más adelante, donde elaboraremos un modelo para la predicción de las zonas de preocupación basándonos en la categoría SAN de los municipios de CA4.

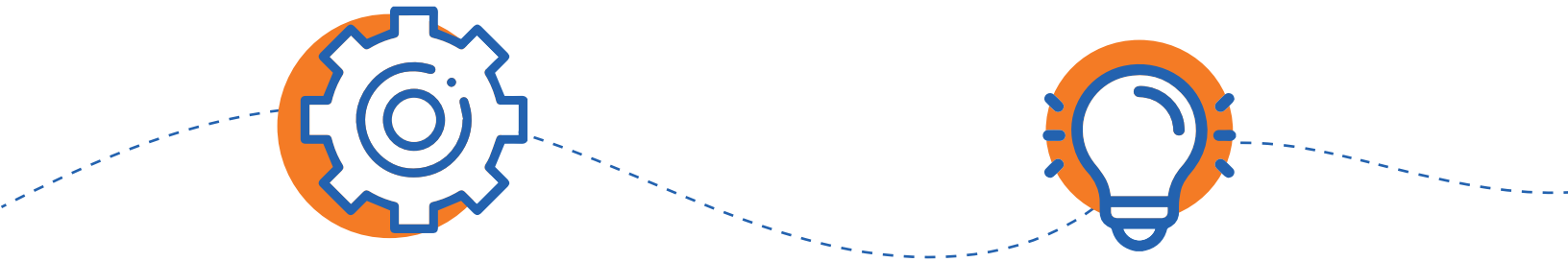
6.1 Instalación de Orange

En primer lugar, se accederá a la web de Orange donde se podrá acceder a las distintas opciones de instalación del programa: **https://orangedatamining.com/getting-started/**. Se hace clic en Download Orange.

En esta pantalla se muestran las distintas versiones para las que está disponible el archivo de instalación. Se hace clic en el icono correspondiente. Se elige el sistema operativo del que dispongamos en nuestra máquina, y pulsamos el botón de descarga.

Automáticamente empezará a descargarse automáticamente el archivo. Cuando finalice la descarga, se hace clic para comenzar con la instalación, y se aprieta el botón donde aparece escrito Next.

A partir de este punto, dejamos el resto de las opciones por defecto, pulsando sobre el botón Next o I agree según corresponda, hasta que aparezca la última pantalla, donde seleccionaremos la opción Finish. De este modo, ya tendremos instalado el software Orange y podremos proceder a explorar la interfaz y realizar los primeros análisis.



6.2 Primeros pasos

Cuando abrimos el software por primera vez, veremos una pantalla en la que el programa nos preguntará si queremos abrir un nuevo fichero o si tenemos algún proyecto previo. Crearemos un nuevo fichero de análisis, para lo cual seleccionaremos la opción New.

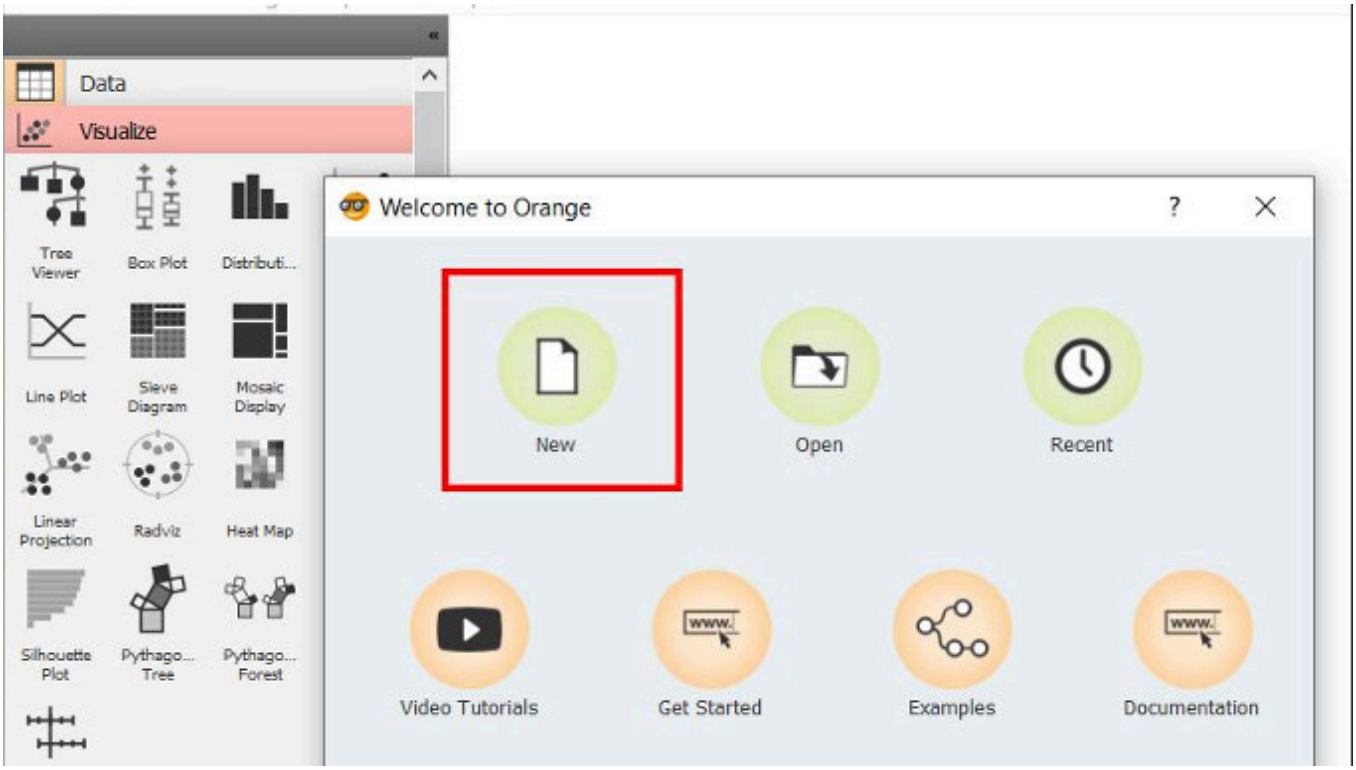


Ilustración 54. Nuevo documento en Orange.

La interfaz está compuesta por dos grandes bloques: el espacio de trabajo a la derecha y el espacio de herramientas a la izquierda, donde tenemos todas las herramientas agrupadas en 5 bloques:

- Datos
- Visualización
- Modelo
- Evaluación
- No supervisado

En el bloque de datos se encuentran las herramientas relacionadas con la limpieza y preprocesamiento de los datos, es decir, las transformaciones necesarias que hay que realizar en los datos que se estén utilizando antes de introducirlos en el modelo de predicción (regresión o clasificación).

En el segundo bloque de herramientas, el de visualización, estarán las herramientas para realizar gráficos. En el tercer bloque se encuentran los diferentes algoritmos supervisados.

En el cuarto bloque hay diferentes métodos de evaluar cómo de buena es la predicción que se ha realizado mediante los diferentes algoritmos de predicción, y en el bloque 5 están los algoritmos de aprendizaje automático no supervisado, mediante los cuales se pueden encontrar patrones ocultos en los datos que se estén usando entre otras funciones.

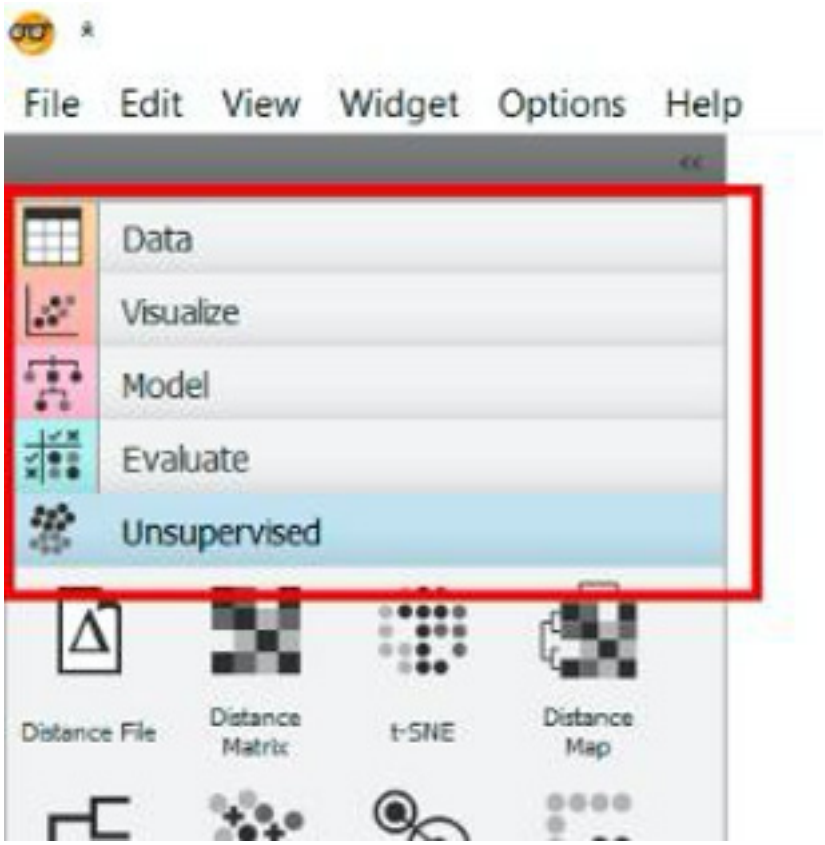


Ilustración 55. Cinco bloques de herramientas.

Por otra parte, en la parte superior izquierda, en la pestaña archivo aparecen las opciones para abrir un nuevo documento, guardar documento, salir…

Y por último a la derecha se encuentra el tablero en blanco donde se irán arrastrando las distintas herramientas y se mostrarán como burbujas que iremos uniendo unas con otras para analizar los datos.

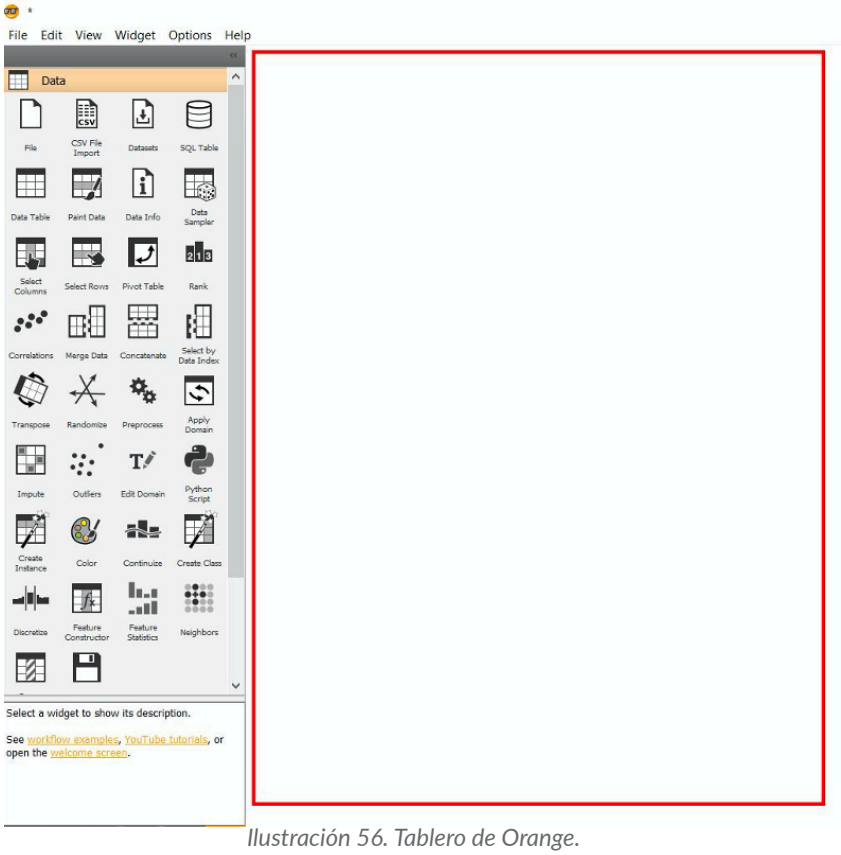


Ilustración 56. Tablero de Orange.

6.3 Ejemplo de aplicación

6.3.1. Bloque de herramientas de datos

Para comenzar a usar el programa, en las herramientas dentro del recuadro rojo en la siguiente imagen se puede importar datos desde un archivo csv con la herramienta “CSV File Import”. El programa también contiene una serie de conjuntos de datos predeterminados que sirven para realizar pruebas sin necesidad de disponer de datos.

Para comenzar a usar cualquier herramienta, se hace clic en esa herramienta, y sin soltar, arrastrar hasta la parte de la derecha, y soltar en el tablero en blanco.

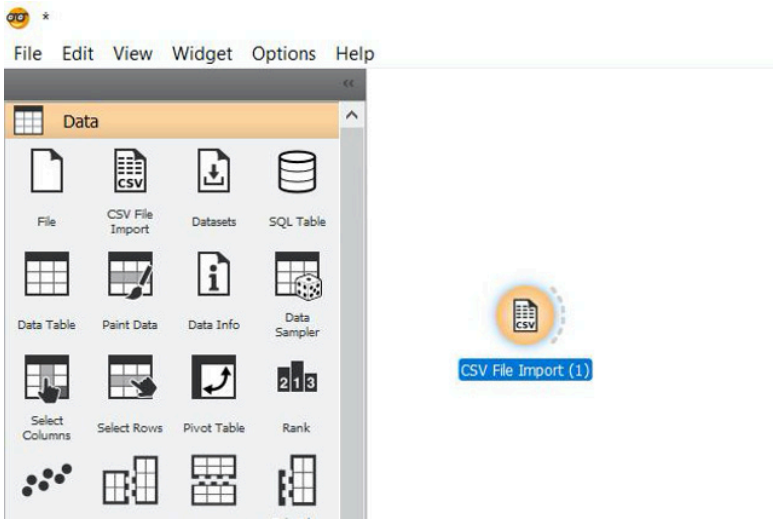


Ilustración 57. Importación de un fichero csv.

Carga de datos:

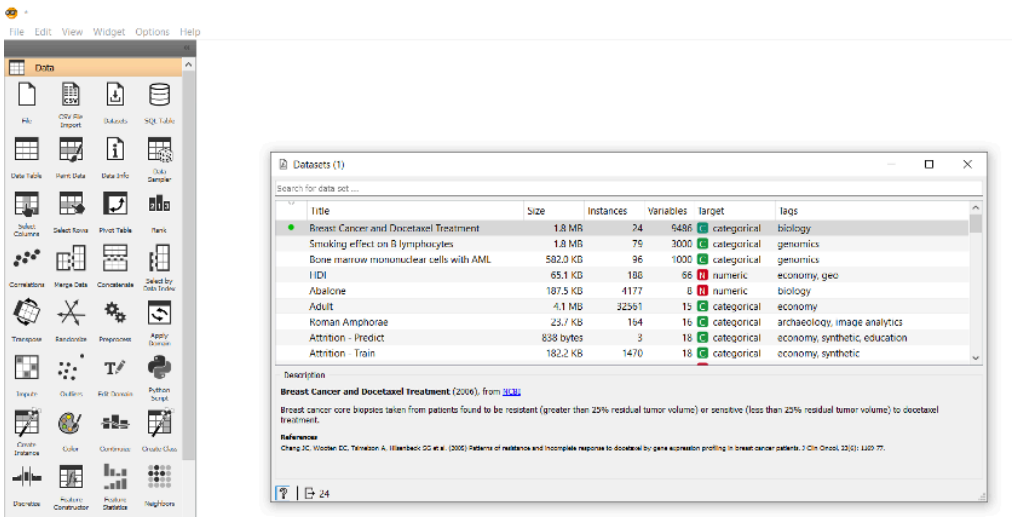


Ilustración 58. Exploración de datos cargados.

Una vez que hemos cargado el conjunto de datos que queremos estudiar, vamos a utilizar la opción Data Table, que permite explorar los datos cargados. Para ello, arrastramos la burbuja correspondiente a Data Table, conectamos la fuente de datos con la nueva burbuja, y aparecerá una nueva ventana donde podremos observar las variables correspondientes a cada uno de los individuos.

Ahora haciendo clic en Data Table se muestra la tabla.

Si se quiere eliminar una relación entre burbujas hay que apretar con el clic derecho en la relación (la línea gris) y seleccionar la opción de eliminar.

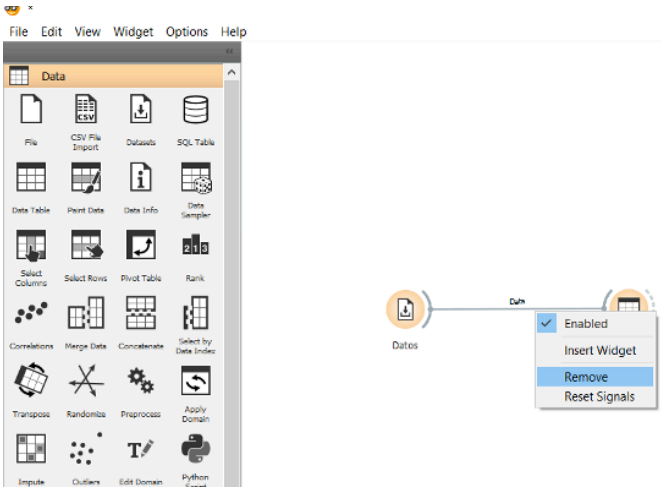


Ilustración 59. Eliminar de conexión.

Las herramientas que se muestran a continuación permiten seleccionar columnas y filas. Esta herramienta es de especial relevancia, dado que, para poder crear un modelo predictivo, es necesario decirle al programa que queremos elegir una variable respuesta o variable de salida. Es decir, le tenemos que decir cuál es la variable que queremos predecir, y con qué variables queremos realizar dicha predicción.

También se dispone de una herramienta llamada Select Rows, con la que podríamos eliminar filas si quisiéramos, pero dado que disponemos de pocos individuos en el conjunto de datos estudiado (24), no vamos a filtrar dichas filas.

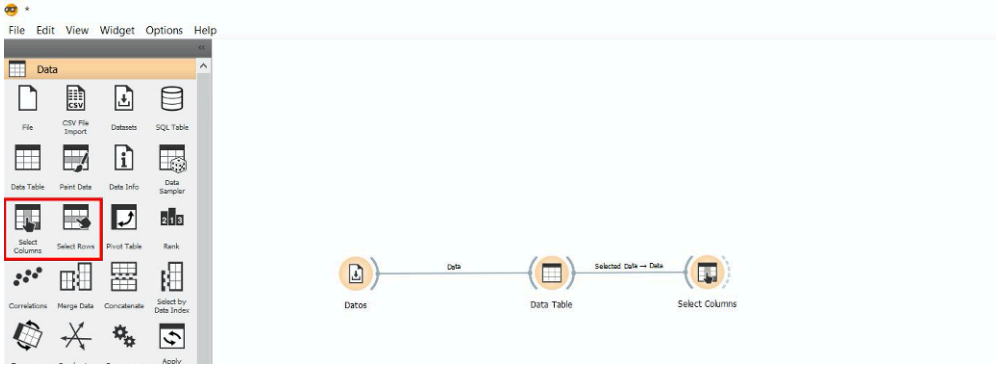


Ilustración 60. Seleccionar columnas.

Haciendo doble clic en la burbuja de *Select Columns* o seleccionar columnas nos aparece el cuadro de la imagen. Por defecto están todas las columnas seleccionadas en el cuadro blanco de la derecha. Si queremos eliminar del posterior análisis alguna de las columnas, simplemente tenemos que arrastrar de derecha a izquierda cada una de ellas. Como se puede observar, por defecto, la variable que se quiere predecir se ha colocado automáticamente en el recuadro *Target*.

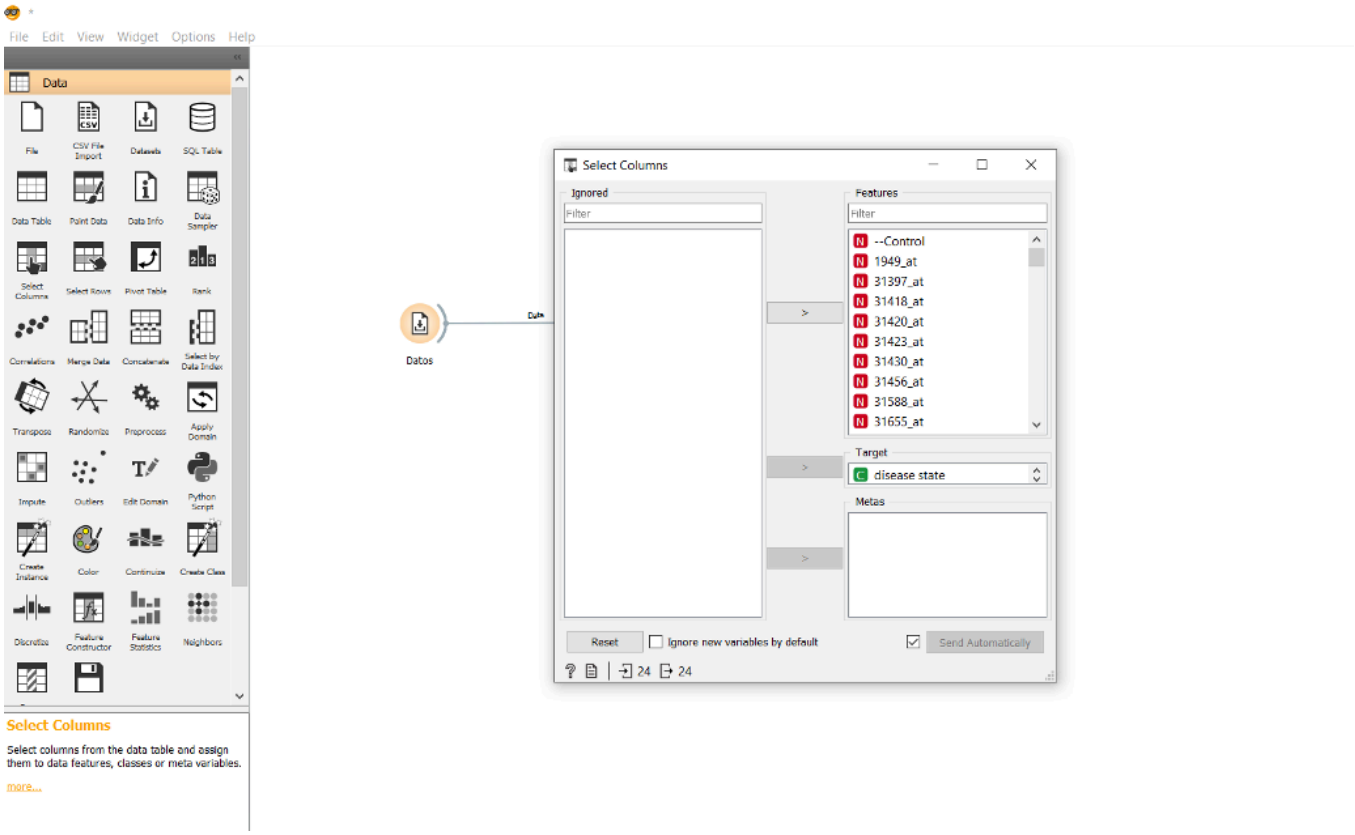


Ilustración 61. Panel de selección de columnas.

6.3.2. Bloque de visualización

A continuación, se pueden usar herramientas de visualización de datos para analizar las características y las particularidades de cada una de las columnas de manera gráfica. En este ejemplo utilizaremos dos tipos de gráficos usados habitualmente para la exploración de datos: los gráficos de dispersión (*scatter plots*) y los gráficos de barras (*bar plots*).

Los primeros son utilizados para confrontar dos variables. Por ejemplo, si analizásemos las características físicas de un grupo de personas, en un gráfico de dispersión podríamos mostrar la relación entre la altura y el peso de dichas personas. En cuanto a los gráficos de barras, suelen utilizarse para observar la distribución o la frecuencia de una sola variable.

Como se puede observar en la imagen, se hace clic y arrastran las herramientas mencionadas y se unen con la burbuja de selección de columnas.

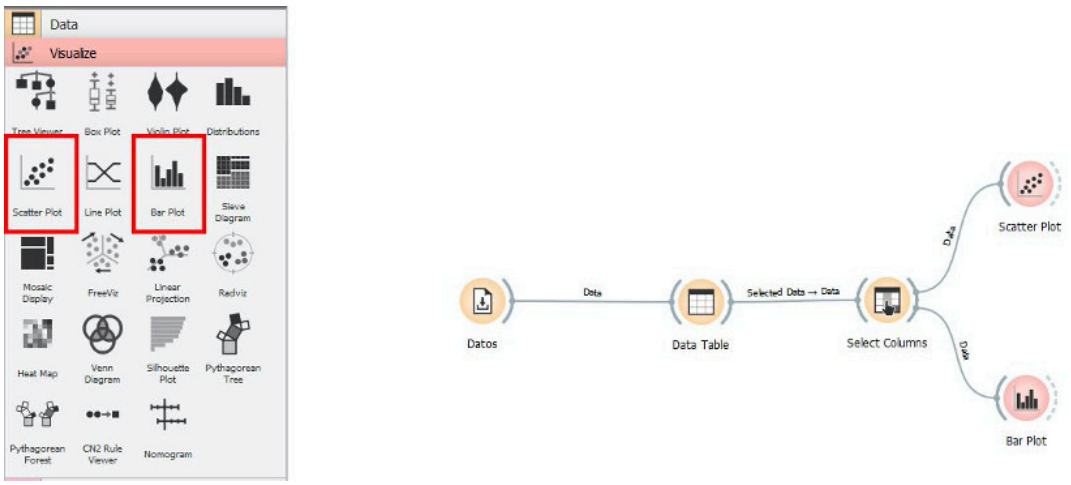


Ilustración 62. Gráficos.

Si se hace doble clic en la burbuja de scatter plot, se abre la ventana donde se puede observar el gráfico a partir de las variables que se determine. Arriba a la derecha, en “Axis X” y “Axis y” se seleccionan las variables que queramos mostrar en el gráfico. En el resto de las opciones de la parte de la izquierda podemos modificar el color, el tamaño y la forma del gráfico.

En la siguiente imagen se ve el diagrama de barras. En las opciones de arriba a la izquierda se pueden seleccionar las variables que se van a graficar.

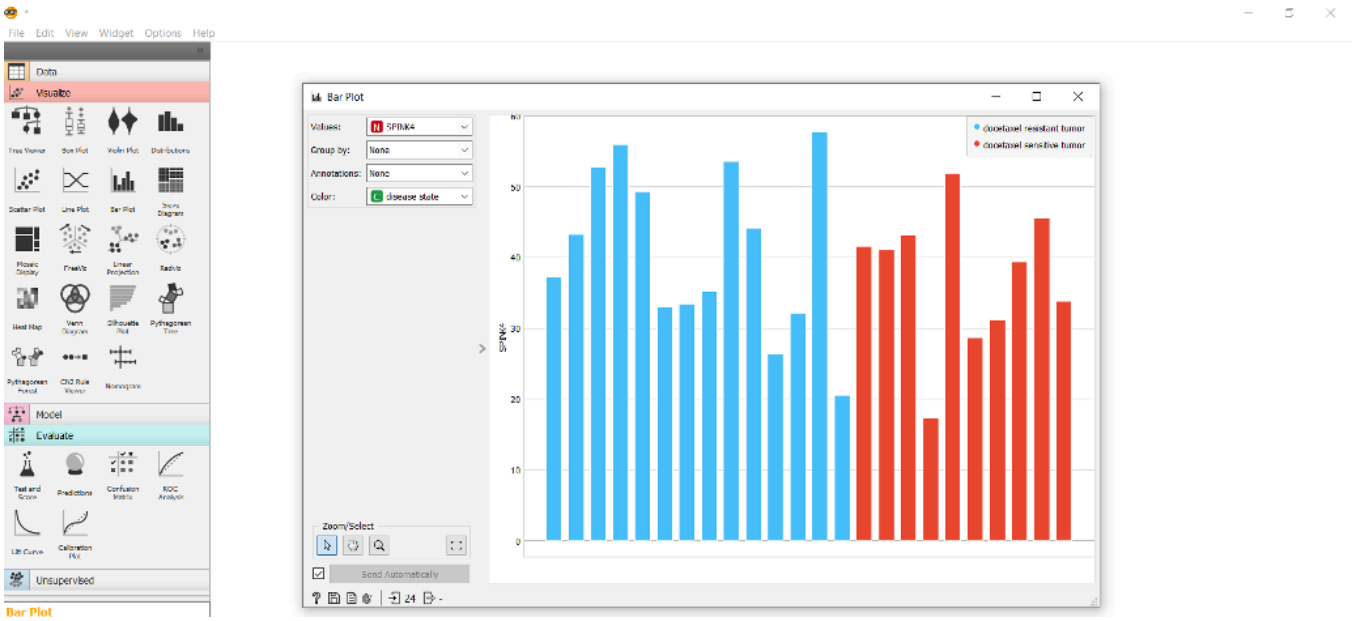


Ilustración 63. Diagrama de barras.

6.3.3. Bloque de modelos

A continuación, se va a proceder a introducir los datos en el modelo de aprendizaje automático. Para ello se va a usar el algoritmo de clasificación de *Random Forest*. Se arrastra la burbuja y se une con la herramienta de selección de columnas. De esta manera, le estaremos indicando al algoritmo cuáles son las columnas que tiene que emplear para la detección de patrones, y cuál es la columna o variable sobre la cual tiene que predecir.

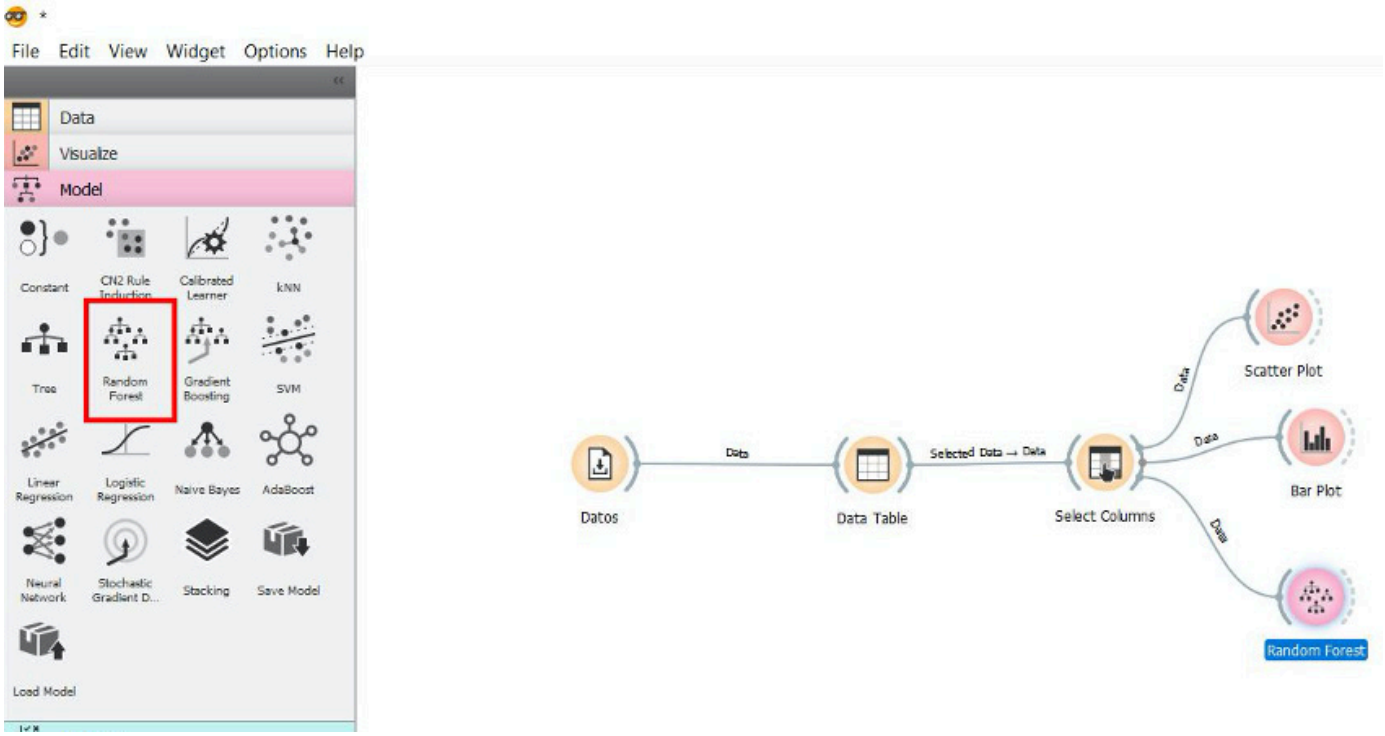


Ilustración 64.Algoritmo Random Forest.

Para la realización de las predicciones podemos seleccionar tantos modelos como queramos (siempre que sean de clasificación), y posteriormente podremos comparar la eficacia de las predicciones de cada uno de ellos.

6.3.4. Bloque de validación

Una vez que se ha elegido un modelo de aprendizaje automático, se procede a la validación del mismo. Para realizar esta validación existen numerosas métricas que se pueden utilizar para determinar la bondad de las predicciones de nuestro modelo, o en qué medida ha sido capaz de adaptarse a la variabilidad de nuestros datos.

Para explicar estas métricas de validación, en primer lugar, se tiene que introducir el concepto de matriz de confusión. En problemas binarios, es decir, en aquellos conjuntos de datos en los que la variable de salida puede tomar dos valores (sí/no, alto/bajo, tiene cáncer/no tiene cáncer, etc.), se puede emplear una matriz en la que vemos la relación entre el número de predicciones realizadas correctamente.

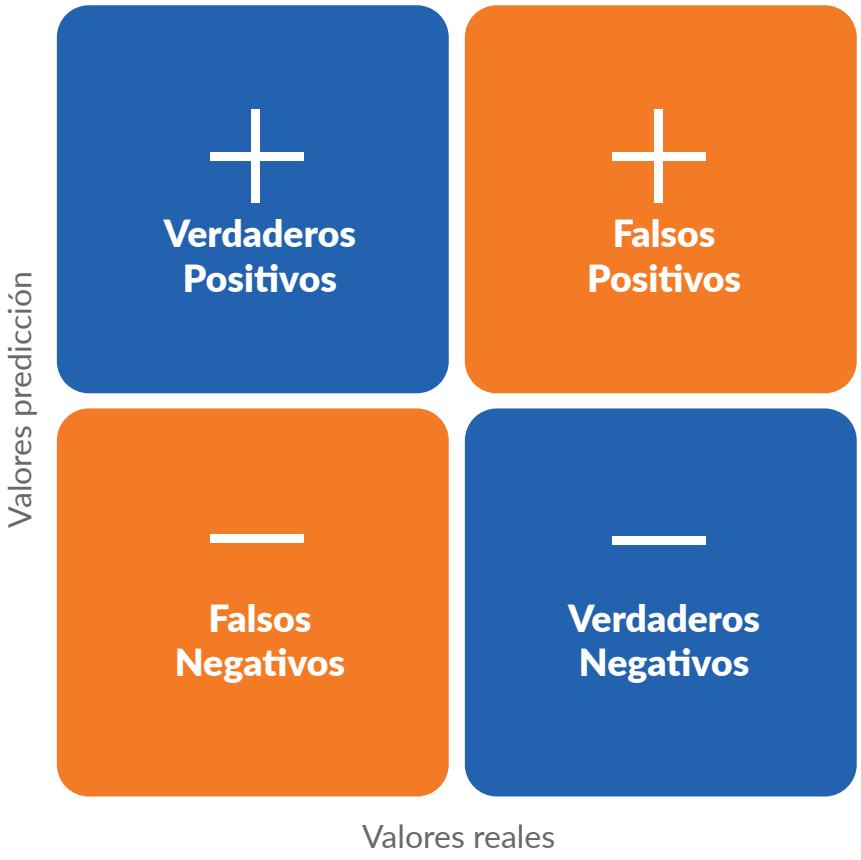


Ilustración 65.Matriz de confusión.

De esta manera, nuestro objetivo a la hora de realizar las predicciones es maximizar los valores de la diagonal verde y minimizar los de la diagonal roja. A partir de esta matriz, se pueden generar algunas métricas que caracterizarán la bondad del modelo.

Se considerará como referencia la precisión en este manual.

Si utilizamos la burbuja *Test and Score* del bloque de validación, podemos evaluar la bondad del modelo de cara a la clasificación. Si conectamos las burbujas relativas a la selección de columnas *(Select Columns)* y al modelo predictivo (*Random Forest*) a la nueva herramienta que hemos añadido, el programa evaluará las diferentes métricas de clasificación para el problema estudiado.

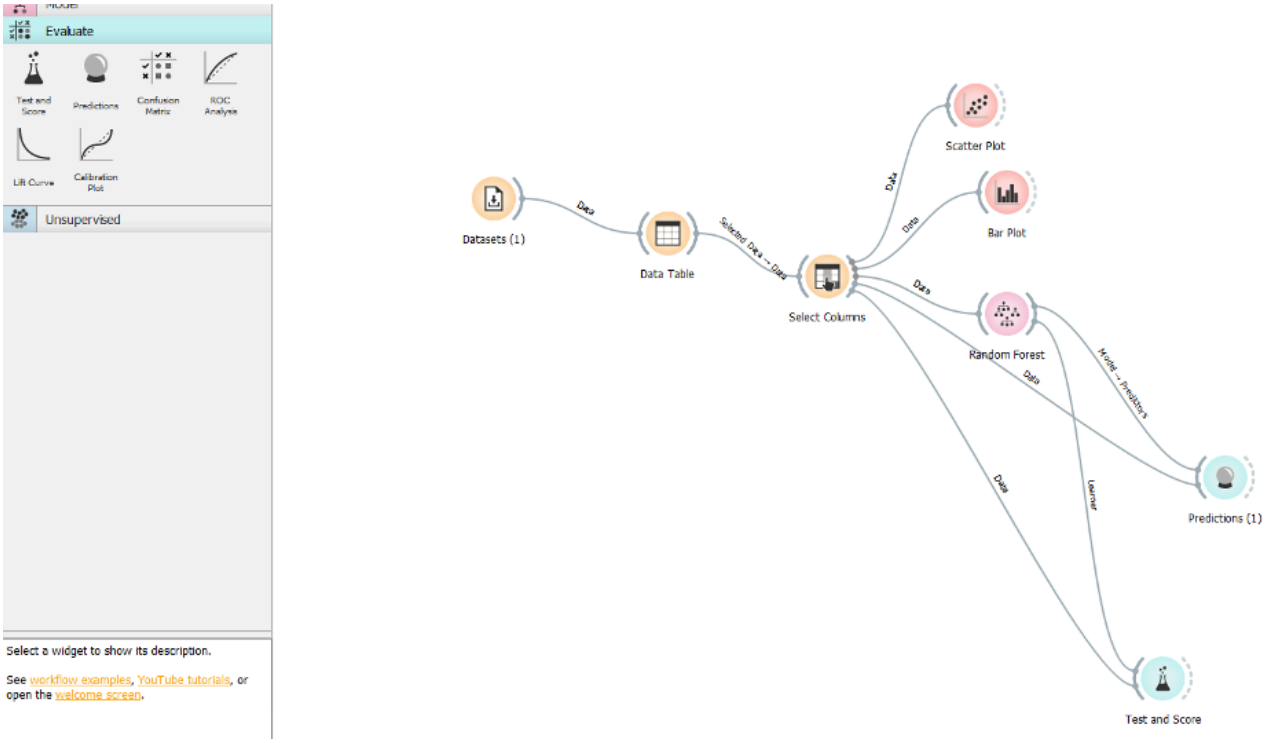
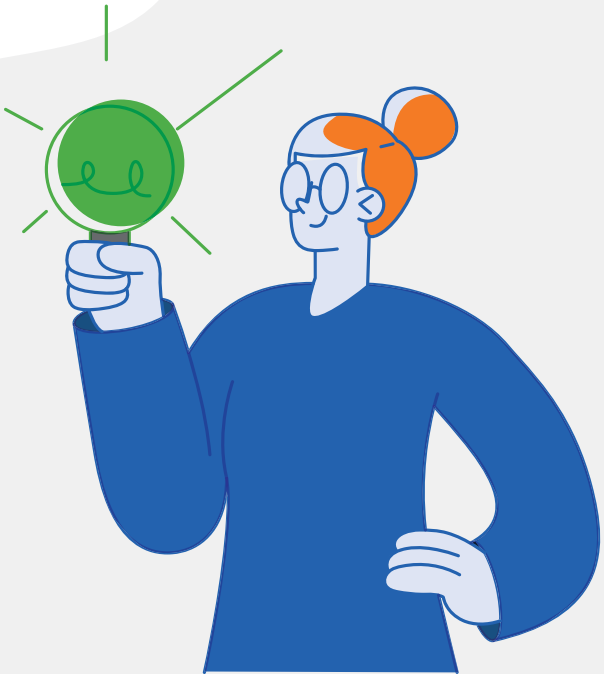


Ilustración 66.Métodos de evaluación.

Haciendo doble clic sobre *Test and Score* podemos comprobar la precisión alcanzada por el modelo. Queda a juicio del analista decidir si la precisión alcanzada puede considerarse suficiente, ya que cada problema que se estudia tiene características particulares, y debe diagnosticar si el problema se encuentra en los datos, en el modelo o en el planteamiento general del problema.

Para observar de manera más detallada cómo ha actuado el modelo ante este problema, podemos utilizar la ya mencionada matriz de confusión. Desde el bloque de validación arrastramos la burbuja *Confusion Matrix* y la conectamos con la burbuja *Test and Score*.



7

PREPARACIÓN DEL MODELO PREDICTIVO

7. Preparación del modelo predictivo

Tal y como hemos visto hasta ahora, el modelo predictivo que alimenta a la Plataforma se basa en el uso de información agroclimática y de datos socioeconómicos con el objetivo de predecir una categoría sintética denominada Zona de preocupación, que se ha generado a partir de las encuestas tomadas en campo por los técnicos del Consorcio de ONG.Anteriormente generamos un conjunto de datos en el que teníamos cuatro bloques de información:

- **Identificadores:**

  Columnas que contenían información básica sobre el municipio, como son el país y el departamento al que pertenece, el nombre del propio municipio y un identificador único.

- **Datos socioeconómicos:**

  Variables que identificaban las características demográficas, económicas e infraestructurales de cada municipio.

- **Datos agroclimáticos:**

  Estadísticas zonales de los índices utilizados (SPI, NDVI, ARVI, SIPI, EVI), tomados en un rango de 12 meses. A la hora de entrenar el modelo predictivo se utilizarán los 12 meses anteriores a la toma de datos en campo, mientras que para realizar las predicciones oportunas se tomarán los 12 meses previos al mes en el que se desea realizar la predicción.

- **Zonas de preocupación:**

  Cada municipio se clasifica en una zona de preocupación de acuerdo con la categoría de Seguridad Alimentaria y Nutricional a la que pertenezca, y puede ser alta o baja.

Para el presente caso, se tendrá en consideración lo siguiente:

- La toma de datos en campo se realizó en octubre de 2020.
- Se tomarán datos agroclimáticos desde octubre de 2019 hasta octubre de 2020 (12 meses) para el entrenamiento.
- Se generarán conjuntos de predicción adicionales en función de la cantidad de meses de los cuales se dispongan datos.
- En esos conjuntos de predicción, se tomarán los 12 meses anteriores para los datos agroclimáticos. Por ejemplo, si queremos realizar la predicción del mes de noviembre, tomaremos los datos agroclimáticos de noviembre de 2019 hasta noviembre de 2020.

7.1 Introducción de datos

Existen múltiples formatos para la introducción de datos en el software Orange, aunque en este manual se utilizará el formato CSV, ya que se trata del formato más común de transferencia de información cuando se dispone de bajos volúmenes de datos.

Empezamos por la introducción de los datos de entrenamiento, es decir, aquellos municipios en los cuales sí tenemos encuestas de campo, y, por lo tanto, conocemos las zonas de preocupación a las que pertenece.

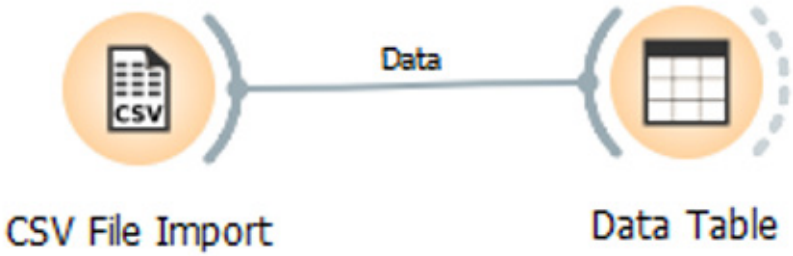


Ilustración 67. Burbujas necesarias para la introducción y la exploración de datos.

De esta manera, necesitaremos utilizar las burbujas de “CSV File Import” para introducir el fichero de datos de entrenamiento y “Data Table” para explorar la tabla y asegurarnos de que la entrada de datos ha sido correcta.

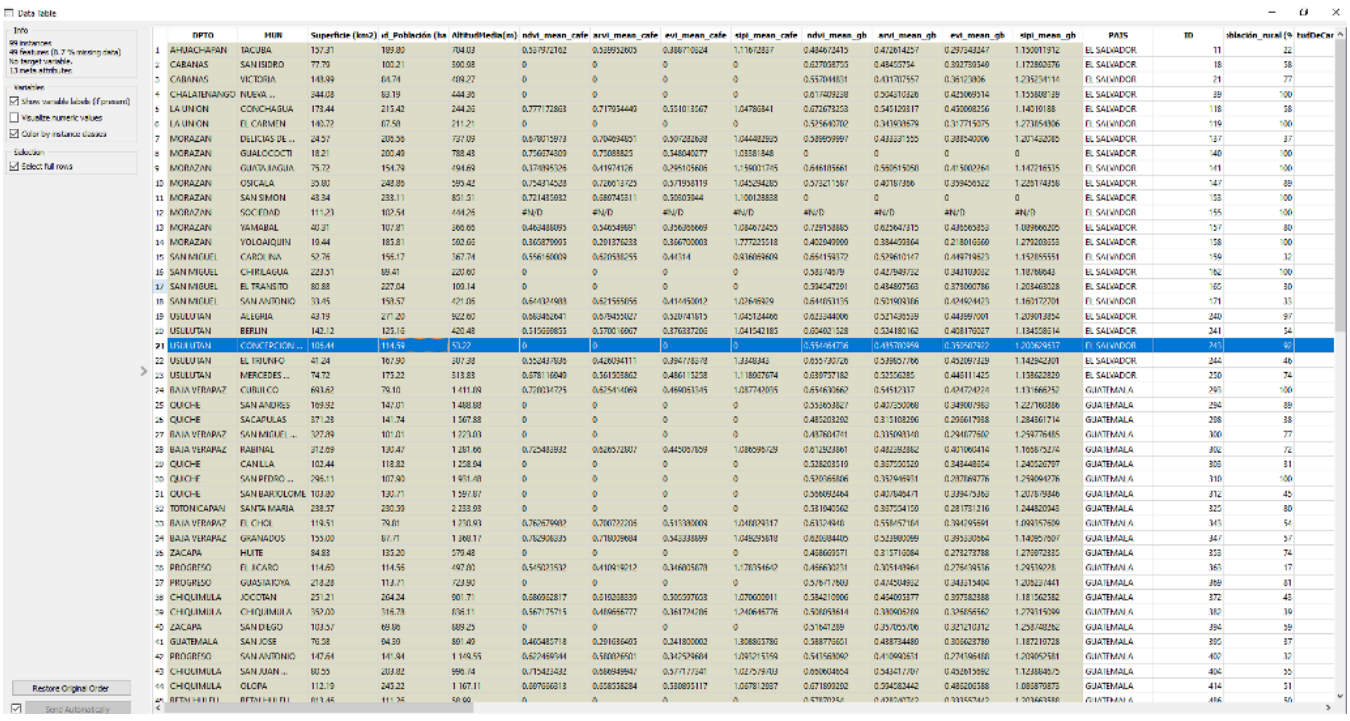


Ilustración 68. Tabla de datos para el entrenamiento.

7.2 Generación del modelo predictivo

Una vez que los datos de entrenamiento han sido introducidos en el software, se procede a la selección de variables, donde le diremos al modelo cuáles son las variables que debe usar para el ajuste de los modelos y cuáles debe usar para la predicción.

Una vez que introducimos la burbuja Select Columns, procedemos a decirle al programa que la variable a predecir o Target es la columna "Zona preocupación", y nos aseguramos de que las columnas "País", "Departamento", "Municipio" e "ID" se encuentren en el recuadro de Metas, ya que esta información no debe ser incluida en el modelo para el cálculo.

A continuación, introducimos la burbuja de Preprocess o preprocesamiento, que nos servirá para imputar datos faltantes y para estandarizar las variables de manera que todas tengan de media 0 y de desviación típica 1, lo cual facilita la tarea de los modelos a la hora de ajustarse a la variabilidad de los datos introducidos.

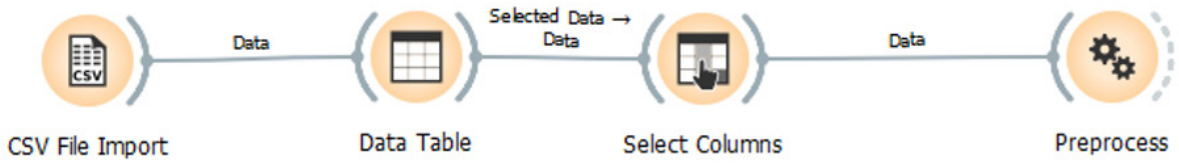


Ilustración 69. Introducción de la burbuja de preprocesamiento.

De cara al preprocesamiento, seleccionaremos como acciones la imputación de datos con la media o el valor más frecuente de cada columna, y estandarizar las variables numéricas.

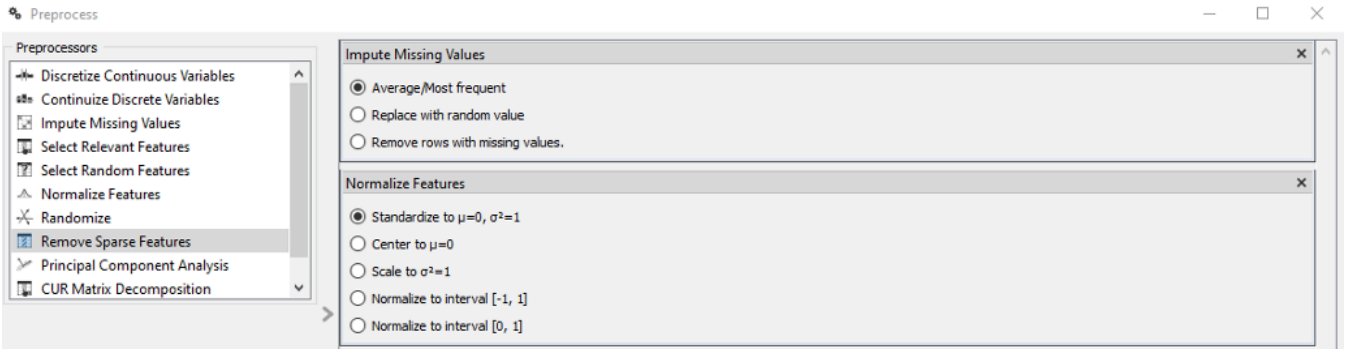


Ilustración 70. Selección de parámetros para el preprocesamiento.

Con estos pasos realizados, ya se pueden introducir las burbujas necesarias para la generación del modelo predictivo, es decir, los diferentes modelos predictivos que queramos probar, así como la burbuja de pruebas y validación.

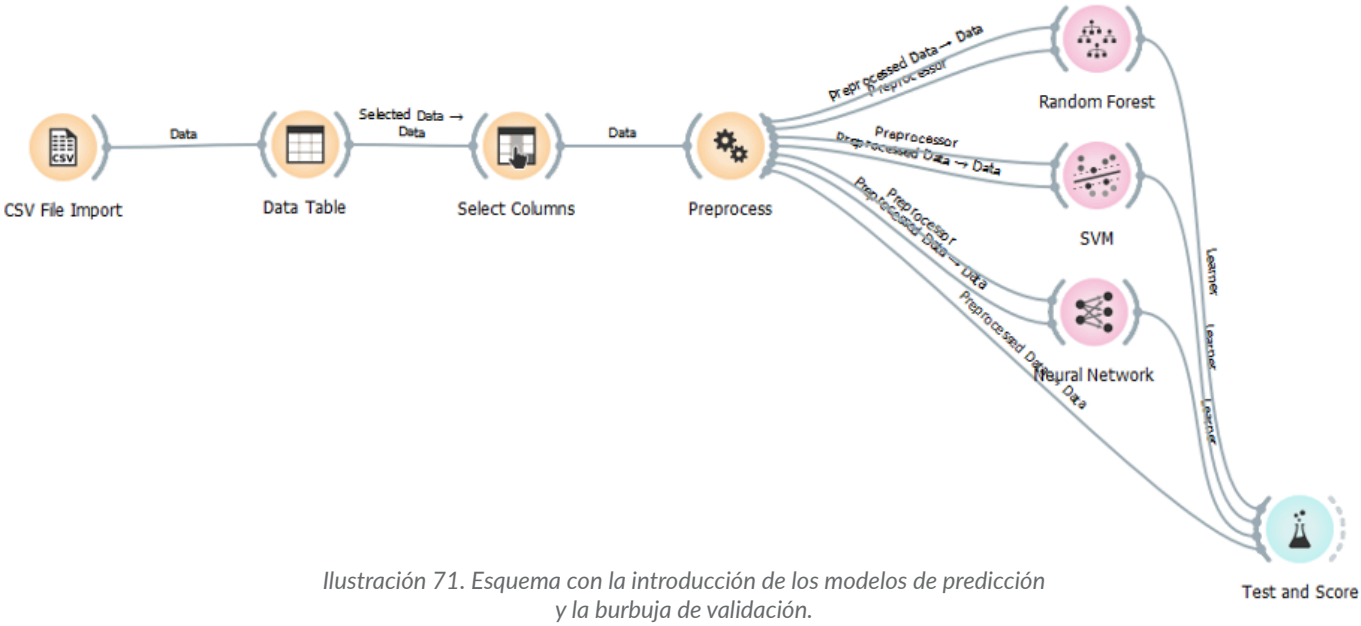


Ilustración 71. Esquema con la introducción de los modelos de predicción y la burbuja de validación.

Si abrimos la burbuja de Test and Score, podemos comprobar cómo ha sido la precisión de cada uno de los modelos a la hora de generar las predicciones. Se puede comprobar que el mejor modelo en términos de precisión ha sido el Random Forest. A partir de este punto, se puede considerar que ha finalizado la fase de entrenamiento de los modelos.

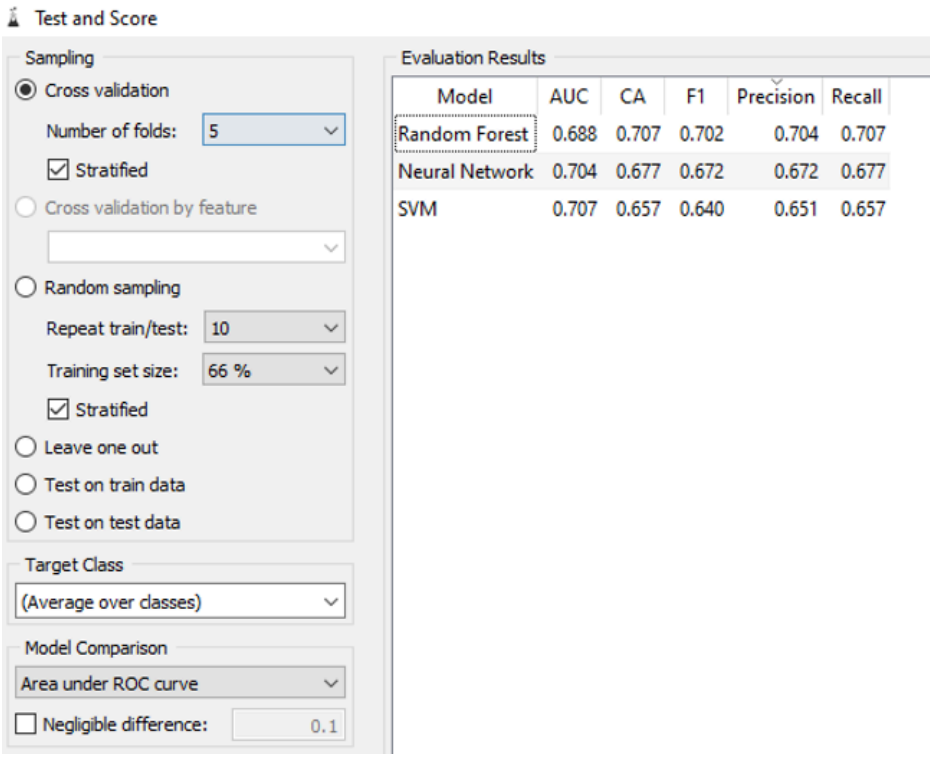


Ilustración 72. Resultados del modelo a partir del conjunto de entrenamiento.

7.3 Introducción del cojunto de predicción

Una vez que hemos entrenado el modelo predictivo con los conjuntos de entrenamiento, procedemos a introducir los conjuntos de predicción, es decir, aquellos registros en los cuales no se han tomado datos en campo, ya sea porque ese mes no hubo campañas de encuestas, o porque la campaña no cubrió determinados municipios.

Por lo tanto, realizamos el mismo proceso que en el caso anterior, con la diferencia de que en este caso incluiremos el archivo relativo al conjunto de predicción, que se diferencia del anterior en que:

- El archivo de predicción tiene la columna de Zona de preocupación vacía.
- El archivo de predicción contiene todos los municipios, se hayan realizado o no encuestas en ellos.

Asimismo, seleccionamos como variable respuesta o variable a predecir “Zona de preocupación”, pese a que en este conjunto no contenga información.

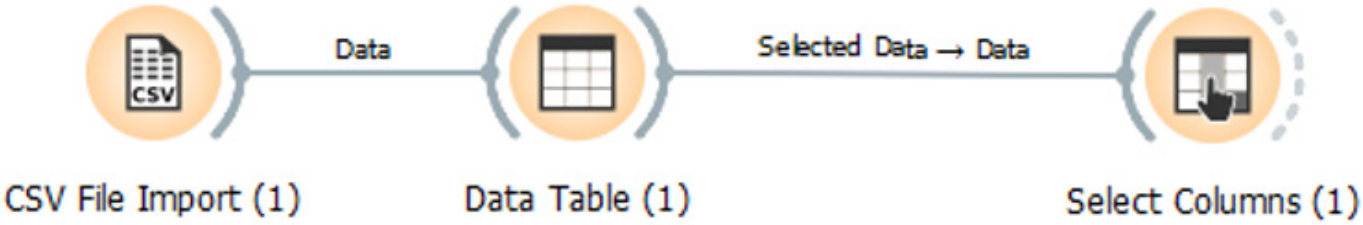


Ilustración 73. Esquema de introducción, exploración y selección de columnas para los datos de predicción.

A continuación, introducimos la burbuja de Predictions, que se enlazará tanto con los modelos predictivos que hemos generado como con el conjunto de datos que acabamos de introducir. De este modo, el esquema general quedaría como se muestra en la siguiente imagen.

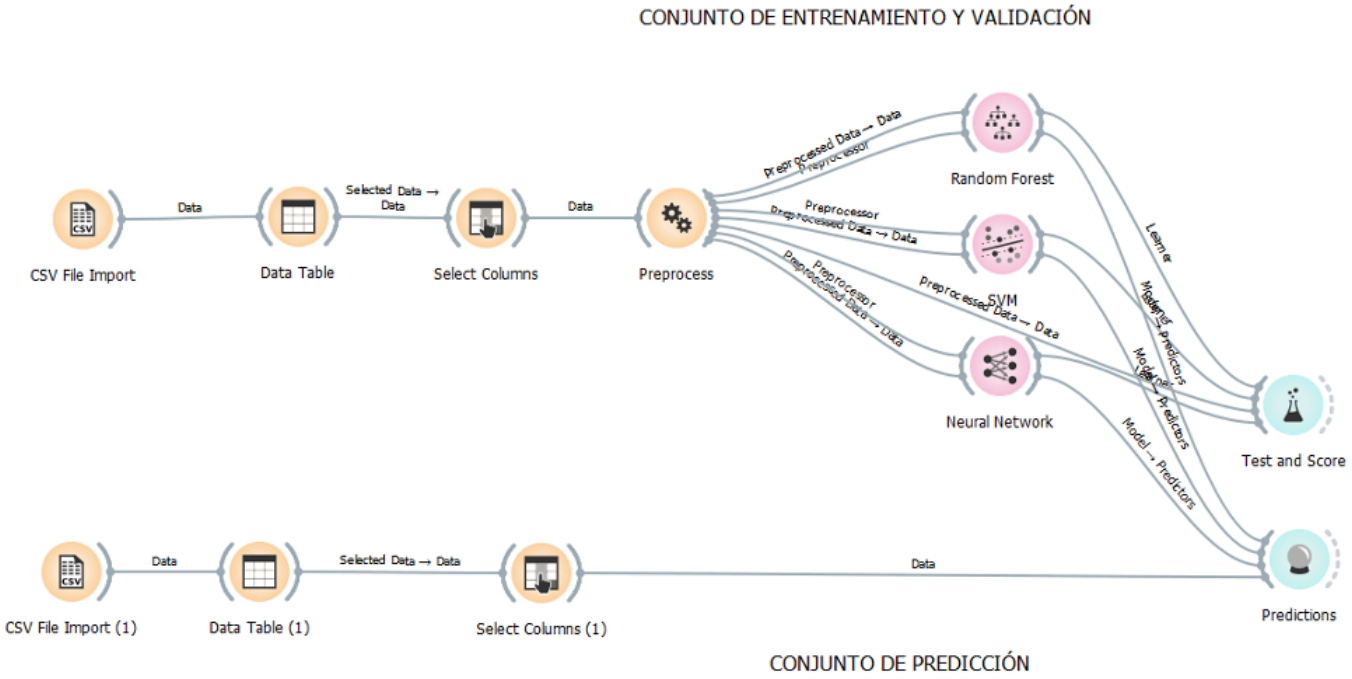


Ilustración 74. Esquema general de la generación del modelo predictivo.

Haciendo doble clic sobre la burbuja Predictions, obtenemos las predicciones realizadas por cada modelo para cada uno de los municipios del conjunto de predicción. Se puede comprobar que, para un mismo municipio, los diferentes modelos dan resultados diferentes.

	Neural Network	SVM	Random Forest	ID	DPTO	MUN	PAIS	Población	Superficie,(km2)	ad_Población,(ha	samientosDeServi	AltitudMedia(m)	Sismicos(adim.)	tilizamientos(adim	ulcanismo(adim.	undaciones(adim	anesY lormentas)(	RatioMujeres
1	0	0	0	1	AHUACHAPAN	AHUACHAPAN	EL SALVADOR	110511	251.09	440.12	4	669.77	4	2	2	1	2	52.21
2	0	0	1	2	AHUACHAPAN	APANECA	EL SALVADOR	8383	40.74	205.79	1	1389.95	4	4	2	1	3	51.23
3	0	0	1	3	AHUACHAPAN	ATIQUIZAYA	EL SALVADOR	33587	62.99	533.19	0	809.45	2	3	0	1	3	51.65
4	0	0	1	4	AHUACHAPAN	CONCEPCION ...	EL SALVADOR	12786	68.31	187.18	6	1111.29	4	3	2	1	1	50.92
5	0	0	0	5	AHUACHAPAN	EL REFUGIO	EL SALVADOR	8171	11.33	721.41	0	696.17	4	4	2	1	3	52.32
6	1	0	1	6	AHUACHAPAN	GUAYMANGO	EL SALVADOR	19037	78.73	241.81	0	287.89	2	3	0	1	3	50.17
7	1	0	1	7	AHUACHAPAN	JUJUTLA	EL SALVADOR	28599	179.8	159.06	1	255.53	4	1	2	1	1	51.11
8	1	0	1	8	AHUACHAPAN	SAN FRANCISC...	EL SALVADOR	42607	266.69	159.76	6	172.7	4	4	2	1	3	51.25
9	0	0	0	9	AHUACHAPAN	SAN LORENZO	EL SALVADOR	9194	42.73	215.17	0	544.04	2	2	0	1	3	49.35
10	0	0	1	10	AHUACHAPAN	SAN PEDRO ...	EL SALVADOR	7773	25.07	310.04	0	570.77	4	4	2	1	2	50.03
11	1	1	1	11	AHUACHAPAN	TACUBA	EL SALVADOR	29858	157.31	189.8	1	704.03	4	4	2	1	3	50.38
12	0	0	0	12	AHUACHAPAN	TURIN	EL SALVADOR	8997	9.99	900.37	2	653.65	2	2	0	1	3	52.62
13	1	0	0	13	CABAÑAS	CINQUERA	EL SALVADOR	1467	31.95	45.92	0	447.09	4	1	2	1	1	48.94
14	0	0	0	14	CABAÑAS	DOLORES	EL SALVADOR	6847	138.84	45.71	0	195.25	4	4	2	1	3	51.73
15	0	0	0	15	CABAÑAS	GUACOTECTI	EL SALVADOR	5530	28.31	195.02	0	623.02	2	3	0	1	3	53.3
16	1	0	0	16	CABAÑAS	ILOBASCO	EL SALVADOR	61510	234.8	261.97	0	554.7	4	1	2	1	1	52.09
17	0	0	0	17	CABAÑAS	JUTIAPA	EL SALVADOR	6534	86.3	76.29	0	332.76	4	4	2	1	3	52.51
18	0	0	0	18	CABAÑAS	SAN ISIDRO	EL SALVADOR	7796	77.79	100.21	0	390.98	2	3	0	1	3	51.65
19	0	0	0	19	CABAÑAS	SENSUNTEPEQUE	EL SALVADOR	40332	310.92	129.72	0	404.59	4	2	2	2	1	53.64
20	0	0	0	20	CABAÑAS	TEJUTEPEQUE	EL SALVADOR	7114	42.4	167.79	0	651.81	4	4	2	1	3	53.08
21	0	0	0	21	CABAÑAS	VICTORIA	EL SALVADOR	12626	148.99	84.74	0	409.27	2	1	0	3	2	53.33
22	0	0	0	22	CHALATENANGO	AGUA CALIENTE	EL SALVADOR	3261	120.59	64.24	0	566.99	4	2	2	3	2	53.76
23	0	0	1	23	CHALATENANGO	ARCATAO	EL SALVADOR	2946	33.46	88.04	1	556.47	4	4	2	1	3	49.59
24	0	0	0	24	CHALATENANGO	AZACUALPA	EL SALVADOR	1136	12.09	93.96	0	423.38	2	2	0	1	3	51.58
25	0	0	0	25	CHALATENANGO	CANCASQUE	EL SALVADOR	1751	30.08	58.22	0	799.79	4	1	2	1	2	49.4
26	0	0	0	26	CHALATENANGO	CHALATENANGO	EL SALVADOR	29271	131.66	222.32	10	428.28	4	4	2	1	3	51.62
27	1	0	0	27	CHALATENANGO	CITALA	EL SALVADOR	4164	72.81	57.19	1	1189.42	4	2	2	1	1	52.83
28	0	0	0	28	CHALATENANGO	COMALAPA	EL SALVADOR	2696	26.24	114.19	0	578.87	4	4	2	1	3	53.6
29	0	0	0	29	CHALATENANGO	CONCEPCION ...	EL SALVADOR	6457	49.39	130.74	0	660.11	4	4	2	1	2	52.78
30	0	0	0	30	CHALATENANGO	DULCE NOMBR...	EL SALVADOR	5051	67.5	74.83	0	895.44	4	4	2	1	3	52.94
31	0	0	0	31	CHALATENANGO	EL CARRIZAL	EL SALVADOR	2464	28.28	87.14	0	743.09	4	1	2	1	1	53.25
32	0	0	0	32	CHALATENANGO	EL PARAISO	EL SALVADOR	10483	57.8	181.36	0	773.94	3	4	2	1	3	51.67
33	0	0	0	33	CHALATENANGO	LA LAGUNA	EL SALVADOR	3923	24.6	159.47	0	804.89	4	4	2	1	3	51.11
34	0	0	1	34	CHALATENANGO	LA PALMA	EL SALVADOR	12233	130.94	81.06	0	1256.23	4	4	3	2	3	52.75
35	0	0	0	35	CHALATENANGO	LA REINA	EL SALVADOR	9525	104.86	90.83	0	608.01	4	4	2	1	3	52.72
36	0	0	0	36	CHALATENANGO	LAS FLORES	EL SALVADOR	1503	25.98	60.94	0	366.08	4	4	3	2	3	48.09
37	0	0	0	37	CHALATENANGO	LAS VUELTAS	EL SALVADOR	940	34.27	27.43	0	820.84	4	4	2	1	3	51.01
38	0	0	0	38	CHALATENANGO	NOMBRE DE ...	EL SALVADOR	4484	42.55	105.37	0	388.57	3	3	2	1	1	52.77
39	0	0	0	39	CHALATENANGO	NUEVA ...	EL SALVADOR	28505	344.08	83.19	0	444.36	3	4	1	2	2	52.87
40	0	0	0	40	CHALATENANGO	NUEVA TRINIDAD	EL SALVADOR	1673	39.73	42.11	0	378.02	3	2	0	3	4	48.71
41	1	0	0	41	CHALATENANGO	OJOS DE AGUA	EL SALVADOR	3667	42.18	86.93	0	654.16	4	3	2	1	1	55.19
42	0	0	0	42	CHALATENANGO	POTONICO	EL SALVADOR	1586	37.73	42.03	2	804.82	3	4	1	2	2	52.21
43	0	0	1	43	CHALATENANGO	SAN ANTONIO ...	EL SALVADOR	1654	31.79	50.33	0	293.00	2	3	0	1	3	50.7

Ilustración 75. Resultados arrojados por los modelos predictivos.

Una vez llegados a este punto, se debe atender a la distribución de las probabilidades para la predicción de las zonas de preocupación. Si recordamos la definición establecida para las zonas de preocupación, se expresan en términos de la probabilidad de que un municipio pertenezca a una categoría SAN de inseguridad moderada o severa. Por lo tanto, lo que nos resulta de interés para el modelo no es tanto la clasificación realizada, sino las probabilidades, sabiendo que los modelos de clasificación empleados funcionan de modo que, si la probabilidad de pertenencia a una clase es superior al 50%, se le asigna dicha clase.

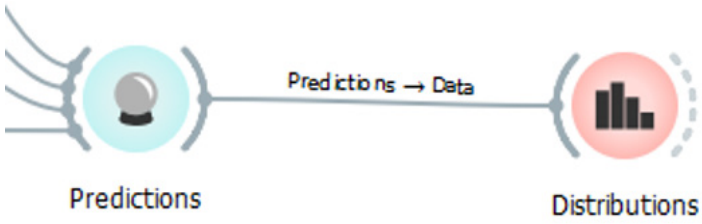


Ilustración 76. Introducción de la burbuja de distribuciones para el estudio de las probabilidades resultantes.

Por lo tanto, a la burbuja *Predictions* le uniremos otra denominada *Distribution*", que nos permitirá ver cómo es la distribución de la probabilidad de las predicciones hechas. Haciendo doble clic sobre la burbuja *Distributions*, tendremos la opción de seleccionar, para cada modelo predictivo, la distribución de probabilidad de pertenencia a la clase 1, que en este caso representa las zonas de alta preocupación.

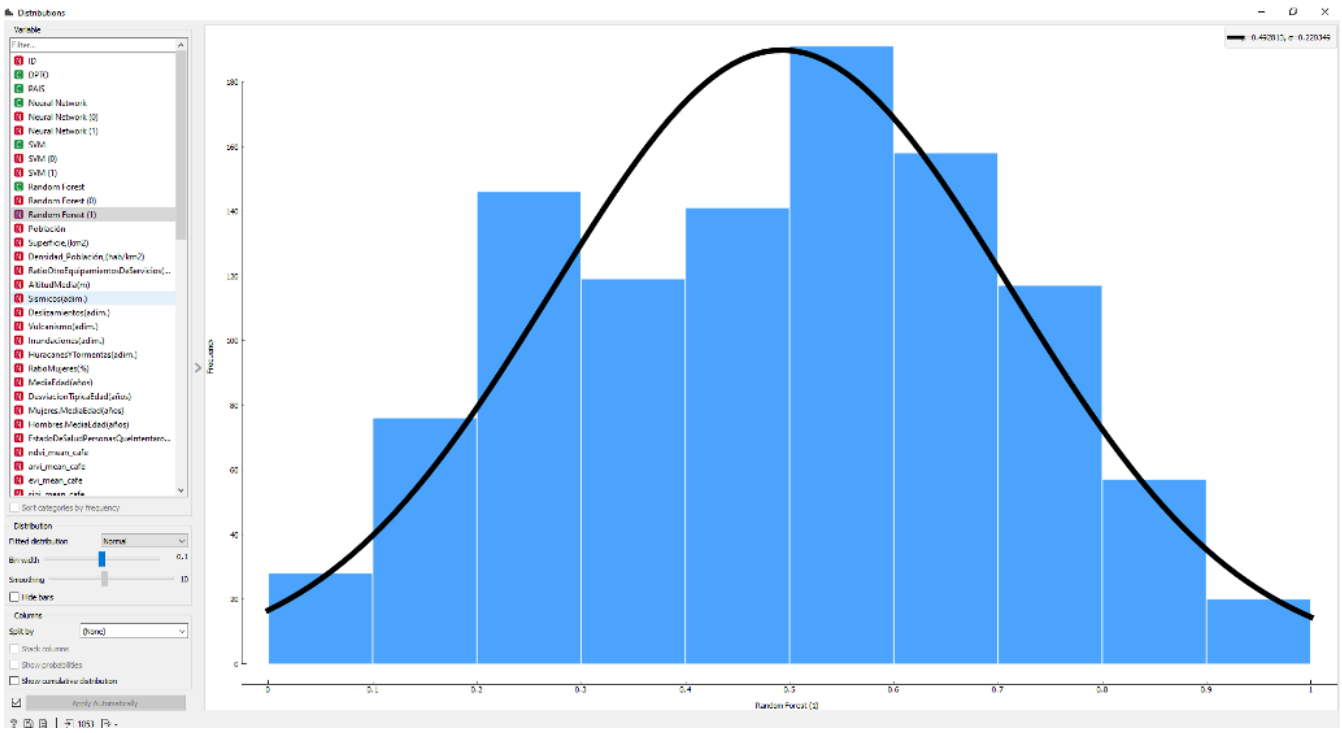


Ilustración 77. Probabilidad de pertenencia a las zonas de preocupación altas para el modelo Random Forest.

Si comparamos las distribuciones resultantes para cada modelo, veremos que existe una gran diferencia en términos de probabilidades. Pese a que los resultados puedan ser similares en cuanto a precisión, se puede observar que las diferencias de distribución son notables, pues el *Random Forest* devuelve una distribución próxima a la normal, mientras que las redes neuronales arrojan una distribución bimodal. Dado que buscamos una distribución más parsimoniosa, debemos elegir aquellos modelos que devuelvan resultados cercanos a una distribución normal, por lo que en este caso nos quedamos con el modelo *Random Forest*.

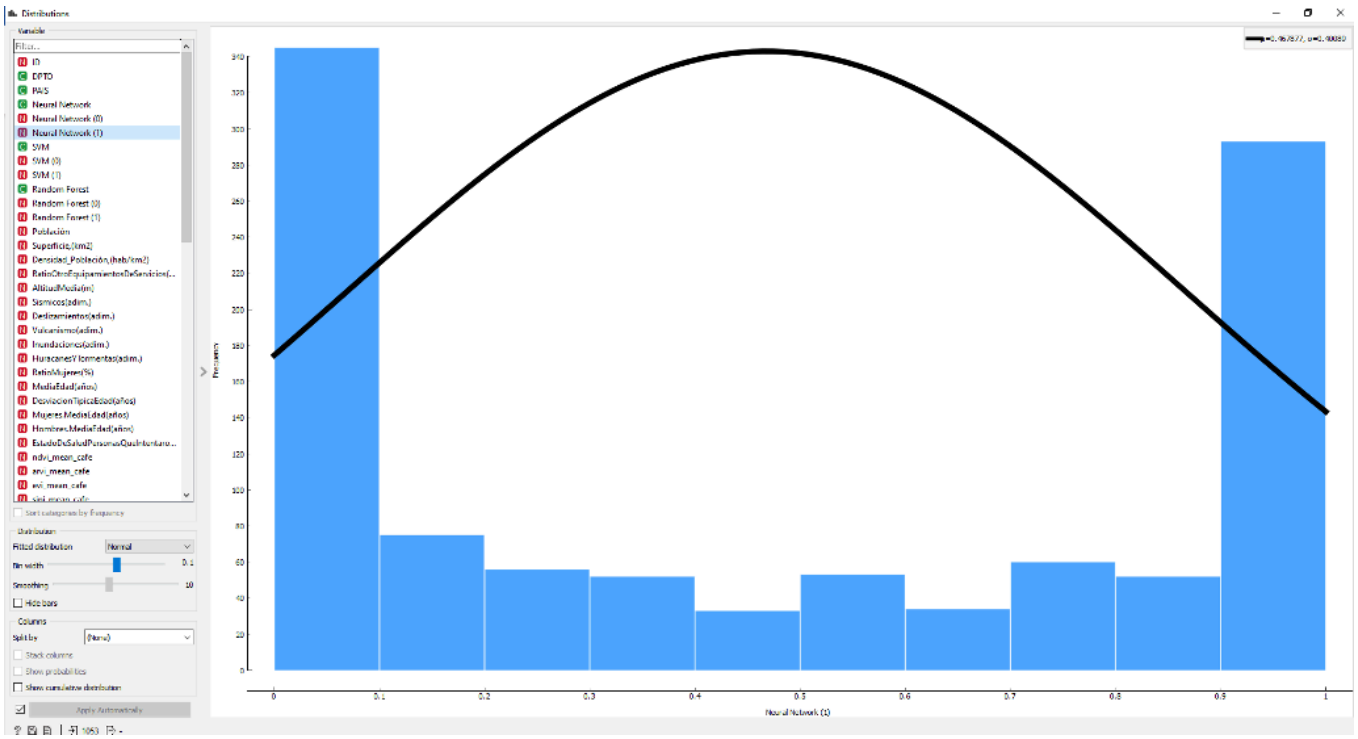


Ilustración 78. Probabilidad de pertenencia a las zonas de preocupación altas para el modelo de Redes Neuronales.

El siguiente paso consistirá en guardar los resultados obtenidos para realizar las transformaciones oportunas sobre las probabilidades obtenidas. Para ello, agregamos la burbuja *Save Data*, y guardamos las predicciones en la carpeta de nuestra elección.

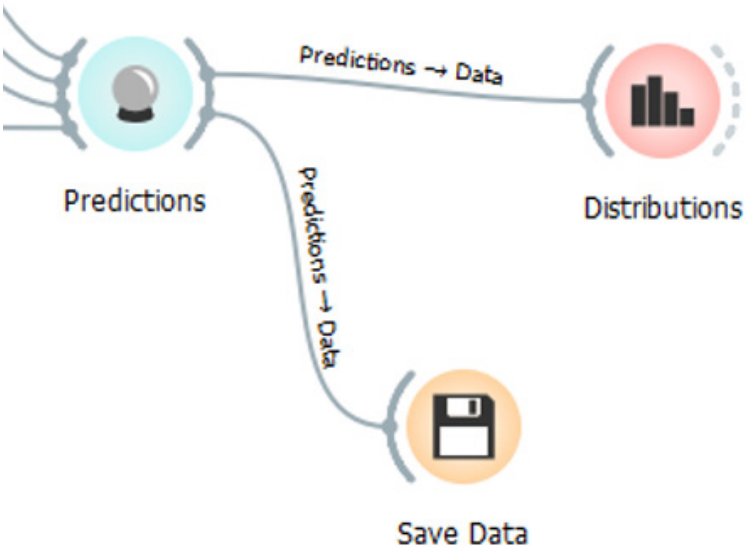


Ilustración 79. Guardado de las predicciones

El esquema de datos obtenidos se puede resumir en la siguiente imagen, donde encontraremos los datos de identificación de los registros, así como la columna de probabilidad de pertenencia a la clase 1 obtenida mediante el modelo *Random Forest*.

ID	DPTO	MUN	PAIS	Random Forest (1)
continuous	AHUACHAPA	string	EL\ SALVADOR GUATEMALA HONDURAS NICARAGUA	continuous
meta	meta	meta	meta	meta
1	AHUACHAPA	AHUACHAPA	EL SALVADOR	0.488888889
2	AHUACHAPA	APANECA	EL SALVADOR	0.50952381
3	AHUACHAPA	ATIQUIZAYA	EL SALVADOR	0.503888889
4	AHUACHAPA	CONCEPCIOI	EL SALVADOR	0.576190476
5	AHUACHAPA	EL REFUGIO	EL SALVADOR	0.455555556
6	AHUACHAPA	GUAYMANG(	EL SALVADOR	0.533333333
7	AHUACHAPA	JUJUTLA	EL SALVADOR	0.6
8	AHUACHAPA	SAN FRANCI	EL SALVADOR	0.6
9	AHUACHAPA	SAN LORENZ	EL SALVADOR	0.428888889
10	AHUACHAPA	SAN PEDRO I	EL SALVADOR	0.588888889
11	AHUACHAPA	TACUBA	EL SALVADOR	0.733333333
12	AHUACHAPA	TURIN	EL SALVADOR	0.463888889

Ilustración 80. Esquema de datos incluyendo las predicciones realizadas.

A continuación, se realiza una transformación sobre los datos para generar cuatro clasificaciones a partir de las predicciones realizadas:

- **Zonas de muy baja preocupación:**

  Probabilidad < Media – 1.5*Desviación típica.

- **Zonas de preocupación baja:**

  Media – 1.5*Desviación típica < Probabilidad < Media.

- **Zonas de preocupación alta:**

  Media < Probabilidad < Media + 1.5*Desviación típica.

- **Zonas de preocupación muy alta:**

  Media + 1.5*Desviación típica < Probabilidad.

El hecho de tomar 1.5 desviaciones típicas con respecto a la media es algo que queda a criterio del analista, ya que el umbral que decidamos tendrá un impacto en la clasificación que realicemos. Podemos plantear como ejemplo usar diferentes valores para la desviación típica y analizar cómo cambiaría la clasificación de los municipios:

	2 desviaciones	1.5 desviaciones	1 desviaciones
Muy baja	0	61	153
Baja	452	391	299
Alta	592	545	378
Muy alta	9	56	223

Ilustración 81. Distribución de los municipios en función de cuántas desviaciones típicas se tengan en cuenta.

Como se puede observar, si utilizamos 2 desviaciones típicas con respecto a la media, apenas tendremos municipios clasificados en los extremos, pero si usamos 1 desviación, obtendremos un escenario muy desfavorable, en el que hasta 223 municipios estarán clasificados como zonas de preocupación muy alta. En este caso, vamos a tomar como criterio 1.5 desviaciones típicas, puesto que es la distribución que más puede asemejarse a la realidad.

Para realizar esta transformación desde la probabilidad a la nueva división entre zonas de preocupación, podemos introducir la siguiente fórmula en Microsoft Excel:

```
=SI(BL2<PROMEDIO($BL$2:$BL$1054)-1.5
        *DESVEST.M ($BL$2:$BL$1054); "Muy baja"; SI (BL2
        <PROMEDIO ($BL$2:$BL$1054); "Baja"; SI (Y (BL2
        >PROMEDIO ($BL$2:$BL$1054); BL2
        <PROMEDIO ($BL$2:$BL$1054) + 1.5
        *DESVEST.M ($BL$2:$BL$1054)); "Alta" ; "Muy alta")))
```

Siendo BL2 la celda en la cual se encuentra la probabilidad de pertenencia a la clase 1 obtenida mediante el *Random Forest* y BL2:BL1054 la columna donde se encuentran dichas probabilidades. De esta manera, cada municipio tendrá asociada una predicción en relación con la zona de preocupación a la que pertenece, y podremos introducirla en la Plataforma, como se verá en los siguientes apartados.

INTRODUCCIÓN A POWER BI

8

8. Introducción a Power BI

8.1 ¿Qué es y para qué sirve Power BI?

Power BI es una herramienta desarrollada por Microsoft para la exploración y el análisis avanzado de datos. Permite la introducción de datos desde diversas fuentes y formatos, ya sea de manera local o desde un punto de conexión remota, para su representación a través de un cuadro de mando con funcionalidades más avanzadas que una hoja de Microsoft Excel.

8.2 Arquitectura de la aplicación

En cuanto a la arquitectura de la aplicación, podemos dividir Power BI en 3 partes fundamentales:

- **Integración de datos:**

  En Power BI, podemos importar datos de diferentes tipos de fuentes de datos en diferentes formatos. En el paso de integración de datos, Power BI reúne los datos (extraídos) de diferentes fuentes de datos y los convierte en un formato estándar. Una vez que los datos se integran en Power BI, se almacenan en un área de almacenamiento común conocida como área de preparación.

- **Procesamiento de datos:**

  Una vez que Power BI integra y almacena los datos en un lugar seguro, los datos brutos requieren cierto procesamiento. Varias operaciones de procesamiento o limpieza transforman los datos brutos, como la eliminación de valores redundantes, etc. Posteriormente, aplicamos las reglas de negocio pertinentes sobre los datos procesados que los transforman de acuerdo con nuestras necesidades. Estos datos transformados se cargan en los almacenes de datos. Así se completa un proceso completo de extracción, transformación y carga.

- **Presentación de datos:**

  En esta fase final, los datos procesados salen del almacén y pasan a las plataformas de Power BI, como Power BI Desktop, para crear informes, cuadros de mando y cuadros de mando.

8.3 Conexiones de datos para Power BI

Como se ha comentado anteriormente, existen múltiples conexiones de datos posibles para Power BI, tanto en local como en remoto. Algunas de esas conexiones son las que se muestran a continuación:

- **Ficheros**

  (Excel, Text/CSV, PDF, JSON), bases de datos, Power BI, Azure, servicios online, etc.

9

INSTALACIÓN DE POWER BI DESKTOP

9. Instalación de Power BI Desktop

La instalación de Power BI Desktop es muy sencilla y consta de los siguientes pasos:

- En primer lugar, nos dirigiremos a la página oficial de Microsoft (**https://powerbi.microsoft.com/es-es/downloads/**) en la que se encuentra el enlace de descarga.

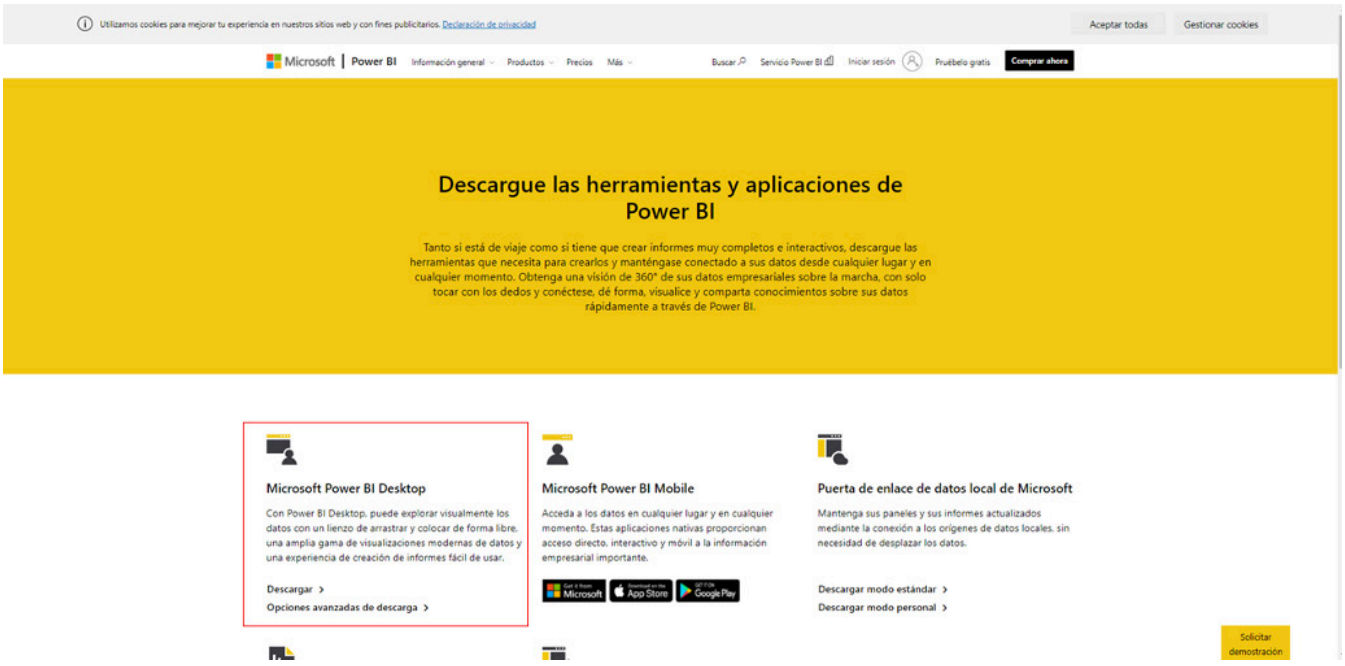


Ilustración 82. Web de descarga de Power Bi.

- En el cuadro señalado en rojo se encuentra el enlace de descarga. Procedemos a clicar en la opción “Opciones avanzadas de descarga”, y nos llevará a la siguiente página, donde cambiaremos el idioma a Español en el menú desplegable y haremos clic sobre el botón “Descarga”.

- Elegimos la versión que se corresponda con nuestro sistema (en este caso elegiremos la versión de 64 bits, que es la más habitual), y pulsamos “Next”

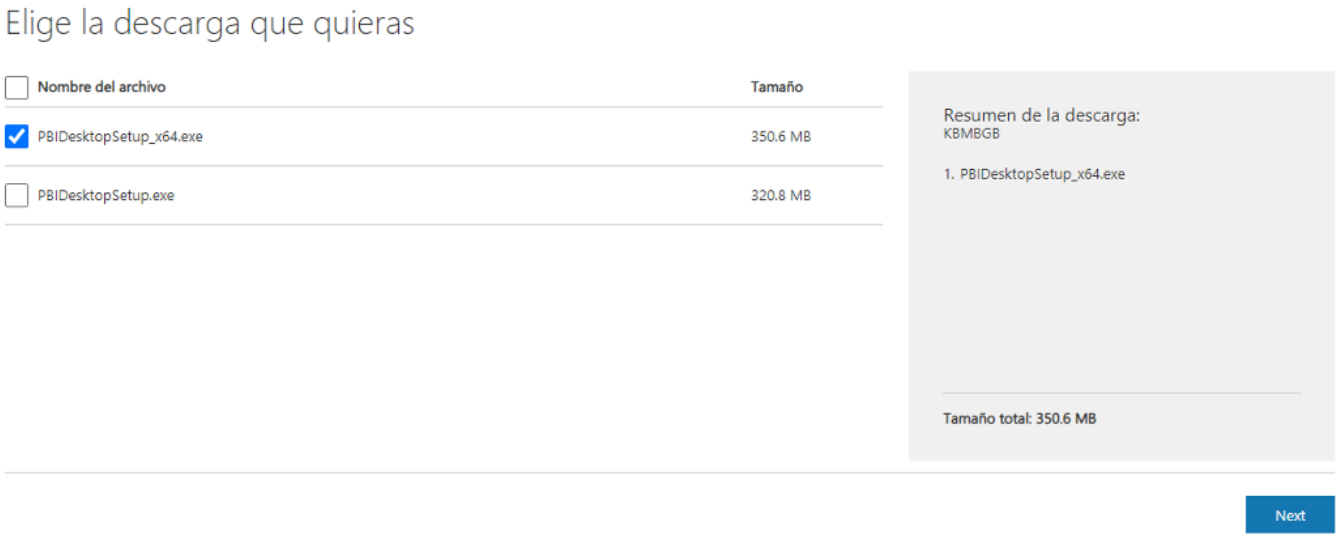
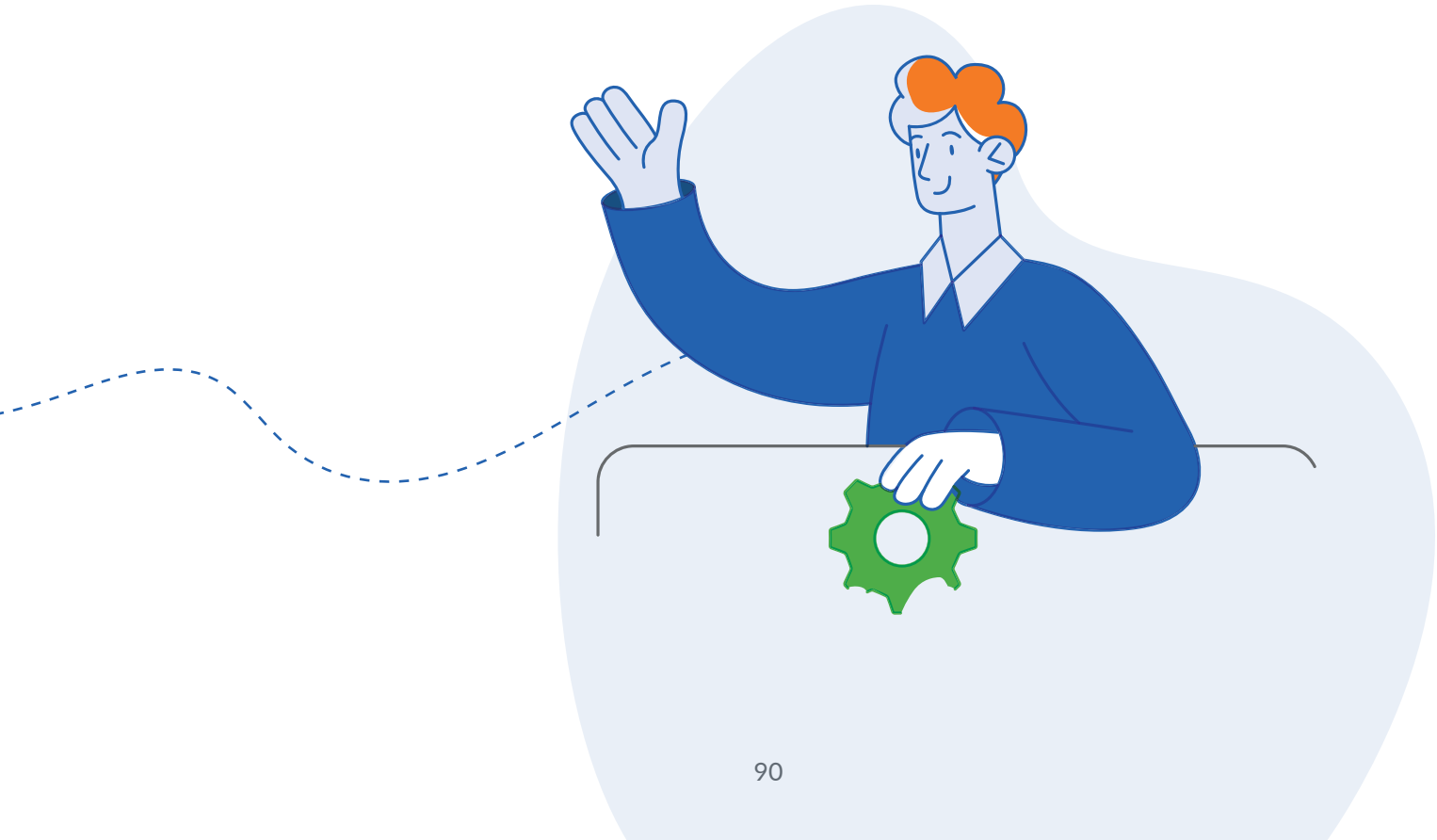


Ilustración 83. Opciones de descarga de Power Bi.

De este modo, se nos descargará un fichero ejecutable, y los siguientes pasos consistirán en aceptar todos los cuadros que nos saldrán.

Una vez finalizada la instalación, podremos abrir la aplicación y comenzar a explorar la interfaz.



PRIMEROS PASOS CON LA INTERFAZ

10

10. Primeros pasos con la interfaz

Lo primero que veremos cuando abramos Power BI será la siguiente imagen, y nos encontraremos con diferentes espacios:

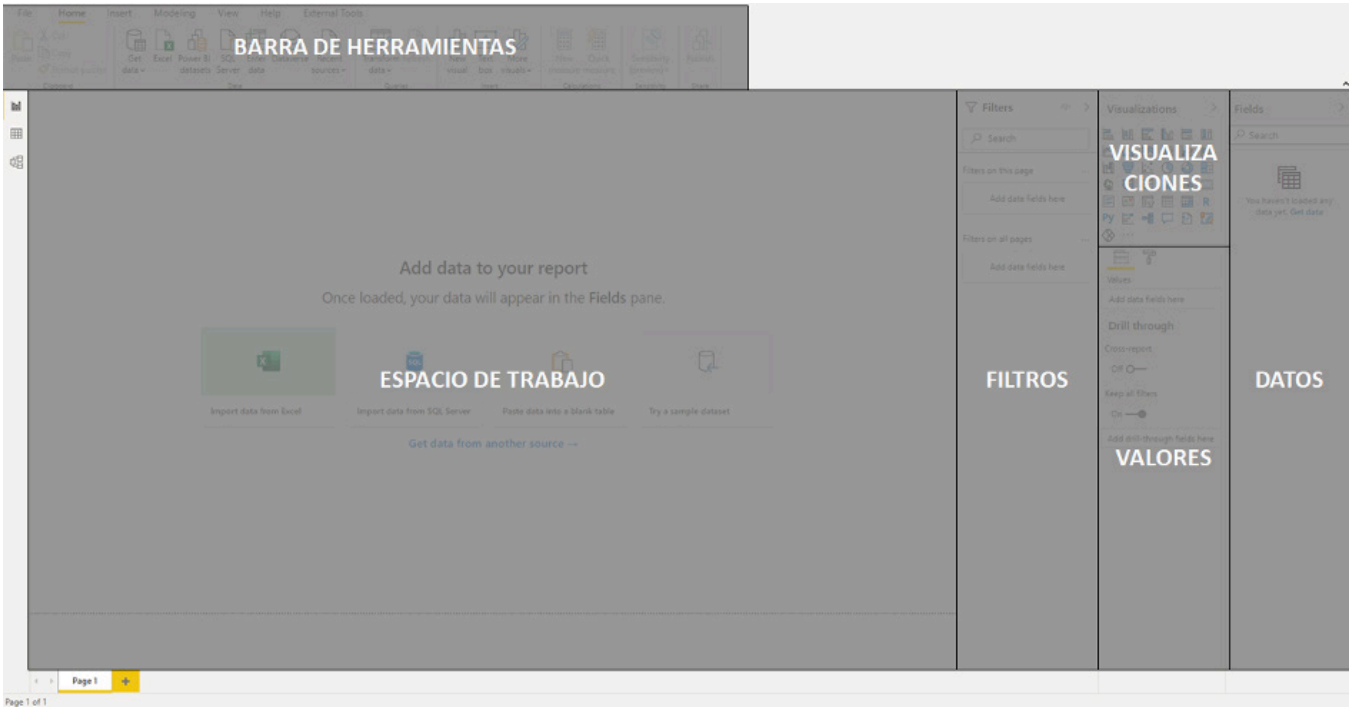


Ilustración 84. Espacios del programa Power Bi.

Barra de herramientas

Se trata de un conjunto de opciones que nos permitirá insertar datos, realizar transformaciones y administrar los diferentes conjuntos de datos y visualizaciones disponibles en nuestro informe.

Datos

En este espacio se encontrarán los diferentes conjuntos que hayamos introducido junto con las columnas que compongan dichos conjuntos de datos.

Visualizaciones

En este apartado se encuentran las diferentes opciones de representación y visualización de datos que ofrece Power BI. Adicionalmente, se pueden importar nuevas visualizaciones desarrolladas por la comunidad.

Valores

Para cada visualización se podrán in troducir una serie de valores diferentes desde el apartado de Datos. En estos valores se introducirán las columnas que queramos representar.

Filtros

En este apartado se pueden incluir diferentes filtros que afectarán a todo nuestro reporte, a una página del mismo o a una sola visualización. Los filtros aplicables afectarán a las variables introducidas en las visualizaciones, y pueden consistir en reglas sencillas o en reglas basadas en otras variables.

Adicionalmente, en la barra lateral izquierda se observan dos iconos que nos permiten acceder a otros espacios, que veremos en el siguiente apartado cuando veamos un caso de uso de la aplicación:

Espacio de datos

En este espacio se pueden explorar los datos de los conjuntos de manera separada. Cada conjunto de datos se podrá ver como una tabla.

Espacio de modelos

Este espacio está dedicado a la modelación de las relaciones entre datos. Es decir, una vez que hayamos introducido diferentes conjuntos de datos, podremos establecer cuáles son las relaciones existentes entre las variables de dos o más conjuntos de datos para realizar operaciones de manera cruzada.

ESTRUCTURA INTERNA CON LA PLATAFORMA

11. Estructura interna de la plataforma

Para poder ver la estructura interna de la Plataforma, es decir, para ver cómo son las fuentes de datos necesarias para generar las diferentes visualizaciones, nos tenemos que dirigir al apartado “Transformar datos”, ubicado en la barra de herramientas de la parte superior de la interfaz de Power BI.

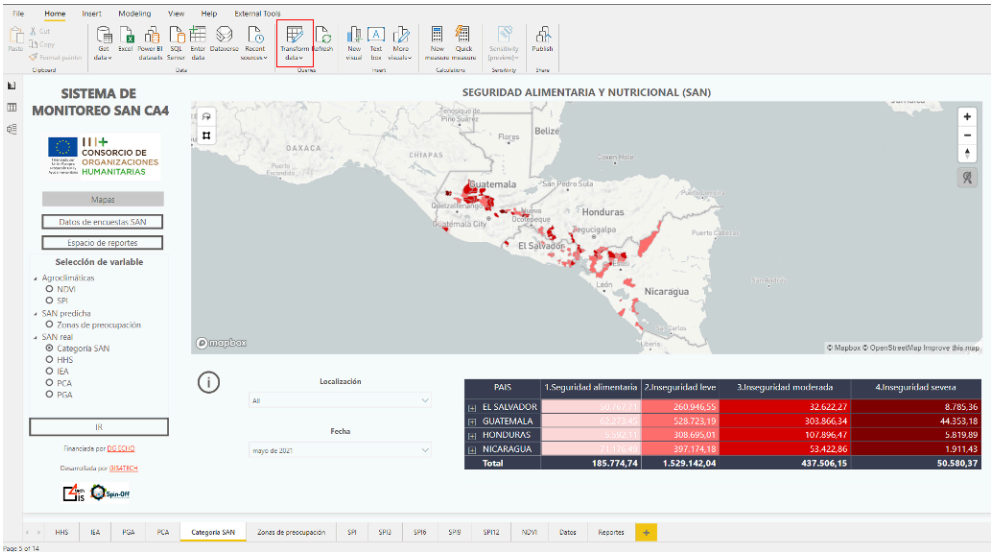


Ilustración 85. Mapa de Seguridad alimentaria y nutricional en Power Bi

En dicho apartado veremos 5 conjuntos de datos, que además podemos subdividir en información principal y secundaria:

Información primaria:

Datos encuestas

Se trata de la información recopilada por los técnicos de campo en las diferentes campañas de tomas de datos realizadas en el marco del Proyecto. Contienen todos los datos resultantes del levantamiento de información, y adicionalmente se incluye **una columna que identifica de manera única al municipio.**

Predicciones

Este conjunto de datos contiene las predicciones realizadas.

Datos agro

En este conjunto se incluyen los índices agroclimáticos correspondientes a los cultivos de grano básico y de café de cada municipio.

Información secundaria:

Población

Se trata de un conjunto en el que se incluye la población de cada uno de lo municipios objeto de estudio.

Reportes

Este conjunto de datos incorpora el nombre y los enlaces de los reportes que se mostrarán en el espacio correspondiente de la Plataforma.

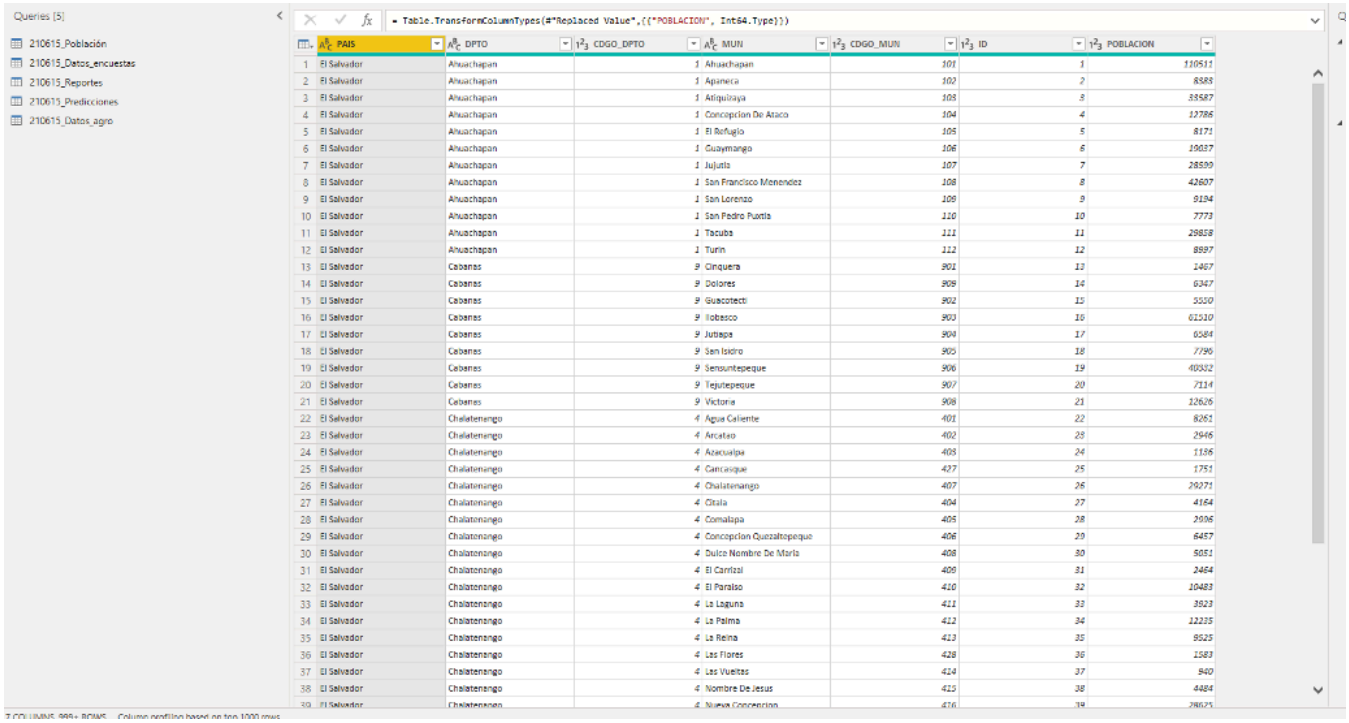
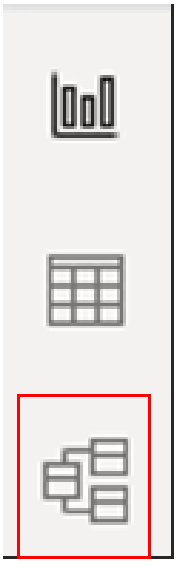


Ilustración 86. Transformación del tipo de datos de una columna.

Otro aspecto importante de la estructura interna de la Plataforma es la relación existente entre los propios conjuntos de datos. Para poder ver este aspecto, nos tenemos que dirigir al apartado “Modelo”, ubicado en la parte lateral izquierda de la interfaz de Power BI.



En este apartado podremos ver los cinco conjuntos de datos fundamentales que alimentan la Plataforma, así como las relaciones existentes entre ellos. Como hemos ido viendo a lo largo de la capacitación, todos los conjuntos de datos (excepto el de Reportes), viene acompañado de un identificador único para los municipios. La razón de que exista este identificador es que nos va a permitir establecer relaciones y uniones entre los diferentes conjuntos.

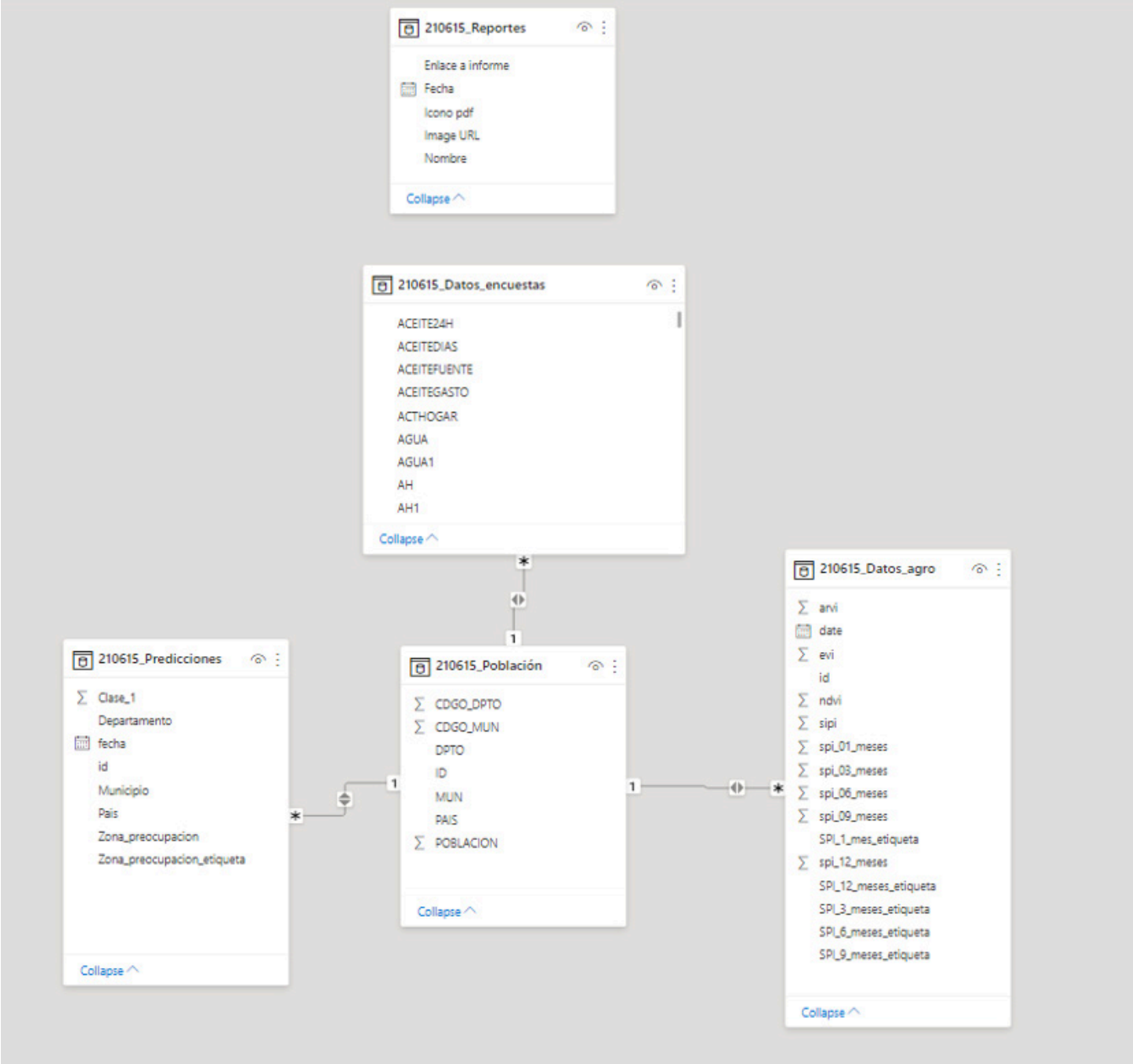


Ilustración 87. Relación entre tablas.

Se puede observar que los conjuntos de encuestas, predicciones y de datos agroclimáticos están vinculados a la tabla Población, ya que contiene valores únicos que identifican a los municipios y aportan información sobre el número de habitantes existente en cada uno de ellos.

Para comprobar la naturaleza de dichas relaciones, podemos hacer doble clic en cualquiera de los vínculos, y veremos cómo para cada conjunto aparece una columna destacada. Estas columnas destacadas son las que le están indicando al programa cómo se relacionan las tablas entre sí, por lo que, como se puede comprobar, el criterio de unión o de homogeneización de la información es el identificador único de los municipios o ID.

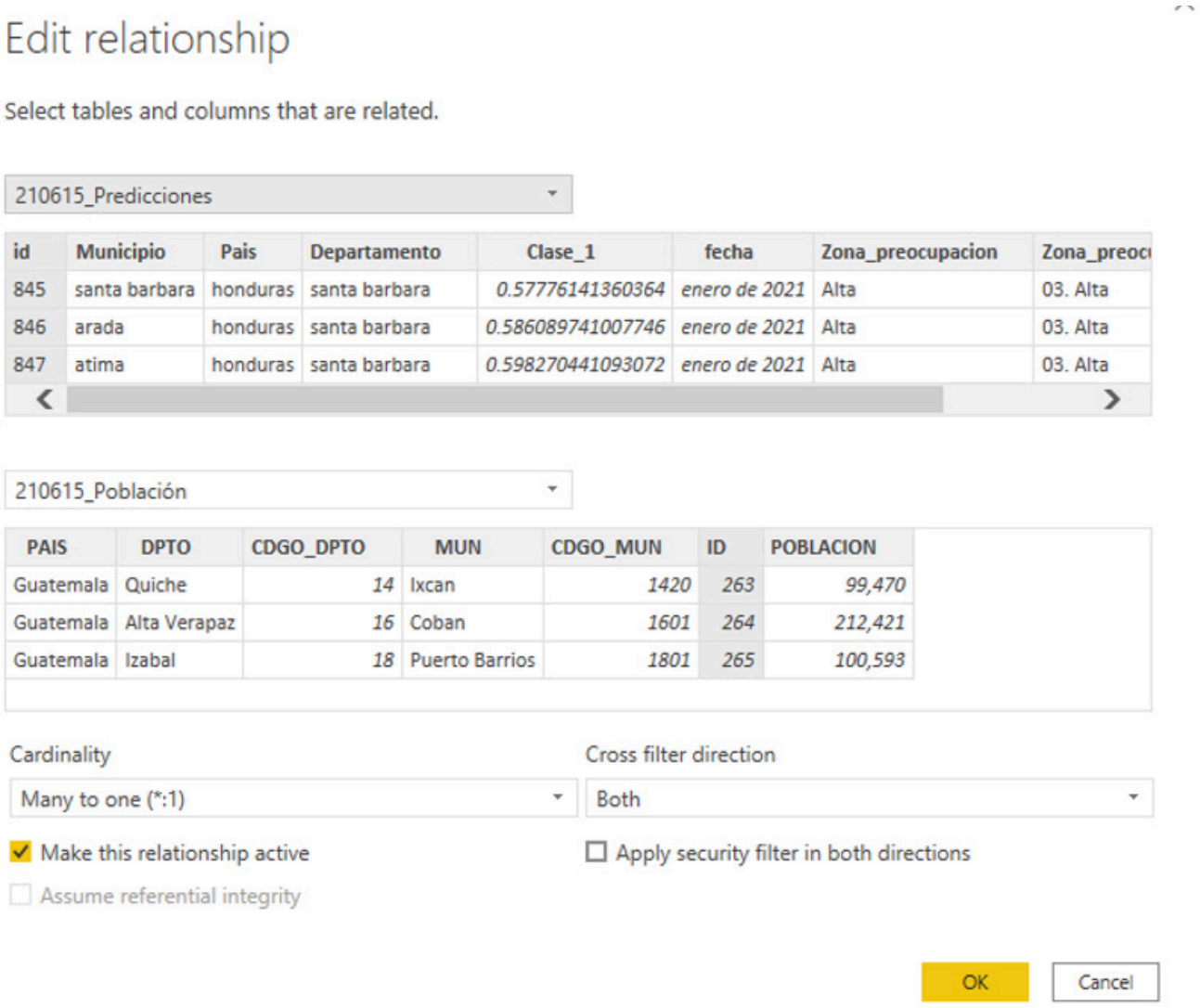


Ilustración 88. Editar relación entre tablas.

Por su parte, el conjunto de datos de Reportes no requiere ningún tipo de relación con el resto de las tablas, puesto que es una fuente auxiliar que alimentará únicamente el espacio de reportes.
Una vez vista la estructura interna de los datos, cabe destacar un par de aspectos de especial importancia:

- La estructura de los datos de origen no debe cambiar. En el caso de que los archivos de los que se alimenta la Plataforma sufran cualquier tipo de modificación (adición o sustracción de alguna columna, cambio de nombre o cambio del tipo de dato contenido), toda la información representada se verá perjudicada ya que el programa no será capaz de leer los datos correctamente.

- No se debe incluir ningún tipo de información personal en la Plataforma en relación con encuestadores o encuestados, ya que supondría una vulneración de las normativas de protección de datos.

RELACIÓN ENTRE LAS VISUALIZACIONES Y LOS CONJUNTOS DE DATOS

1.2

12. Relación entre las visualizaciones y los conjuntos de datos

A continuación, se muestra una tabla que incluye, para cada espacio, grupo y tipo de variable representada, cuáles son las fuentes de información utilizadas.

Espacio	Grupo	Pestaña	Descripción	Fuentes de datos
Mapas	SAN Real	HHS	Incidencia del hambre en el hogar. Muestra un mapa con la categoría resultante del índice HHS y la evolución temporal de la población potencialmente incluida en cada categoría.	•Datos_encuestas •Población
		IEA	Índice de estrategias de afrontamiento. Ídem de lo anterior	•Datos_encuestas •Población
		PGA	Porcentaje de gasto en alimentos. Ídem de lo anterior	•Datos_encuestas •Población
		PCA	Puntaje de consumo de alimentos. Ídem de lo anterior	•Datos_encuestas •Población
		Categoría SAN	Categoría SAN a la que pertenece un municipio	•Datos_encuestas •Población
	SAN predicha	Zonas de preo-cupación	Muestra la clasificación en función de la zona de preocupación predicha (Muy baja / Baja / Alta / Muy alta), así como la población potencialmente perteneciente a cada categoría tanto a nivel temporal como a nivel puntual	•Predicciones •Población
	Agro-climáticas	SPI	Índice estandarizado de precipitación. Muestra la anomalía de precipitación en relación con los cultivos para un periodo de meses determinado	•Datos_agro •Población
		NDVI	Índice de vegetación de diferencia normalizada. Muestra el vigor de la vegetación a lo largo del tiempo, referido a los cultivos de grano básico y de café.	•Datos_agro •Población
Datos de encuestas SAN			Resultados de las encuestas tomadas en campo, filtrables por fecha y localización	•Datos_encuestas
Reportes			Reportes relacionados con la información de la Plataforma, incluyendo el nombre, la fecha de inclusión del reporte y el enlace para su descarga	•reportes

ACTUALIZACIÓN DE DATOS DE LA PLATAFORMA

1.3

13. Actualización de datos de la plataforma

Lo primero que hay que tener en consideración a la hora de actualizar la información representada en la Plataforma es que las relaciones entre los conjuntos de datos y su inserción en las respectivas visualizaciones ya está preconfigurada, de manera que los técnicos encargados de la actualización no deberán preocuparse por realizar ninguna modificación en este sentido.

Los únicos aspectos que se deben tener en cuenta para una correcta actualización de datos en la Plataforma son los siguientes:

- Respetar la estructura de datos, incluyendo posición y nombre de las columnas.
- Mantener el formato de los datos, no incluyendo cadenas de texto donde se deban disponer números.
- Siempre que se introduzcan nuevos datos agroclimáticos, de encuestas o de predicciones, deben estar acompañados de la fecha a la que hacen referencia.

A modo de ejemplo, veremos cómo se deben actualizar los diferentes conjuntos de datos para poder incluirlos en la Plataforma.

Supongamos que tenemos los datos de la plataforma actualizados al mes de abril de 2021, y que ya disponemos de datos agroclimáticos y de predicciones a mayo de 2021. Además, contamos con un nuevo levantamiento de datos, por lo que podremos actualizar los tres conjuntos de información.

Para una mejor comprensión, estableceremos dos ficheros de datos para almacenar las fuentes:

- **01. Info**

  Esta carpeta incluirá las carpetas desde las cuales se alimentará la Plataforma.

- **02. Actualización**

  Aquí se incorporarán los ficheros a medida que dispongamos de nueva información.

A continuación, se muestra la información relativa a los datos agroclimáticos, para los cuales tenemos datos hasta abril de 2021.

A	B	C	D	E	F	G	H	I	J	K
id	date	spi_12_meses	sipi	spi_09_meses	evi	spi_06_meses	spi_01_meses	spi_03_meses	ndvi	arvi
1041	01/04/2021	0.51697975	1.1509352	0.81743383	0.37421045	2.550773	-0.17210057	0.010578878	0.62901336	0.42266303
1042	01/04/2021	0.92934704	1.2133085	1.1911128	0.3264582	3.0370746	-0.039840903	0.13983038	0.5565188	0.33317435
1043	01/04/2021	1.6757839	1.2955774	1.4542996	0.3073986	3.3073888	1.3129948	1.445214	0.48610187	0.26193622
1044	01/04/2021	1.5893048	1.3598858	1.4491165	0.28762087	3.508429	1.0601757	1.3346525	0.4234502	0.19632077
1045	01/04/2021	1.318134	1.233511	1.1706926	0.3134307	2.8273098	0.74665165	0.9705736	0.5088923	0.32420906
1046	01/04/2021	1.2710298	1.1622661	1.1702198	0.3864635	2.485948	2.619533	2.7549436	0.5995931	0.38917324

Como se puede comprobar, la última información disponible es la relativa al mes de abril de 2021, por lo que incorporaremos la información que hayamos calculado mediante las técnicas vistas en esta capacitación. Para ello, nos dirigiremos a la carpeta “02. Actualización”, donde tendremos una hoja de Excel con los datos agroclimáticos del mes de mayo de 2021.

A	B	C	D	E	F	G	H	I	J	K
id	date	spi_12_meses	sipi	spi_09_meses	evi	spi_06_meses	spi_01_meses	spi_03_meses	ndvi	arvi
1	01/05/2021	-0.34083614	1.2735066	0.03267419	0.30865848	-0.5006281	-0.89222884	-0.57033956	0.5026799	0.2895212
2	01/05/2021	-0.18194036	1.5838214	-0.000446528	0.1872837	-0.6195571	-0.83252877	-0.68305767	0.25089023	0.14272279
3	01/05/2021	-0.048453297	1.2548177	0.45581767	0.31663126	-0.31413683	-0.7695593	-0.45275357	0.5339828	0.31118846
4	01/05/2021	-0.21697715	1.076929	-0.10004171	0.3262576	-0.588691	-0.82694864	-0.62894416	0.4668236	0.33295402

Para actualizar la información agroclimática ubicada en la Plataforma, copiamos los datos de este segundo fichero y los incorporamos en el mismo orden al primero, teniendo cuidado de que las columnas coincidan perfectamente. Una vez que hemos copiado y pegado la información, debemos asegurarnos de que la ruta al archivo desde la Plataforma se está leyendo correctamente. Desde el archivo en formato .pbix que contiene los datos de la Plataforma, desplegamos el apartado “Archivo” de la barra de herramientas superior y nos dirigimos a “Configuración de fuentes de datos”.

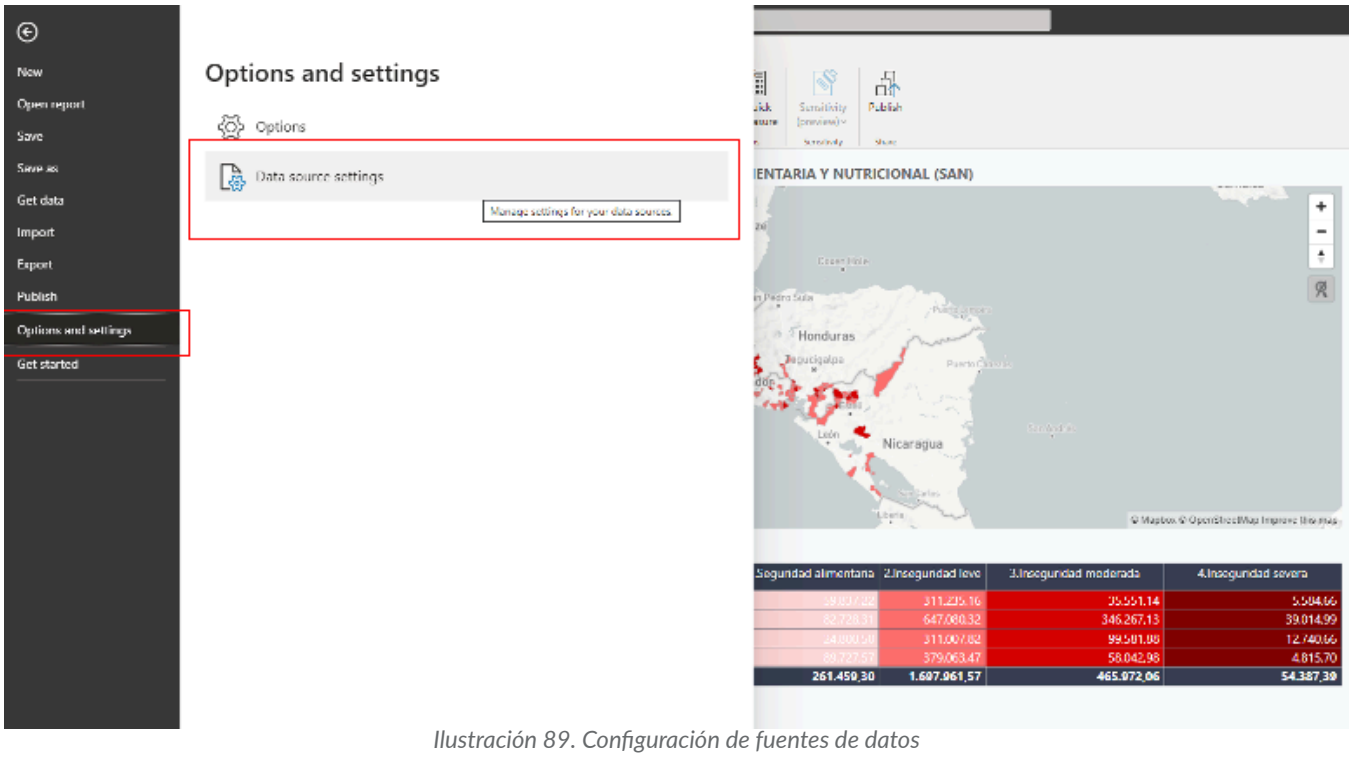


Ilustración 89. Configuración de fuentes de datos

A continuación, aparecerá una ventana en la que nos saldrán los ficheros desde los cuales se está alimentando la Plataforma. Dado que estamos actualizando los datos agroclimáticos, seleccionaremos dicha fuente y haremos clic en el botón “Cambiar fuente”.

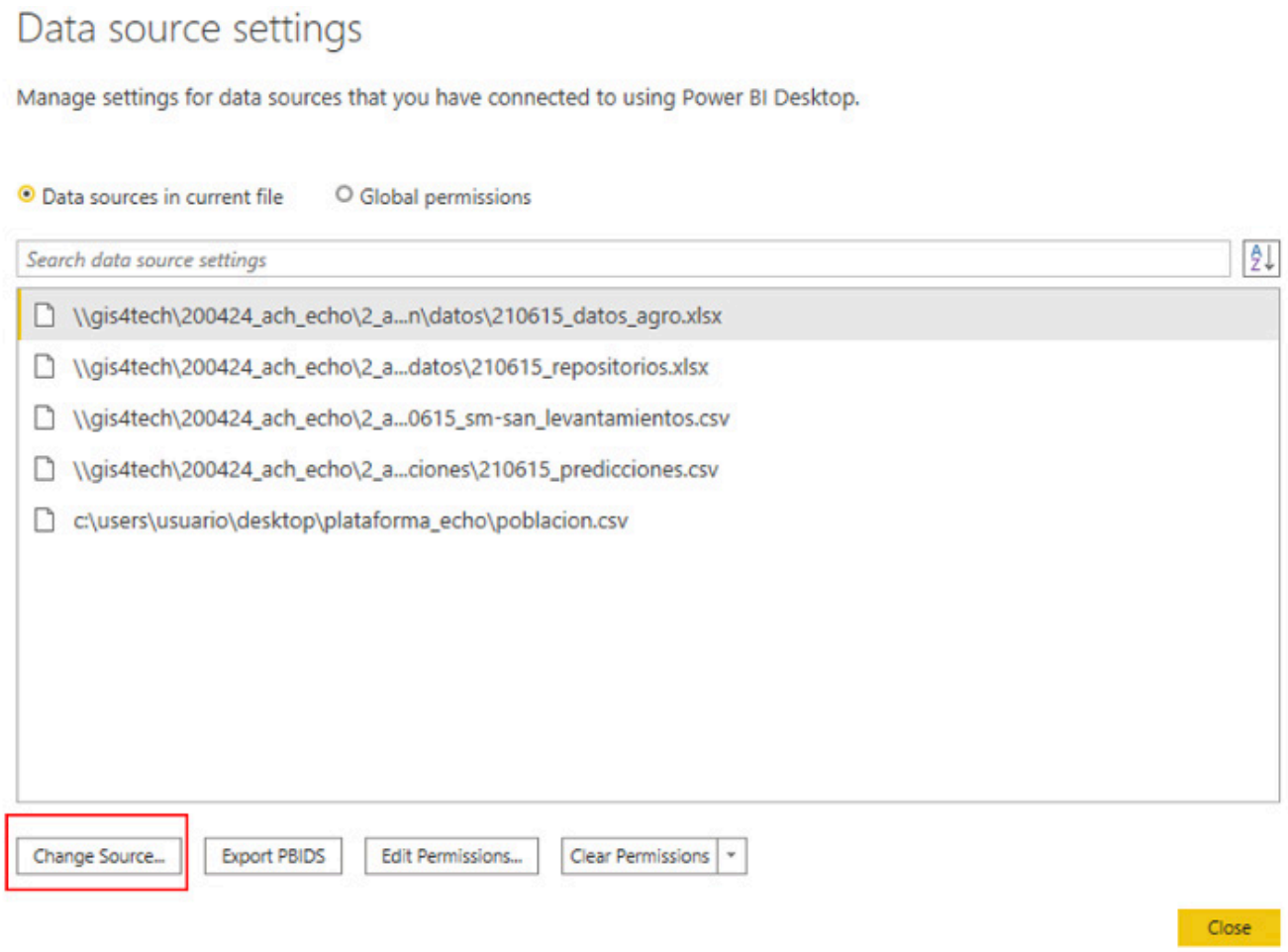


Ilustración 90. Cambiar fuente de datos.

De esta manera, nos aparecerá una ventana de exploración en la que buscaremos el fichero con los datos agroclimáticos actualizados, y lo seleccionaremos.

Este proceso se hace de manera análoga para todos los conjuntos de datos, por lo que procedemos al último paso para la actualización, que consiste en pulsar el botón “Actualizar”, ubicado en la barra de herramientas superior de la interfaz, concluyendo así la actualización de datos.

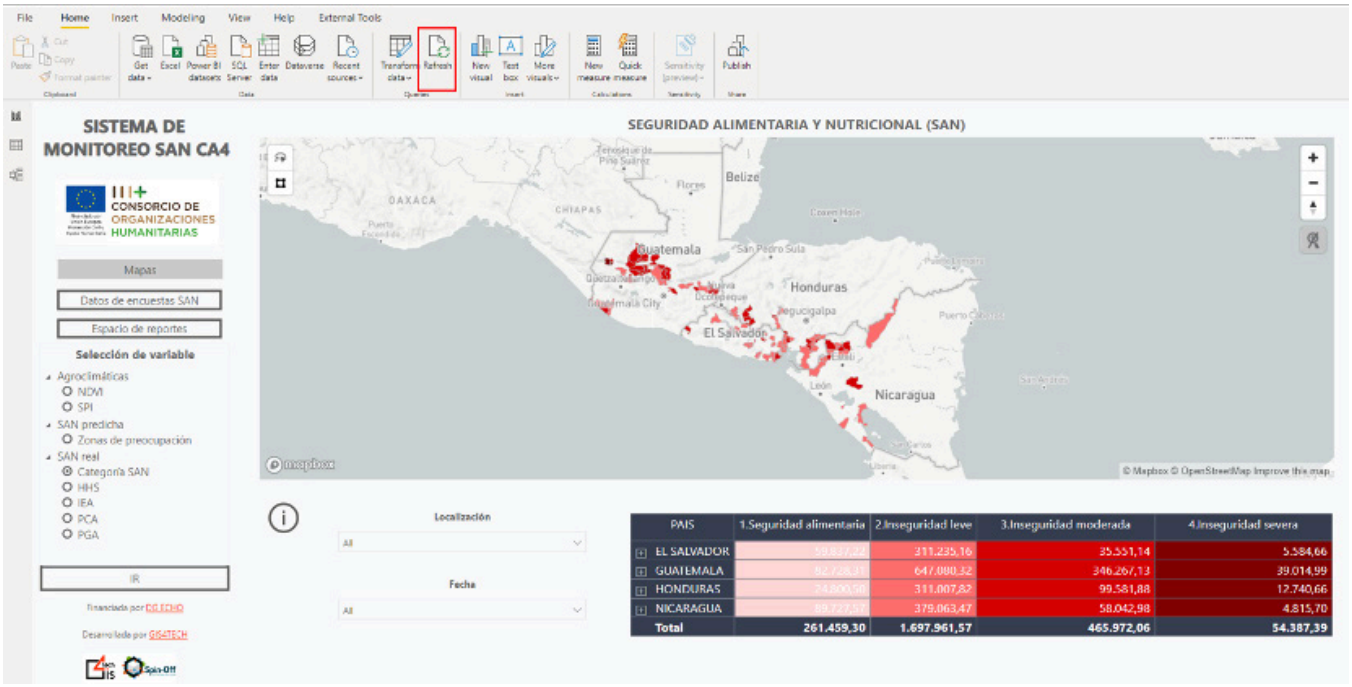


Ilustración 91. Actualizar datos.

PUBLICACIÓN DE LA PLATAFORMA

14

14. Publicación de la plataforma

Una vez que ya hemos actualizado los datos contenidos en la Plataforma, debemos proceder a su publicación en la web. Para ello, es imprescindible tener, como mínimo, una licencia de Power BI Pro. En la barra de herramientas superior de la interfaz se encuentra el botón “Publicar”, donde procedemos a hacer clic.

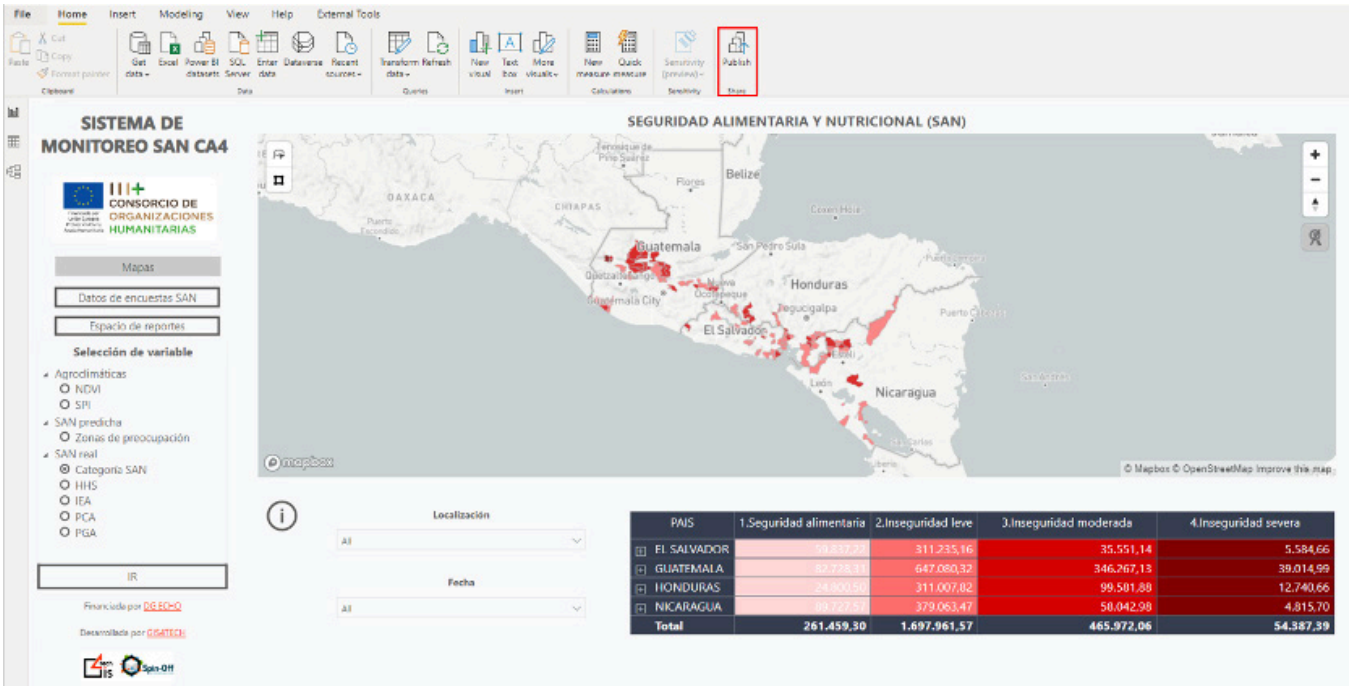


Ilustración92. Publicar panel.

A continuación, aparecerá una ventana en la que se mostrarán los espacios de trabajo asociados a la cuenta de Power BI en caso de haber creado alguno. Por defecto, aparece uno denominado “Mi espacio de trabajo”. Lo seleccionamos y pulsamos el botón “Seleccionar”.

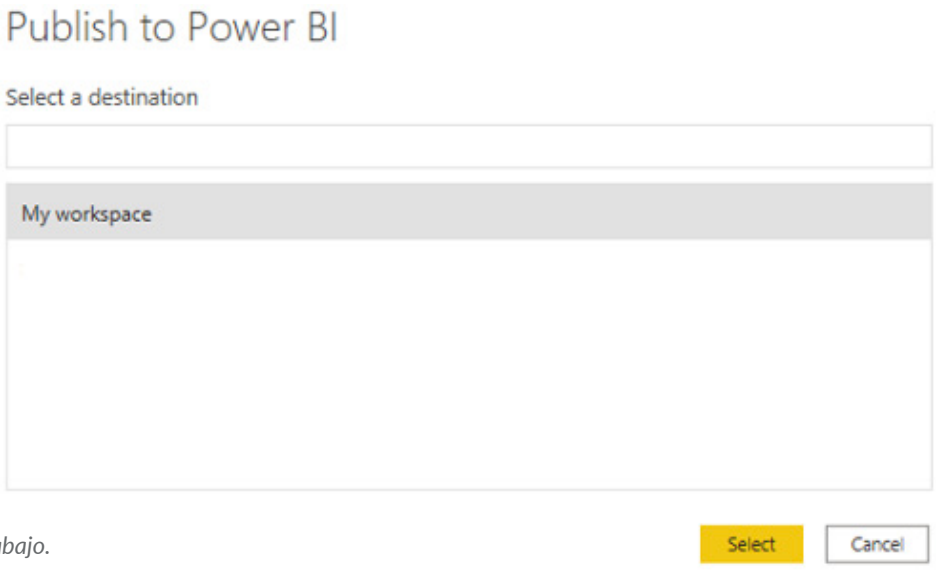


Ilustración 93. Selección del espacio de trabajo.

Tras un proceso de carga, nos aparecerá un enlace en el que debemos hacer clic, que nos llevará a nuestro navegador predeterminado.

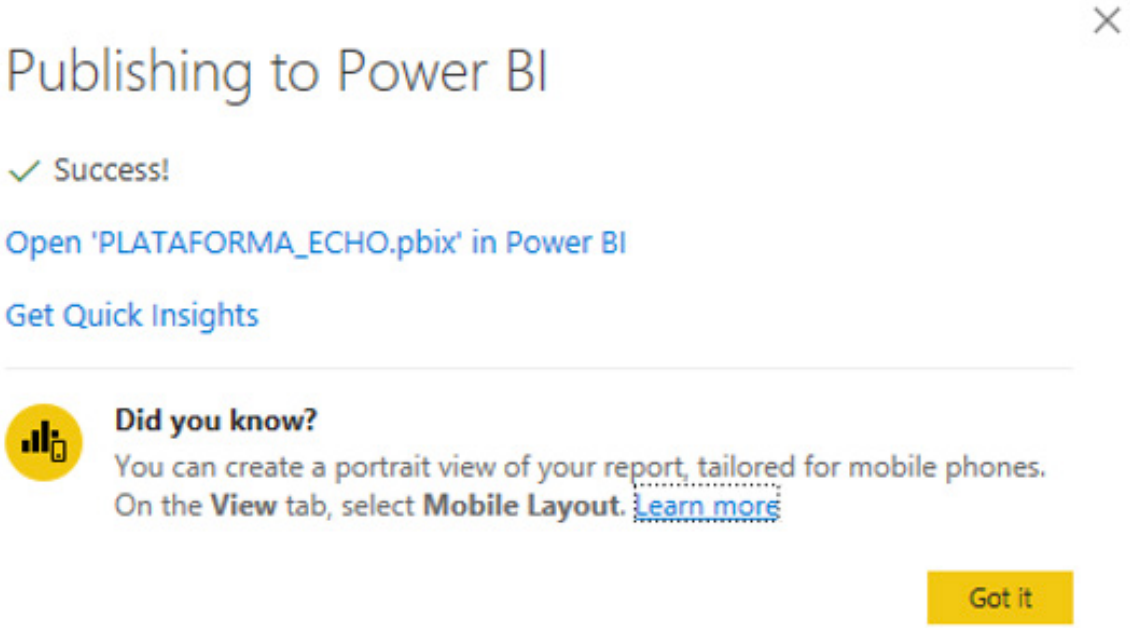


Ilustración 94. Publicar panel.

En la ventana abierta en el navegador, podremos observar el cuadro de mando que hemos creado en local y que hemos publicado. El siguiente paso consiste en dirigirse a la pestaña “Archivo” > ”Insertar informe” > “Publicar en la web (público)”, lo cual nos generará un enlace que podremos copiar y pegar en la ventana del navegador.

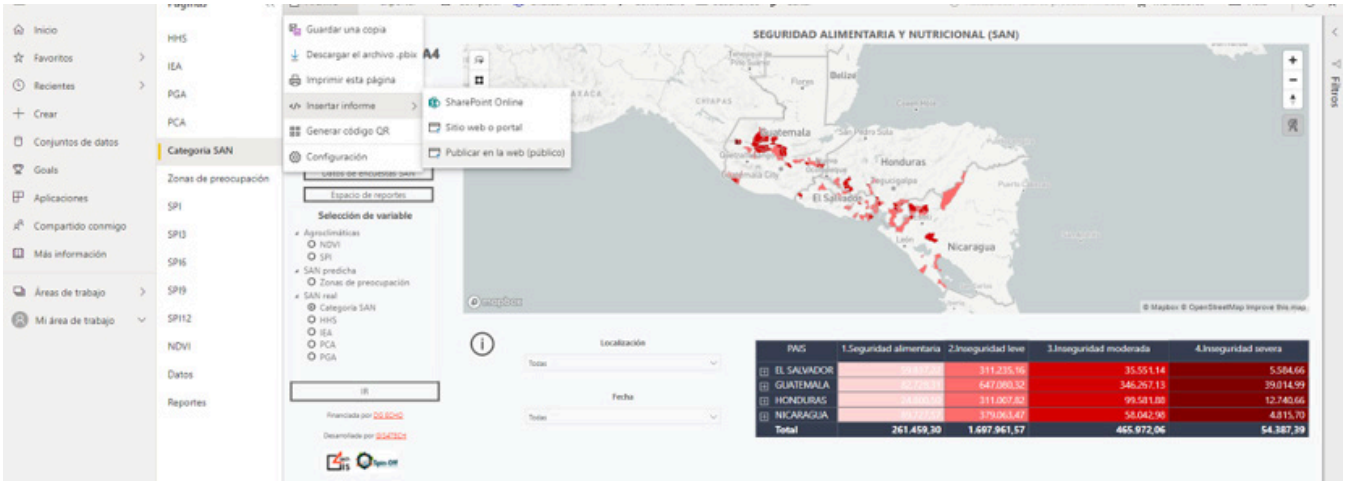


Ilustración 95. Publicar panel en la web.

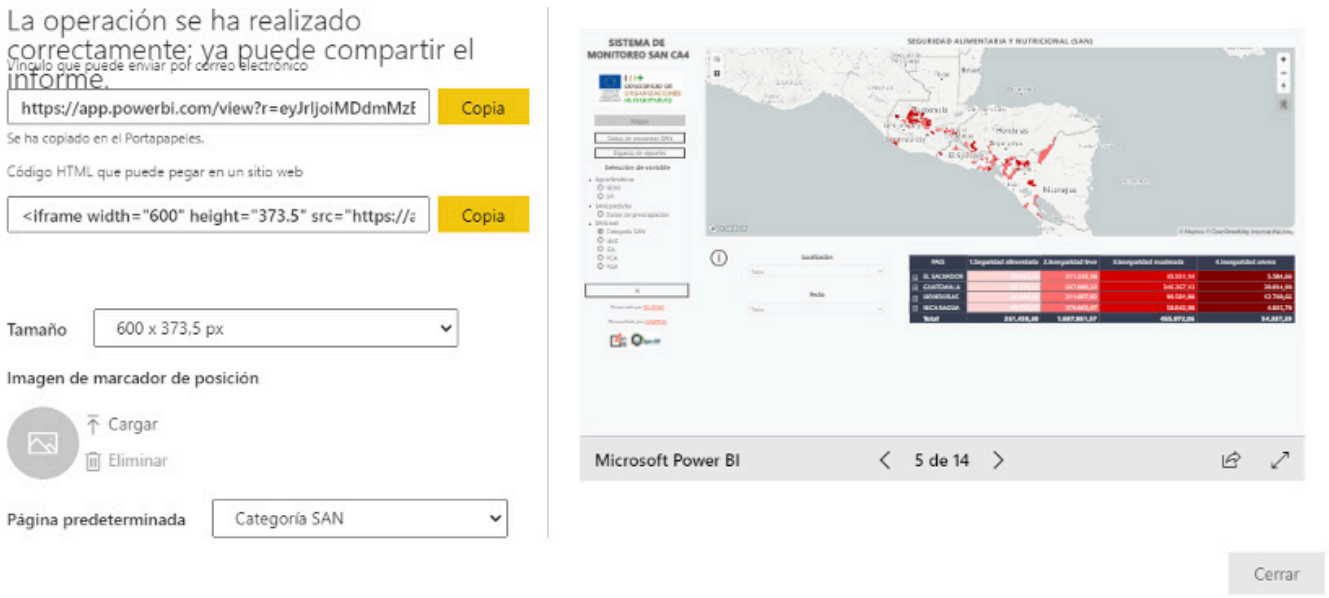
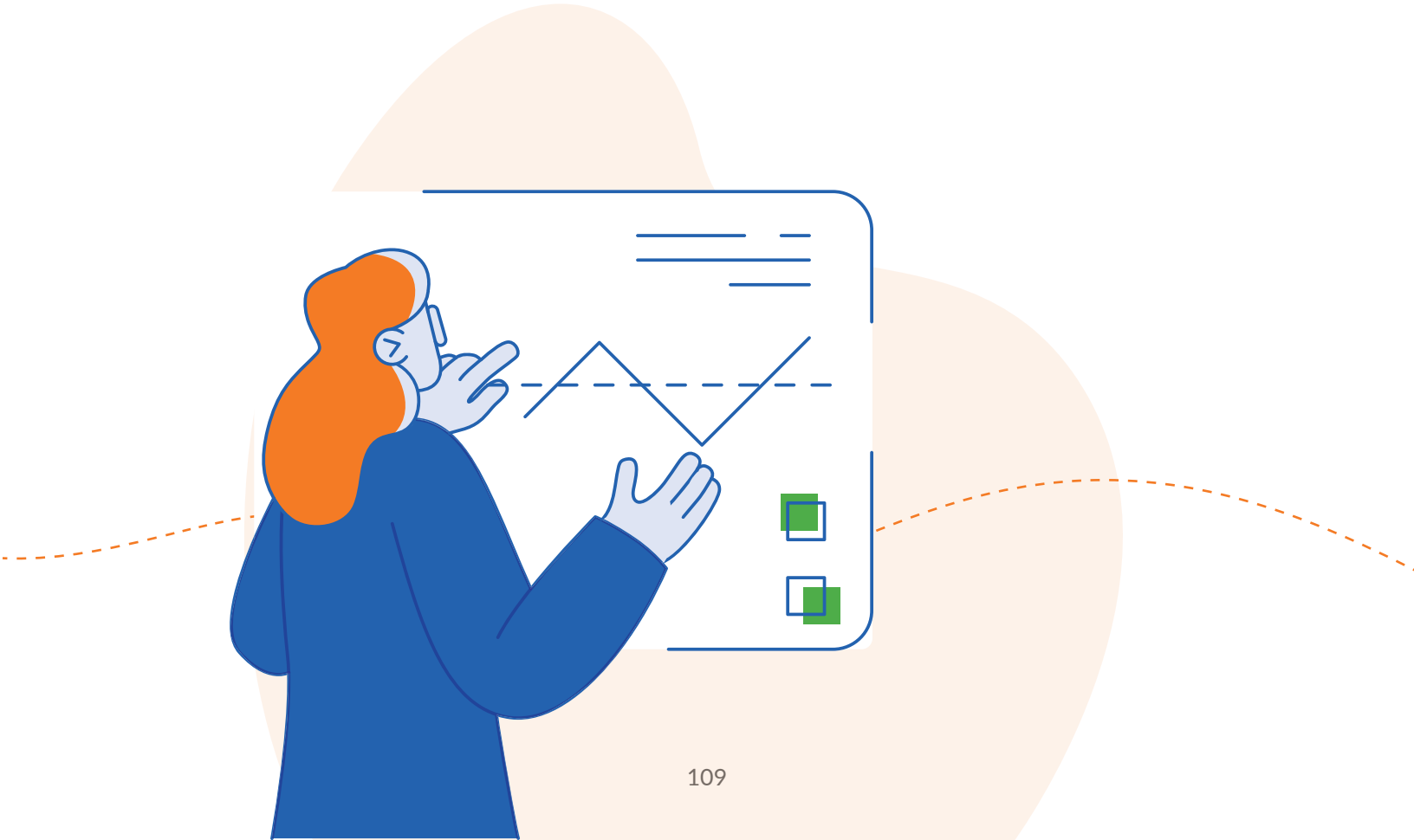


Ilustración 96. Generación del enlace del informe.

Una vez que tengamos este enlace, podremos acceder a la Plataforma desde cualquier ubicación o navegador, y podremos compartir los resultados del monitoreo de la SAN y las zonas de preocupación con cualquier usuario público.