MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA CARACTERIZACIÓN MULTIDIMENSIONAL DE LA SEGURIDAD ALIMENTARIA EN EL DEPARTAMENTO DEL CAUCA

Tesis de Trabajo de Grado
Modalidad: Trabajo de Investigación

David Santiago Restrepo Rodriguez
100614021014
Luis Enrique Pérez Maldonado
100615020647

Director: PhD. Diego Mauricio López Gutiérrez
Co-Director. PhD. Rubiel Vargas Cañas.

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones Departamento de Telemática
Línea de Investigación en e-Salud

Popayán, agosto de 2021

LISTA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Capítulo 1. Introducción

1.1. Planteamiento Del Problema

En muchos lugares del mundo, especialmente en los países de bajos y medianos ingresos (LMIC), la desnutrición y la inseguridad alimentaria han sido foco de preocupación. La Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), define la desnutrición como un estado patológico causado, ya sea por una dieta deficiente (en uno o varios nutrientes esenciales) o por una mala asimilación de los alimentos [1]. Así mismo, esta organización, define la inseguridad alimentaria como la carencia de acceso regular a suficientes alimentos inocuos y nutritivos para un crecimiento y desarrollo normales, y para llevar una vida activa y saludable.

Países con escasos recursos como Colombia enfrentan hoy día desafíos relacionados con la malnutrición (desnutrición aguda, retraso en talla y sobrepeso especialmente en escolares), situaciones que afectan el crecimiento económico y social del país y limita las posibilidades de desarrollo de los individuos [2]. Según datos de la Encuesta Nacional de Salud y Nutrición (ENSIN) [3], la desnutrición crónica en Colombia alcanzó al 10,8% de la población en el año 2015. Así mismo, la prevalencia de desnutrición aguda en menores de 5 años, se incrementó entre el año 2010 y 2015 de 0,9% a 2,3%. También es importante tener en cuenta que para el año 2015, casi 1 de cada 3 niñas y niños indígenas (29.6%) sufría desnutrición crónica, proporción tres veces mayor que el promedio nacional.

En el departamento del Cauca los datos a nivel de nutrición son aún más preocupantes. El Cauca ocupa el tercer lugar en desnutrición en menores de 5 años y así mismo en obesidad en adultos y adolescentes. En el año 2010 los indicadores de desnutrición aguda para menores de 5 años en el departamento se encontraba en 4.9%, mientras que esta prevalencia a nivel nacional se encontraba en 3.4%. Así mismo y de manera alarmante, la desnutrición crónica en menores de 5 años se encuentra casi 10 puntos por encima del nivel nacional con un 23.1% para el Cauca y un 13.2% para el nacional. Uno de los factores que incrementan las altas tasas de desnutrición crónica en el departamento, es que el Cauca es el departamento que alberga mayor porcentaje de población indigena del país con 190.069 personas (cerca del 20% del total departamental) [4].

Ante lo anterior, surge la necesidad de entender las posibles variables que están afectando los problemas de malnutrición en el departamento del Cauca. Para esto,

se puede tomar como referencia un estudio sobre seguridad alimentaria de la FAO, [30] en la cual se describen las 4 dimensiones fundamentales para caracterizar la seguridad alimentaria: disponibilidad física de los alimentos, acceso económico y físico a los alimentos, la utilización de los alimentos y la estabilidad en el tiempo de las tres dimensiones anteriores. Una herramienta para cuantificar estas dimensiones es el índice Global Food Security Index (GFSI)[5] el cuál, como se explica en el estado del arte, es el más completo y el que mejor se acopla a las 4 dimensiones de la FAO. A pesar de la pertinencia de este índice para caracterizar la seguridad alimentaria de una región, país o comunidad; el análisis del estado del arte de este proyecto revela que la mayoría de las investigaciones se han enfocado en caracterizar solo la primera dimensión del modelo de seguridad alimentaria propuesto por la FAO (disponibilidad física de los alimentos), dejando de lado las otras dimensiones. Esto puede darse debido a la complejidad y costos para acceder a los datos para realizar un análisis completo en todas las dimensiones.

A su vez el uso cálculo de un índice como el GFSI implica dificultades en cuanto al costo y tiempo que conlleva la recolección de los datos, razón por la cual es un índice que solo se calcula para países. En el caso del departamento del Cauca, no existe ningún dataset multidimensional con el cual se pueda realizar el cálculo del GFSI, por lo cual se hace necesario la construcción de uno.

Además el proceso de calcular un GFSI se hace un proceso muy manual, y debido a la cantidad de datos que este requiere es un proceso engorroso, por esta razón el uso de modelos de aprendizaje automático puede hacer de esta tarea un proceso más simple, fácil de automatizar y rápido.

1.2. Pregunta de investigación e hipótesis

Teniendo en cuenta las problemáticas antes mencionadas, se plantea la siguiente pregunta de investigación: ¿Cuál es la situación de seguridad alimentaria del departamento del Cauca teniendo como referencia las cuatro dimensiones fundamentales propuestas por la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) y el Global Food Security index (GFSI)?

Para responder esta pregunta se plantea la hipótesis que es posible realizar una caracterización integral de la situación de la seguridad alimentaria del departamento del Cauca que aborde las cuatro dimensiones fundamentales propuestas por la FAO mediante el cálculo del GFSI, realizando un análisis multimodal de diversas fuentes de datos abiertas como las disponibles en censos sociodemográficos, agrícolas, nutricionales, de salud pública, y datos de otras

modalidades como imágenes satelitales. Este análisis multimodal y multidimensional estaría apoyado en diferentes técnicas de aprendizaje automático.

1.3. Motivación

Tanto a nivel nacional como internacional, el interés por la seguridad alimentaria y nutrición cobra mucha importancia debido a que está directamente relacionado con el progreso del país y de la sociedad. En el contexto nacional, Colombia es un país que cuenta con diferentes características dependiendo de la región o zona que se analice. Estas características afectan a la situación de seguridad alimentaria de forma diferente en cada región o zona del país. Debido a lo anterior, se hace necesario identificar la situación de seguridad alimentaria en cada región del país, con el fin de obtener información para apoyar en la toma de decisiones de los dirigentes de cada región en cuanto a la nutrición y seguridad alimentaria.

El presente proyecto de investigación es relevante tanto para la comunidad internacional como la nacional, ya que apoya la toma de decisiones de los dirigentes de los países o regiones a partir de un índice que permite identificar la situación de nutrición de seguridad alimentaria del país o región en cuestión.

1.4. Objetivos

1.4.1. Objetivo General

Determinar la situación de la seguridad alimentaria del Departamento del Cauca mediante la propuesta de un modelo de aprendizaje automático basado en la integración de imágenes satelitales, datos de nutrición, sociodemográficos y del censo nacional agropecuario.

1.4.2. Objetivos específicos

1. Caracterizar la situación del Departamento del Cauca con relación a aspectos nutricionales, agropecuarios, sociodemográficos y socioeconómicos haciendo uso de técnicas de análisis de datos.
2. Construir un dataset que integre las principales variables que influyen en la seguridad alimentaria con fuentes provenientes de repositorios abiertos.
3. Desarrollar un modelo de aprendizaje automático que calcule el GFSI de municipios en el departamento del cauca a partir del dataset creado y los datos disponibles en el GFSI.

1.5. Metodología

La metodología usada para guiar las actividades necesarias para alcanzar los tres objetivos específicos es la metodología CRISP-DM[1]. Esta es una metodología cíclica, por esta razón todas las fases explicadas a continuación no dejan de ser relevantes en ningún momento ni se finalizan definitivamente. En cualquier momento se puede regresar a una fase anterior. Las fases son:

Fase 1 Comprensión del negocio: En esta fase se revisa la literatura, se entiende el problema y se definen los pasos a seguir para brindar una solución. La revisión de la literatura contempló los métodos, fuentes de datos y la forma de medir la seguridad alimentaría (por ejemplo, a través de un índice de seguridad alimentaria). Las actividades en esta fase permitieron definir el objetivo de la investigación y los pasos a seguir.

Fase 2 Comprensión de los datos: En esta fase se recolectan las diferentes fuentes de datos, se analiza su contenido y su posible uso en las etapas posteriores. Como resultado de la etapa anterior para este caso específico se obtiene un índice de seguridad alimentaria que cumpla con las condiciones descritas por la FAO, y se analizan los datos identificados en la etapa anterior con el fin de poder prepararlos y generar un único dataset. Las actividades en esta fase, permiten alcanzar los resultados planteados en el objetivo específico 1.

Fase 3 Preparación de datos: Al tratarse de datos de diferentes fuentes cada tipo de dato tiene su resolución espacial y temporal, además de su estructura, por lo tanto se debe hacer una limpieza de datos y un pre-procesamiento para poder conformar el dataset final. Las actividades en esta fase, permiten alcanzar los resultados planteados en el objetivo específico 2.

Fase 4: Modelado: Se toma como base el dataset creado en la anterior fase para poder calcular el índice de seguridad alimentaria, esto se hace con un modelo a nivel de país debido a los datos existentes actualmente en el GFSI y se extrapola a nivel de municipio usando el modelo. Las actividades en esta fase, permiten alcanzar parcialmente los resultados planteados en el objetivo específico 3.

Fase 5 Evaluación: Se valida el funcionamiento del modelo creado en la fase 4 para verificar su correcto funcionamiento, esto refuerza el uso de un modelo de cálculo de GFSI a nivel de país debido a que es la única forma de validarlo ya que

[1] P. Chapman *et al.*, “CRISP-DM 1.0 Step-by-step data mining guide,” 2000

no existe un cálculo de este a nivel de departamento. Las actividades en esta fase, permiten alcanzar los resultados planteados en el objetivo específico 3.

La sexta fase (despliegue) en la cual se hace la puesta en escena del modelo en un entorno real, no se tiene en cuenta debido a que está por fuera del alcance del trabajo de grado.

1.6. Contenido de la Monografía

La monografía se redactó tomando como base tanto la metodología CRISP-DM, como los objetivos específicos del proyecto, de forma que a medida que se desarrolla la monografía se resuelve cada uno de los objetivos. Cabe aclarar que si bien la metodología CRISP-DM no es una metodología en la cual pasada una fase está ya se dá por finalizada, sino que se pueden realizar varias iteraciones o volver a una fase anterior en un punto del proyecto, aquí se explicará lo hecho en cada fase de forma general y no significa que cada fase se completará con una única iteración o el orden en el que se detalla cada actividad dentro de los capítulos sea estrictamente como se realizó el proyecto.
En el capítulo 2 se reportan los resultados de la fase de Comprensión de negocio. En el capítulo 3 se describen los resultados de la fase de Comprensión de los Datos. En el capítulo 4, se presentan los resultados de la fase de Preparación de los Datos. El capítulo 5 reporta los resultados de las fases de Modelado y Evaluación, según la metodología de CRISP-DM. Finalmente el Capítulo 6 presenta las conclusiones y trabajo futuro de este trabajo de grado.

Capítulo 2

Comprensión del Negocio

2.1. Antecedentes

Para definir los antecedentes se hizo uso de un mapeo sistemático de literatura existente (Estado del Arte). Esta búsqueda del estado del arte se dividió en una parte técnica, desde el lado del uso de datasets de múltiples fuentes y con datos multimodales como pueden ser las imágenes satelitales, y por otro lado la búsqueda de un índice de seguridad alimentaria acorde a lo descrito por la FAO.

2.1.1. Trabajos realizados en seguridad alimentaria con machine learning

Para la primera parte se hizo un mapeo sistemático de la literatura haciendo uso de Scopus y Web of Science. Con ayuda de la cadena de búsqueda TITLE-ABS-KEY ("food security" AND ("deep learning" OR "machine learning")). Se buscaron proyectos en los que se haya usado técnicas de aprendizaje automático que hagan uso al menos dos modalidades de datos, como son imágenes satelitales o fotos aéreas y datos categóricos o numéricos procedentes de censos y encuestas. Para la escogencia de los artículos se utilizaron ciertos criterios de inclusión y de exclusión, entre ellos están: artículos de 2016 en adelante con la exigencia de tener mínimo una citación (excepto para los del 2021), se tuvo en cuenta que los artículos realizaron una correcta explicación de los algoritmos o métodos utilizados y resultados, también que se haga uso de imágenes satelitales y/o fotos aéreas y que se incluyan conceptos como nutrición, agricultura y seguridad alimentaria. Entre los criterios de exclusión se descartaron artículos por fuera del rango de inclusión, que no traten la temática mencionada anteriormente, y que los métodos usados sean desactualizados o que no generen resultados esperados. Con base a lo anterior, se identificaron 23 trabajos [7-29] . Para cada uno de ellos se analizó el área temática del trabajo, el tipo de datos, los algoritmos usados. A continuación se resumen las características encontradas en los artículos seleccionados:

- El uso de datos multimodales y de múltiples fuentes como imágenes satelitales, datos meteorológicos, censos y encuestas para entrenamiento de algoritmos de aprendizaje automático, así como el uso de algoritmos

basados en redes neuronales como redes neuronales convolucionales (CNN) o redes neuronales recurrentes (RNN); son temáticas que han aumentado exponencialmente su aparición en el campo de la seguridad alimentaria entre los años 2020 y 2021 (ver tabla 1).

Año	Número de artículos.
2021	26 (Hasta Abril 18)
2020	50
2019	26
2018	5
2017	2
2016	2

Tabla 1. Artículos con RNN y CNN clasificados por años.

- La mayoría de los trabajos hacían uso de algoritmos de aprendizaje automático para predicción o clasificación del rendimiento de cultivos. Sin embargo, los datos usados eran siempre imágenes satelitales, fotos aéreas, datos meteorológicos y datos socioeconómicos (en un solo caso) (ver tabla 2) o una combinación de los anteriores tal y como se muestra en la tabla 3.

	Imágenes satelitales o imágenes aéreas.	Datos metereológicos.	Información histórica de cultivos.	Datos socioeconómicos.
Número de artículos.	22	13	6	1

Tabla 2. Datos usados en los 23 artículos seleccionados.

	Categoría					
	1	1 y 2	1 y 3	2 y 3	1, 2 y 3	1, 2 y 4
Número	11	5	1	2	3	1

de artículos.						

Tabla 3. Combinaciones de datos en los 23 artículos seleccionados.

Donde:

- Categoría 1: imágenes satelitales o imágenes aéreas.
- Categoría 2: datos meteorológicos.
- Categoría 3: información histórica de cultivos.
- Categoría 4: datos socioeconómicos.

- Los trabajos se centraron en su mayoría en un tipo de cultivo específico, en muy pocas ocasiones se usaron más de uno, exceptuando algunos casos puntuales en los que se hiciera clasificación de cultivos para un país o zona específica.

- La mayoría de imágenes para realizar el entrenamiento de los algoritmos de Machine Learning (ML) se obtienen de bases de datos de los satélites, Sentinel-1, Sentinel-2, MODIS, Worldview 3. Algunas imágenes con características adicionales que se encuentran en Google Earth Engine (GEE), e imágenes provenientes de vehículos aéreos no tripulados (UAV). Las fuentes de las imágenes y el número de artículos que las usan se pueden ver en la tabla 4.

	Sentinel-1.	Sentinel-2.	MODIS.	Google Earth Engine.	Worldview 3.	vehículo aéreo no tripulado. (UAV)	No especifica.
Número de artículos.	4	2	4	2	2	4	9

Tabla 4. Fuentes de imágenes satelitales o aéreas usadas en los 23 artículos seleccionados.

En cuanto a las dimensiones de la seguridad alimentaria propuesta por la FAO [30], la mayoría de artículos se refieren a la dimensión de disponibilidad física de los alimentos (aunque en muchos casos incluso este campo no cubre del todo, por el hecho de enfocar los trabajos a un solo cultivo). Particularmente se enfocan en la predicción del rendimiento de los cultivos.

2.1.2. Índices de seguridad alimentaria

Por otro lado se realizó una búsqueda de índices de seguridad alimentaria, dónde se encontraron algunos índices como el Índice mundial del hambre (GHI) (31), la Escala de experiencia de inseguridad alimentaria (FIES) de la FAO (32) o el Global Food Security Index [5]. Estos índices pueden ser usados para el hacer una medida de la escala de seguridad alimentaria, sin embargo se debe tener en cuenta la premisa de que se busca un índice que cumpla con las 4 dimensiones descritas por la FAO. De los índices de seguridad alimentaria encontrados el que más destaca es el Índice Global de Seguridad Alimentaria (GFSI) introducido por la Economist Intelligence Unit (EIU). Si bien este índice cumple con las dimensiones y ha demostrado ser acertado, es un índice que en la mayoría de los casos se aplica solamente nivel de países, por la cantidad de datos necesarios y el costo que requiere conseguir estos datos.

2.2. Objetivo del negocio

Implementar un modelo de aprendizaje automático que involucre la integración de imágenes satelitales, datos de nutrición, sociodemográficos y del censo nacional agropecuario para determinar el índice de seguridad alimentaria del Departamento del Cauca.

La utilidad del modelo será para identificar de forma íntegra la seguridad alimentaria en una región específica y ayudará a tomar las respectivas medidas a los encargados de la toma de decisiones.

Cumplir con los objetivos de negocio se definieron 3 objetivos de negocio:

1. Lograr una caracterización y comprensión de los datos disponibles públicamente para el departamento del Cauca, de forma que se pueda entender a grandes rasgos los aspectos tanto nutricionales como de seguridad alimentaria en el Cauca.
2. Generar un conjunto de datos multidimensional público que describa la situación del departamento del Cauca con respecto a los aspectos

nutricionales y de seguridad alimentaria, para su posterior uso en investigaciones y cálculos del GFSI.
3. Entrenar y validar el funcionamiento de un modelo, de forma que se pueda calcular un índice multidimensional que refleje la situación de seguridad alimentaria en el Cauca, teniendo en cuenta las 4 dimensiones descritas por la FAO.

2.3. Criterios de éxito

Para lo cual se debe:

1. Encontrar los posibles datos que describan la situación de seguridad alimentaria en el departamento del Cauca y realizar las respectivas gráficas y estadísticas con las cuales se pueda describir la situación nutricional y de seguridad alimentaria. Estadísticas como promedio, desviación estándar, valores mínimos y máximos deben ser sacados para al menos las características más importantes de cada conjunto de datos localizado.
   Además, gráficas y tablas que puedan reflejar la situación a nivel de municipio o de departamento deben ser mostradas al menos para las características consideradas más importantes de cada conjunto de datos.
2. Hacer uso de técnicas de preprocesamiento e integración de datos para lograr generar un conjunto de datos multidimensional pre-procesando los datos abiertos disponibles y publicar el conjunto de datos en un sitio de acceso público como un repositorio en github o mendeley data de forma que pueda ser usado en un futuro por los encargados de la toma de decisiones, otros investigadores o en para cumplir el objetivo del negocio de este proyecto.
3. Entrenar un modelo capaz de calcular el GFSI para una región específica dados unos datos de entrada. Debido a que no hay trabajos que previamente hayan realizado el cálculo del GFSI haciendo uso de modelos de machine learning, se realizó una estimación de un error que podría ser considerado bueno teniendo en cuenta el rango de valores que puede tomar el modelo como salida (entre 0 y 100). Por lo tanto se consideró que el modelo al predecir valores entre 0 y 100 debe tener un error absoluto medio (MAE) menor a 3.

La relación entre los criterios de éxito, tareas y objetivos del negocio se puede ver en la tabla 5 a continuación:

Objetivo del negocio específicos	Criterios de éxito	Tareas a realizar
Lograr una caracterización y comprensión de los datos disponibles públicamente para el departamento del Cauca, de forma que se pueda entender a grandes rasgos los aspectos tanto nutricionales como de seguridad alimentaria en el Cauca.	Encontrar los posibles datos que describan la situación de seguridad alimentaria en el departamento del Cauca, y realizar las respectivas gráficas y estadísticas con las cuales se pueda describir la situación nutricional y de seguridad alimentaria. Estadísticas como promedio, desviación estándar, valores mínimos y máximos deben ser sacados para al menos las características más importantes de cada conjunto de datos localizado.	Identificar las posibles fuentes de datos, sus proveedores, formatos y métodos de adquisición.
		Realizar la lectura de los respectivos datos haciendo uso de un lenguaje de programación como python.
		Realizar el preprocesamiento respectivo en caso de ser necesario y aplicar técnicas estadísticas, generar tablas y gráficos para la descripción de los datos.
Generar un conjunto de datos multidimensional público que describa la situación del departamento del Cauca con respecto a los aspectos nutricionales y de seguridad alimentaria para su posterior uso en investigaciones y cálculos del GFSI.	Hacer uso de técnicas de preprocesamiento e integración de datos, para lograr generar un conjunto de datos multidimensional pre-procesando los datos abiertos disponibles, y publicar el conjunto de datos en un sitio de acceso público como un repositorio en github o mendeley data, de forma que pueda ser usado en un futuro por los encargados de la toma de decisiones, otros investigadores o en para cumplir el objetivo del negocio de este proyecto.	Realizar el preprocesamiento necesario en los datos, e identificar una variable y formato en común para la integración
		Realizar la integración de los conjuntos de datos en base a una variable común, y generar un archivo que contenga el conjunto de datos.
		Generar un diccionario que explique las variables que contiene el conjunto de datos creado, y proceder a la publicación en un repositorio público como Mendeley data o github.
Entrenar y validar el funcionamiento de un modelo, de forma que se pueda calcular un índice multidimensional que refleje la situación de seguridad alimentaria en el Cauca, teniendo en cuenta las 4 dimensiones descritas por la FAO.	Entrenar un modelo capaz de calcular el GFSI para una región específica dados unos datos de entrada. Debido a que no hay trabajos que previamente hayan realizado el cálculo del GFSI haciendo uso de modelos de machine learning, se realizó una estimación de un error que podría ser considerado bueno teniendo en cuenta el rango de valores que puede tomar el modelo como salida (entre 0 y 100). Por lo tanto se consideró que el modelo al predecir valores entre 0 y 100 debe tener un error absoluto medio (MAE) menor a 3.	Localizar un índice de seguridad alimentaria (GFSI) que cumpla con las dimensiones descritas por la FAO.
		Entrenar uno a varios modelos para el cálculo de un índice de seguridad alimentaria.
		Definir una métrica de error para evaluar el modelo (MAE) y validar el funcionamiento del modelo en un conjunto de datos de evaluación.

Tabla 5. Relación criterios de éxito, tareas y objetivos del negocio.

2.4. Inventario de recursos

2.4.1. Personal

Los recursos humanos participantes del proyecto son David Santiago Restrepo y Luis Enrique Pérez, estudiantes de ingeniería electrónica y telecomunicaciones, los cuales se encargaron de todas las tareas definidas en cada fase de la metodología CRISP-DM. Ambos estudiantes contaron con la supervisión del PhD. Diego Mauricio López Gutiérrez y del PhD. Rubiel Vargas Cañas.

2.4.2. Datos

La realización del proyecto hizo necesaria la recolección de datos provenientes de fuentes públicas o de fácil acceso, con el fin de cumplir con el objetivo del negocio y calcular el índice de seguridad alimentaria para el Departamento del Cauca. Entre los datos obtenidos se encuentran imágenes satelitales provenientes de la plataforma de Google Earth Engine, datos de nutrición obtenidos de la Encuesta Nacional de Situación Nutricional (ENSIN), datos de Censos Agrícolas obtenidos del DANE, datos de salud obtenidos a partir del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), y datos meteorológicos provenientes de WorldClim y Google Earth Engine. En la sección 3 se especifica el origen de los datos, y sus características.

2.4.3. Recursos Hardware

Los recursos hardware utilizados en el proyecto se encuentran en la Tabla 6.

Recursos	**Descripción**
Computador Personal	Hp Laptop 15, procesador i5 octava generación y 8 GB de RAM.
Computador Personal	Macbook Air, Apple Silicon M1 y 8 GB de memoria RAM.

Tabla 6. Recursos Hardware.

2.5. Riesgos y contingencias

Los posibles riesgos y contingencias son los siguientes:

1. El hardware necesario para almacenar y procesar las imágenes satelitales puede no ser suficiente para la cantidad de imágenes y la resolución del Cauca. En este caso múltiples contingencias pueden ser usadas en base a los recursos disponibles.
   Si se poseen los recursos suficientes para hacer el análisis en un solo municipio como Popayán, se puede proceder a hacer el análisis de este municipio como referencia y en una investigación futura realizar el proceso para los demás municipios.
   Otra opción es buscar recursos online como google drive y conectarlo con google colab, sin embargo hay que tener en cuenta que la versión gratuita de Google Colab se reinicia cada 12 horas por lo cual no se puede entrenar un modelo con grán cantidad de imágenes.

2. Los datos necesarios para calcular el GFSI pueden no estar disponibles en su totalidad, ante este riesgo se puede hacer uso de los datos del GFSI de Colombia a escala nacional, o buscar en diversas fuentes estos datos para el Cauca a nivel departamental e imputar estos datos para los municipios.

3. Puede presentarse una pérdida de la información por error humano o por problemas de hardware o software malicioso. Para evitar esto se debe hacer uso de herramientas de almacenamiento en la nube que permiten guardar una copia y registro de la información, en caso de que algún problema suceda localmente se podrá restaurar a la versión de la nube.

2.6. Objetivo de la minería de datos

Con el fin de cumplir con el objetivo del negocio, el objetivo de la minería de datos definido es:

Realizar un modelo de regresión capaz de calcular el índice de seguridad alimentaria GFSI para el Departamento del Cauca a partir de los datos obtenidos de imágenes satelitales, censo agropecuario, encuestas de nutrición, datos meteorológicos.

2.7. Criterios de éxito de la minería de datos

Debido a que el proyecto consta de la integración de imágenes satelitales con un dataset de metadatos, los criterios de éxito estarán dados por las métricas de desempeño de los modelos de extracción de características de imágenes satelitáles y del modelo de cálculo del GFSI, por lo tanto serán:

1. Para darse por exitoso el modelo resultante se debe poder extraer al menos una característica de las imágenes satelitales con una métrica F1-Score de al menos 70%.

2. Para darse por exitoso el modelo encargado de calcular el GFSI del departamento del Cauca se debe tener en cuenta que el índice GFSI tiene un valor entre 0 y 100. Por lo tanto el modelo para calcular el GFSI con un error de 5 o menos sería un caso exitoso.

2.8. Plan del proyecto

La Tabla 7 que se encuentra a continuación contiene el plan general del proyecto. Cada Fase se encuentra explicada en los siguientes capítulos de la monografía.

Fase	**Duración (meses)**
Comprensión del negocio	1
Comprensión de los datos	1
Preparación de los datos	3
Modelado	2
Evaluación	1

Tabla 7. Plan general del proyecto, correspondiente a la metodología CRISP-DM.

2.9. Evaluación inicial de herramientas y técnicas

2.9.1. Lenguajes de programación:

El proyecto al ser un proyecto de minería de datos, se proponen principalmente 2 lenguajes de programación que pueden ser usados. El lenguaje R y Python. Por un lado el lenguaje R el cual es el lenguaje más usado en ciencia de datos y tiene nativamente muchas funciones para análisis, procesamiento y visualización de datos. Por otro lado el lenguaje Python, el cual es un lenguaje de programación ampliamente utilizado debido a que es un lenguaje multipropósito, pero también es un lenguaje que cada vez tiene más fuerza en la ciencia de datos debido en gran parte al soporte de la comunidad y de grandes empresas como Google con su librería Tensorflow o Facebook con Pytorch para deep learning o otras librerías como Pandas, o Numpy para pre-procesamiento y lectura de datos, entre muchas otras.

De los dos lenguajes el elegido es Python debido a el soporte y potencia sobre todo con librerías de deep learning. Esto debido a que da más potencia y aprovechan mejor los recursos computacionales con la conexión a hardware como unidades de procesamiento gráfico (GPU), las cuales son unidades encargadas de hacer procesos computacionales asociados a matrices y arreglos, de forma que funcionan bien para hacer los cálculos matriciales y tensoriales que ocurren en las redes neuronales en paralelo, lo cual es muy útil a la hora de tener que procesar grandes cantidades de datos como es el caso de las imágenes satelitales, ya que nos reduciría el tiempo de entrenamiento.

Además en algunos casos como para el uso de imágenes satelitales la plataforma a usar (Google Earth Engine) tiene su soporte mayormente en Javascript, por lo cual este lenguaje también fue seleccionado para ser usado en este proyecto.

2.9.2. Librerías y/o frameworks:

- Lectura y pre-procesamiento de datos tabulares: Pandas es la librería más usada para lectura y pre-procesamiento de datos tabulares en Python. Presenta facilidad de sintaxis para lectura y visualización de datos, además de métodos que pueden ser usados para pre-procesamiento de los datos.

- Álgebra lineal: Numpy es una librería de Python que permite hacer operaciones entre arreglos de una o múltiples dimensiones, además de ser una librería que trabaja a bajo nivel convirtiendo la sintaxis de Python a C, lo cual da más eficiencia a las operaciones y consume menor tiempo de ejecución.

- Visualización: Matplotlib es la librería más usada en términos de visualización de datos en Python, debido a que permite realizar múltiples tipos de gráficas, además de hacer personalizar la forma como se muestran, tamaños, colores, etiquetas, entre otras características de la figura.
  Por otro lado la librería Seaborn en Python también es ampliamente empleada por su facilidad de uso, sin embargo es una librería construida sobre Matplotlib, por lo tanto no da un poder de personalización tan grande, además de no presentar todos los tipos de gráficas.

- Machine learning: Como librería de Python para usar modelos de machine learning Scikit-learn provee una sencilla sintaxis y una grán cantidad de modelos y operaciones que se pueden a cada conjunto de datos.

- Deep learning: Como librerías de deep learning las librerías más populares en Python son Tensorflow y Pytorch. Estas dos últimas son librerías principalmente de diferenciación automática lo cual permite calcular gradientes y usar algoritmos de “backpropagation” para optimizar los pesos de las neuronas, pero que además han añadido funciones para añadir capas de neuronas, con diferentes tipos de neuronas, funciones de activación, algoritmos de optimización, funciones de pérdida, etc.
  Las dos librerías proporcionan las mismas funciones a la hora de implementar una red neuronal, además de dar soporte a GPU. Si bien las dos librerías permiten trabajar a bajo nivel con tensores y cálculos de derivadas, a bajo nivel Pytorch presenta mayor simplicidad para la creación de modelos y visualización de resultados y operaciones, sin embargo, Tensorflow presenta una API de alto nivel llamada Keras que presenta una sintaxis muy simple para la creación y entrenamiento de modelos de deep learning.
  Si bien las dos librerías son completas y facilitan el trabajo con modelos basados en redes neuronales, en este proyecto se le dará principal preferencia a la librería Tensorflow. La razón es principalmente los recursos hardware que hay disponibles, ya que Pytorch solo presenta soporte a GPUs de Nvidia con lenguaje Cuda, mientras que Tensorflow por en su versión 2.5 para Python 3.9 presenta soporte a GPUs con lenguaje Metal para los computadores Macbook con chip M1.

- Análisis y descarga de imágenes satelitales: Para esta tarea se usó la plataforma Google Earth Engine, la cual es una plataforma de google que puede ser usada para la investigación con datos satelitales de libre acceso. GEE provee petabytes de datos en tiempo real, los cuales pueden ser procesados y analizados usando recursos computacionales de Google. La plataforma está construida principalmente para javascript y la mayoría de la documentación se encuentra en este lenguaje, sin embargo también está la opción de conectarse mediante Python usando la API. Para este trabajo se seleccionaron para trabajar la API de GEE además de la librería Geemap que ayuda a trabajar con GEE en Python y visualizar los resultados.

- Lectura y procesamiento de imágenes: En Python hay 3 librerías principales para esta tarea. La librería PIL propia de Python, Opencv y Skimage. Para este caso las librerías seleccionadas fueron PIL, debido a que es de Python y presenta el mejor soporte para la versión más reciente de Python 3.9. Por

otro lado también se usó la librería Skimage para los casos en los que es necesario leer imágenes con más de una banda, caso que es muy común en imágenes satelitales donde también se toman imágenes con frecuencias por fuera del espectro visible.

2.9.3. Almacenamiento y herramientas de trabajo colaborativo:

En proyectos de desarrollo software dónde diferentes personas trabajan con el mismo código o conjunto de datos, se debe poder compartir el trabajo hecho y las modificaciones casi en tiempo real. Una herramienta que ayuda a mantener un registro de lo que se ha hecho, roles y que además da acceso en tiempo real a cualquier miembro del equipo del proyecto es Github. Github es una herramienta para la creación y gestión de repositorios de forma local y online, por lo tanto es perfecta para este trabajo. Por otro lado Github también presenta cierto límite en cuanto a el tamaño de datos que pueden ser subidos a la plataforma, archivos de más de 100 MB, no pueden subirse de forma convencional y se debe usar Git LFS, pero aún así con archivos de tamaños superiores a 1GB se hace difícil el subirlos a Github online debido a que dependen de las características de la red usada, tamaño de búfer, etc. Por lo tanto para estos casos también se usó Google Drive para almacenar y compartir este tipo de archivos muy pesados.

2.9.4. Técnicas a usar:

Las técnicas a usar se dividirán en las diferentes fases y tareas durante el desarrollo de este proyecto.

1. Primero se debe hacer un análisis de los datos disponibles en el Cauca, de esta forma se podrá entender cuales son los principales componentes de los conjuntos de datos, las variables que pueden usarse y la forma como se puede juntar estos conjuntos de datos en un único dataset.

2. Para la creación del dataset se debe hacer un pre-procesamiento de los datos y juntarlos usando una variable común, la cual puede ser una localización geográfica o una marca de tiempo. El pre-procesamiento requerido dependerá del análisis de los datos.

3. Para el uso de imágenes satelitales con metadatos hay varias opciones.
    1. Se puede usar métodos de co-learning para usar dos entradas de diferente modalidad (imágenes y datos) a un único modelo y entrenar el modelo con los dos tipos de datos al mismo tiempo.

2. También se puede hacer uso de Autoencoders para codificar las imágenes a un vector latente, ó técnicas como entropía condicional o usar las componentes principales de Principal Component Analysis (PCA) para reducir generar datos de menor tamaño que puedan ser usados en conjunto con los metadatos.
3. Se puede entrenar un modelo de red neuronal convolucional, que extraiga características de las imágenes detectando si un elemento se encuentra o no presente en las imágenes.
4. Se puede hacer una clasificación a nivel de píxel (Segmentación semántica) para detectar la cantidad de pixeles de cierta clase en la imágen y conocer qué clases se encuentran y cuales predominan en las imágenes.

De los anteriores métodos el primero requiere de una grán cantidad de datos para poder llegar a aprender y generalizar, además de que este método al igual que el segundo método son poco expresivos en cuanto al significado de los valores que se están usando de la imágen y su peso, por lo cual para el cálculo del índice de seguridad alimentaria no sería ideal.

Los métodos 3 y 4 por otro lado son métodos más expresivos ya que podríamos saber qué elemento se encuentra presente en la imágen. Por lo tanto serán los modelos a ser testeados.

Por un lado el método 3 requiere un conjunto de datos etiquetado para ser entrenado y evaluar el modelo. Por lo tanto se hace uso del dataset “Planet” de Kaggle, el cual nos da imágenes satelitales de regiones incluso del Cauca con etiquetas de características presentes en la imágen como ríos, cultivos, carreteras, entre otras.

En cuanto al método 4 para hacer la segmentación se puede usar técnicas no supervisadas como k-means o Gaussian Mixture para generar clusters con los píxeles. También se pueden usar enfoques supervisados haciendo uso de clasificación a nivel de píxel generando un conjunto de datos con píxeles etiquetados y usarlos para entrenar modelos de machine learning clásicos, como redes neuronales, árboles de decisión, etc. o usando métodos semi-supervisados como generar imágenes segmentadas con técnicas no supervisadas y luego entrenar un modelo supervisado como U-Net.

4. Para el cálculo del GFSI debido a que no existe un GFSI para el Cauca, se debe entrenar modelos de machine learning con los datos a nivel de países, esto con el fin de que el modelo aprenda los diferentes pesos de cada una de las variables. Luego se deben filtrar los datos necesarios para el índice en los

diferentes municipios del Cauca y realizar el pre-procesamiento necesario para que estos datos sean como los datos del GFSI usados para el entrenamiento (normalización, cambios de escala, entre otros).

Capítulo 3

Comprensión de los Datos

3.1. Marco Conceptual

Los siguientes conceptos teóricos son necesarios para comprender el proyecto porque dan explicación a la recolección de los datos y al objetivo que busca realizar el proyecto. Por lo tanto, a continuación se presentan los conceptos más relevantes relacionados a seguridad alimentaria, con el fin de identificar el índice de seguridad alimentaria que se menciona en el proyecto.

3.1.1. Seguridad Alimentaria

Según la cumbre mundial sobre la alimentación (1996), ”la seguridad alimentaria existe cuando todas las personan tienen, en todo momento, acceso físico, social y económico a alimentos suficientes, inocuos y nutritivos que satisfacen sus necesidades energéticas diarias y preferencias alimentarias para llevar una vida activa y sana”, por lo tanto, la seguridad alimentaria está directamente relacionada con la nutrición y los condicionantes necesarios para que una persona pueda llevar una vida saludable.

La organización de las naciones unidas para la alimentación y la agricultura (FAO) define cuatro dimensiones que deben cumplirse para garantizar la seguridad alimentaria, las dimensiones son: disponibilidad física de los alimentos, acceso económico y físico a los alimentos, la utilización de los alimentos y la estabilidad en el tiempo de las tres dimensiones anteriores.

La disponibilidad física de los alimentos corresponde a la cantidad de alimentos que hay en el comercio, es decir, se relaciona con el nivel de producción de alimentos con los que cuenta el país y las existencias de los mismos, con el fin de garantizar que los alimentos alcancen a toda la población.

El acceso económico y físico a los alimentos se relaciona a la capacidad que tienen las personas para adquirir alimentos, esto depende de factores económicos

y de facilidades de acceso, es decir, el hecho de que haya suficiente comida para toda la población no garantiza que todas las personas puedan adquirir los alimentos, por lo tanto se afecta directamente a la seguridad alimentaria.

La utilización de los alimentos tiene relación con los nutrientes que necesita el organismo para poder funcionar adecuadamente. Para garantizar la seguridad alimentaria, es necesario tener la energía y nutrientes suficientes provenientes de los alimentos, para ello establecer buenas prácticas de salud y alimentación es una opción, estas prácticas tienen relación con la correcta preparación de los alimentos, diversidad de la dieta y calidad de los alimentos.

La estabilidad en el tiempo de las tres dimensiones anteriores busca garantizar que la seguridad alimentaria se cumpla en todo momento porque aunque se garantice el acceso, la disponibilidad y la utilización de los alimentos en la actualidad, pueden ocurrir fenómenos en el futuro que afecten a cualquiera de las dimensiones anteriores. Algunos de estos fenómenos pueden estar relacionados con las condiciones climáticas, como por ejemplo las temporadas de sequía o inundaciones. Así mismo, existen factores económicos que pueden afectar la adquisición de los alimentos, como por ejemplo el desempleo o el aumento en el precio de los alimentos.

3.1.2. Global Food Security Index (GFSI)

El GFSI tiene en cuenta las dimensiones de disponibilidad de los alimentos, acceso a los alimentos, calidad e inocuidad de los alimentos y recursos naturales y resiliencia. Este índice se calcula para 113 países, y corresponde a un modelo de puntuación dinámico cualitativo y cuantitativo construido a partir de 59 indicadores que miden la seguridad alimentaria. El objetivo del índice es valorar cuales son los países más y menos vulnerables a la inseguridad alimentaria. El índice incluye otros indicadores cualitativos relacionados a las políticas del gobierno, a factores ambientales y de recursos naturales, que generalmente no son considerados para el cálculo de la seguridad alimentaria.

Las dimensiones definidas por el GFSI son similares a las definidas por la FAO, pero consideran otros factores que generalmente no son tenidos en cuenta para el cálculo de índices de seguridad alimentaria. Las dimensiones del GFSI son:

- Acceso: Se relaciona con la capacidad de los consumidores para adquirir alimentos, la vulnerabilidad frente a la variación de los precios de los alimentos y la presencia de programas y políticas para lidiar con las variaciones de los precios de los alimentos.

- Disponibilidad: Mide la suficiencia de alimentos disponibles en el país, el riesgo de la interrupción de suministro de alimentos, la capacidad nacional de expandir la producción agrícola.
- Calidad e inocuidad de los alimentos: Mide la variedad y calidad nutricional de los alimentos.
- Recursos naturales y resiliencia: Busca medir el impacto del cambio climático en cada país frente a la seguridad alimentaria, su impacto en los recursos naturales y cómo se adapta un país frente a estos cambios.

3.1.3. Minería de datos

La minería de datos es el proceso utilizado para analizar grandes volúmenes de datos y recolectar información útil mediante tareas como el encontrar patrones, descubrir tendencias y ganar conocimiento de cómo los datos podrían ser empleados, para esto se hace uso de métodos matemáticos, debido a que los datos se encuentran en grandes cantidades o hay relaciones complejas entre estos. Por tanto se dice que la minería de datos es interdisciplinaria, porque mezcla los campos de estadística, aprendizaje de máquina e inteligencia artificial. Los resultados del análisis de datos pueden utilizarse para tomar decisiones o predecir un resultado [34].

Para realizar el procedimiento de minería de datos es necesario saber cómo recolectar, almacenar y extraer información útil de los datos [33]. Una de las metodologías más utilizadas para análisis de datos es la CRISP-DM, la cual se mencionó anteriormente y es la utilizada para el desarrollo del trabajo de grado.

3.2. Recolección de datos

Debido a que el objetivo de este trabajo es hacer un análisis con datos abiertos y que cumplan con las 4 dimensiones de la FAO, el proceso de recolección de datos constó de diferentes pasos:

3.2.1. Localización de las diferentes fuentes de datos:

Para localizar los diferentes tipos de datos se debe tener en cuenta las diferentes dimensiones de la FAO y los datos públicos que pueden ayudar para extraer variables de cada dimensión.

Teniendo en cuenta que las dimensiones de la FAO incluyen la disponibilidad física de los alimentos, acceso económico y físico a los alimentos, la utilización de los alimentos y la estabilidad en el tiempo de las tres dimensiones anteriores. Los datos localizados abiertos fueron:

- Datos nutricionales y alimentarios: Estos datos se obtuvieron de la Encuesta Nacional de la Situación Nutricional (ENSIN), la cual es una encuesta realizada por el instituto Colombiano de bienestar familiar cada aproximadamente 5 años y contiene información como las vitaminas consumidas al día por una persona, alimentos consumidos, porciones, ingresos, entre muchas otras variables. Esta encuesta está principalmente relacionada con la dimensión de la FAO de utilización de los alimentos.

- Censo Agrario: Los datos del Censo Nacional Agropecuario son datos recolectados por el Departamento Administrativo Nacional de Estadística. Actualmente el último censo se realizó en el año 2018, sin embargo los datos aún no están disponibles, por lo tanto se trabajó con los datos del censo del año 2014. Estos datos contienen información sobre los agricultores y cultivos alrededor del país. Información como área de cultivo, métodos de cultivo, inversión, entre otras puede ser encontrada aquí. Estos datos corresponden a la dimensión de disponibilidad física de alimentos de la FAO.

- Registros de salud: Usualmente conseguir registros de salud es una tarea compleja y que retrasa muchos proyectos de investigación, sin embargo en Colombia el sistema de vigilancia en salud pública (SIVIGILA) ofrece los conjuntos de datos públicos con el número de registros hospitalarios y sus causas en cada municipio de colombia. De esta forma se pueden obtener datos de interés como desnutrición aguda en menores de 5 años o muertes por desnutrición.
  Estos registros son una buena fuente de información para encontrar la ubicación de los casos en Colombia o en el Cauca y detectar patrones o hacer regresiones de estos datos. Estos corresponden a la dimensión de estabilidad en el tiempo de las otras 3 dimensiones de la FAO.

- Datos Meteorológicos: Los datos meteorológicos pueden ser datos que influyan a los problemas de seguridad alimentaria al afectar el acceso al agua, las carreteras o los cultivos, además de tener un importante impacto en análisis de datos futuros como variables del cambio climático. Estos datos pueden obtenerse de fuentes públicas nacionales como el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) en Colombia. Sin embargo estos son datos que pueden presentar problemas debido a que son puntos geográficos (el lugar donde está la estación que hizo la medida) y no necesariamente representan las características de toda una región del

tamaño de un municipio. Por esta razón se buscaron otras fuentes alternativas como WorldClim o Google Earth Engine.

WorldClim es una página web que posee información en formato Raster (mapas) de características como temperatura o precipitación a través del mundo. Esta información se encuentra mensual para cada uno de los valores entre 1960 y 2018, además de también tener valores de elevación el cual es un mapa único debido a que no varía en el tiempo. Google Earth Engine por otro lado provee más variables y con diferente resolución espacial y temporal dependiendo del satélite que se esté usando y la variable que se quiera medir. Un ejemplo de la interfaz gráfica de Javascript de Google Earth Engine para extraer información de temperatura en Popayán puede verse en la Figura 1.

Figura 1. Ejemplo de interfaz de javascript de google earth engine para visualizar y extraer información de temperatura del satélite Modis.

Para este caso se decidió usar WorldClim como fuente debido a que la resolución espacial y temporal eran suficientes para hacer una estimación de lo que se buscaba. Estos datos estarían relacionados con las dimensiones de disponibilidad física de los alimentos, además de la sostenibilidad en el tiempo debido a la resolución temporal.

- Imágenes satelitales: Las imágenes satelitales se pueden obtener de diferentes plataformas dependiendo del objetivo. En este caso la resolución

tanto espacial como temporal eran muy importantes, sin embargo se debe tener en cuenta la premisa de que deben ser datos abiertos. Ante estas limitaciones la plataforma Google Earth Engine nos ofrece una solución en la cual se puede acceder a grandes cantidades de datos de satélites como Landsat-8 o Sentinel-2 de una forma sencilla a través de la API de Python o Javascript. En este caso el satélite que mejor se adapta es Landsat-8, el cual debido a la fecha de lanzamiento tiene imágenes desde mediados de 2013, a una resolución espacial de 30 m/px que se puede llegar a escalar hasta 15 m/px y una resolución temporal de aproximadamente 15 días entre cada imágen, además de que la colección de Google Earth Engine de este satélite provee imágenes con un pre-procesamiento desde 2013, mientras que en otros casos como Sentinel-2 si bien las imágenes tienen mayor resolución temporal (5 días) y espacial (10 m/px), solo están disponibles desde mediados de 2017. Por esta razón se usaron imágenes de Landsat-8 las cuales cubrirán las dimensiones disponibilidad física de los alimentos con características como cultivos, carreteras y fuentes de agua, y la sostenibilidad en el tiempo debido a la resolución temporal de las imágenes.

- Otras fuentes de datos: Estos datos fueron datos que se fueron buscando en diversas fuentes a medida que se fueron necesitando para poder cubrir algunas variables necesarias en el GFSI que no se encontraban dentro del dataset creado. Aquí se encuentran datos como medida de pobreza que se consiguió haciendo uso de los metadatos en un archivo shapefile, datos de conflicto armado, datos de reportes de entidades del Cauca o incluso datos nacionales usados para calcular el GFSI a nivel nacional por la EIU.

3.2.2. Adquisición de los datos y lectura:

La adquisición de los diferentes datos depende de la entidad. En algunos casos los datos se pueden obtener como archivos limpios en formato csv, o excel, aunque en otros los datos provienen en diferentes formatos como archivos raster, shapefiles, sql, entre otros. Por lo cual el proceso de recolección de datos requiere identificar estos datos y saber leerlos e interpretarlos. Si bien para hacer la extracción de algunos datos como las imágenes satelitales se requirió de un pre-procesamiento, estos no serán explicados en esta sección.

- Datos nutricionales y alimentarios: Estos datos se encuentran originalmente en formato ‘.data’ y ‘.sql’. Debido a que son datos de una base de datos relacional, estos datos vienen divididos en varios archivos y tienen un diagrama para comprender su estructura, así como un archivo que contiene

un diccionario que relaciona las variables con su significado. Estos archivos fueron convertidos a formato csv usando el software STATA. Una vez convertidos a formato csv pueden ser leídos usando la librería pandas de Python.

- Censo Agrario: Los datos del Censo Nacional Agropecuario se encuentran en archivos csv, los paquetes de archivos con los datos se descargan individualmente por departamento, por lo tanto en este caso solo se descargó el paquete de datos del Departamento del Cauca. Los datos descargados contienen varios archivos csv y un diccionario de variables que explica el significado de cada variable, además de las variables que pueden ser usadas como valores para relacionar los archivos.
  Estos archivos pueden ser fácilmente leídos usando la librería Pandas de Python.

- Registros de salud: Estos archivos son tablas de excel que contienen los informes semanales por municipio. Se genera un archivo por año, por lo cual se debe descargar el archivo correspondiente a cada año y aplicar el procesamiento necesario para cada uno de estos archivos.
  Cada archivo puede ser leído con el uso de la librería Pandas de Python, aunque el pre-procesamiento necesario para juntar los registros de salud es más complejo que el de otros datos como el Censo Nacional Agropecuario debido a que en los archivos de SIVIGILA hay cambios de formato cada año en algunas variables.

- Datos Meteorológicos: Los datos meteorológicos como se dijo anteriormente se usaron de la página web WorldClim, la cual nos permite descargar los archivos raster de cada variable a nivel global. Estos archivos pueden ser leídos y visualizados usando diferentes herramientas. En este caso se usó R con las librerías raster, maptools y rgdal, además de un csv que provee el DANE con las coordenadas de cada municipio y un archivo shapefile que contiene los polígonos de cada municipio y metadatos con información de cada polígono para leer los archivos raster y extraer los valores de las variables para cada municipio en formato csv.
  Un ejemplo de un archivo raster visualizado en R junto con el archivo shapefile de los municipios de Colombia puede verse en la figura 2.

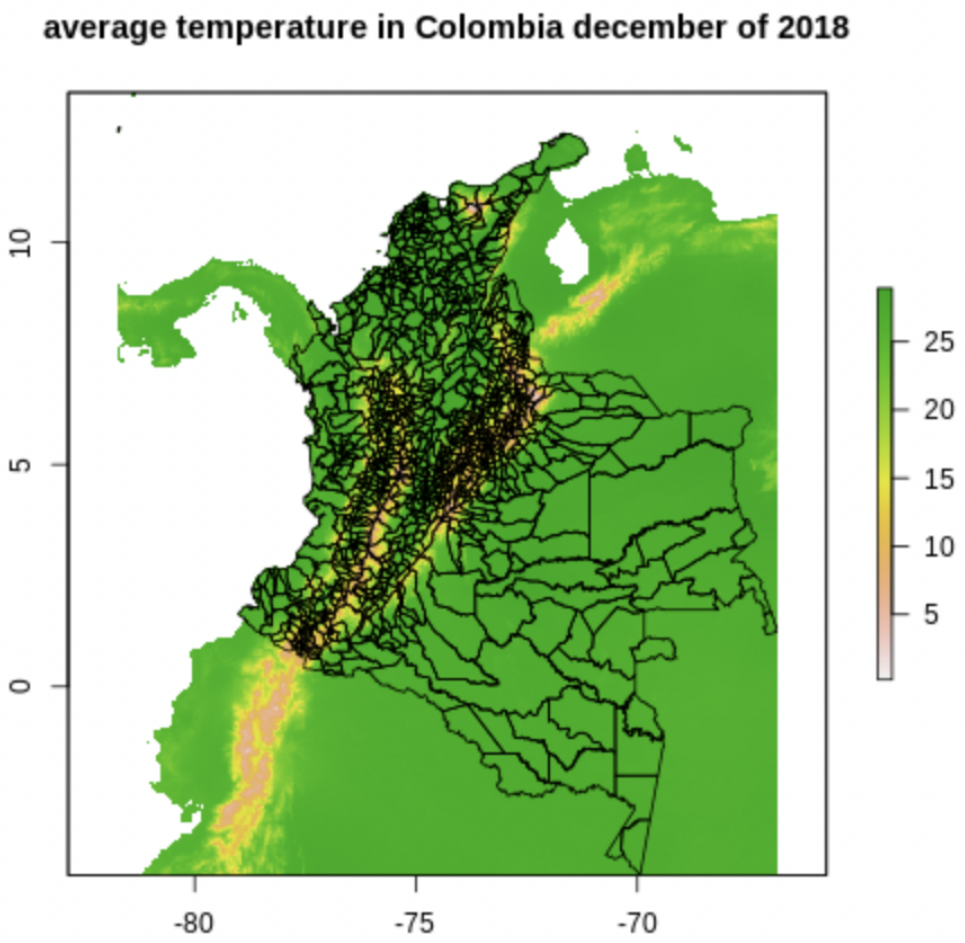
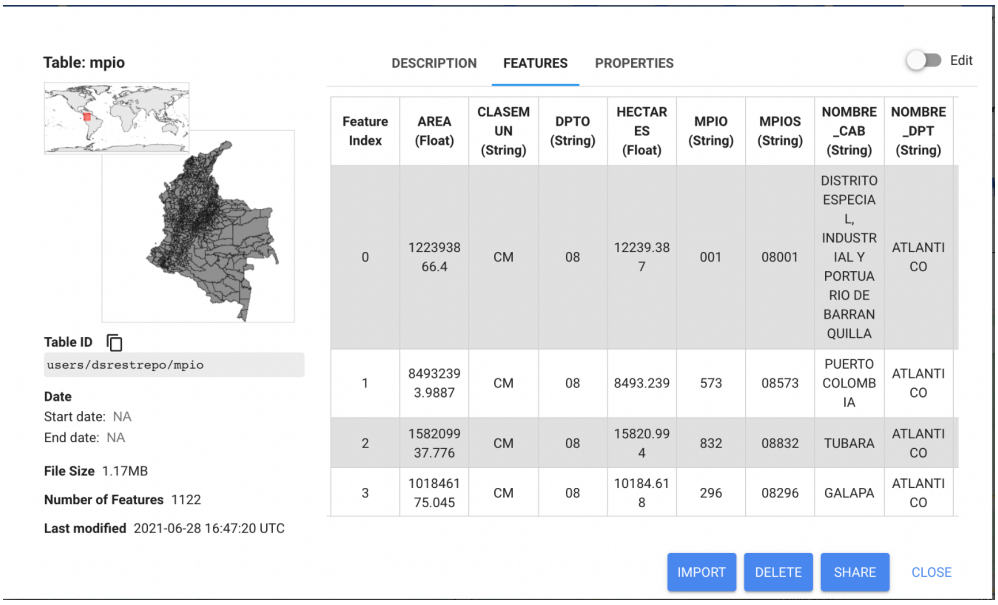


Figura 2. Raster y shapefile de precipitación en Colombia en diciembre de 2018.

- Imágenes satelitales: Las imágenes satelitales se pueden acceder mediante la plataforma GEE, aquí se puede hacer una solicitud al servidor para obtener una colección de imágenes usando un identificador único de colección. Una vez se obtiene el identificador de la colección a usar, se puede filtrar las imágenes por temporalidad o localización, de forma que las imágenes sean solo las correspondientes a una región geográfica específica. Para filtrar por localización se usó un archivo “shapefile” que contiene las coordenadas geográficas de cada municipio en Colombia y se puede cargar individualmente en GEE usando una clase de la plataforma la cual es un ee.Polygon, básicamente un polígono ubicado en el mapa, un ejemplo del archivo usado se puede ver en la figura 3. Cuando ya se tiene el conjunto de imágenes a descargar y se ha hecho su correspondiente

pre-procesamiento que se explicará más adelante, se puede hacer la solicitud al servidor para descargar estas imágenes directamente al computador o extraerlas a una carpeta en google drive en formatos como “png”, “jpg” o “tif”.



Feature Index	AREA (Float)	CLASEMUN (String)	DPTO (String)	HECTARES (Float)	MPIO (String)	MPIOS (String)	NOMBRE_CAB (String)	NOMBRE_DPT (String)
0	122393866.4	CM	08	12239.387	001	08001	DISTRITO ESPECIAL, INDUSTRIAL Y PORTUARIO DE BARRANQUILLA	ATLANTICO
1	84932393.9887	CM	08	8493.239	573	08573	PUERTO COLOMBIA	ATLANTICO
2	158209937.776	CM	08	15820.994	832	08832	TUBARA	ATLANTICO
3	101846175.045	CM	08	10184.618	296	08296	GALAPA	ATLANTICO

Figura 3. Shapefile de los municipios de Colombia.

- Otras fuentes de datos: Para este tipo de datos las técnicas usadas fueron muchas, principalmente fue una imputación manual de los datos en el conjunto de datos usados para realizar las predicciones en el modelo de GFSI. La razón de realizar esta tarea manual es que en muchos casos estas variables se encontraban en informes presentados por diferentes entidades en formatos como pdf, lo cual dificulta la lectura de este tipo de archivos. Sin embargo hay casos en los que los tipos de datos se pudieron añadir automáticamente al conjunto de datos como es el caso de la medida de pobreza multidimensional, la cual se encontraba en los metadatos de un archivo shapefile y se transformó a un archivo csv usando la librería geopandas de Python ó los datos de conflicto armado que se encontraban en un archivo excel, el cual se leyó y pre-procesó usando la librería pandas de Python.

3.3. Descripción de los datos

Los datos obtenidos se usaron en la construcción de un dataset que permita el cálculo del índice de seguridad alimentaria definido por el GFSI, para ello fue necesario encontrar datos que cubran las cuatro dimensiones dispuestas por la FAO. Estos datos provienen de imágenes satelitales, censos agropecuarios, encuestas de nutrición, datos meteorológicos, los cuales se encuentran explicados a continuación en las siguientes secciones. Los trabajos de obtención de los datos y la construcción del dataset fueron llevados a cabo por David Santiago Restrepo y Luis Enrique Pérez.

3.3.1. Descripción de los datos: Imágenes satelitales.

La plataforma de Google Earth Engine ofrece un dataset de imágenes satelitales tomadas por el satélite Landsat 8. Este dataset contiene imágenes capturadas desde Abril del 2013 hasta el presente. Las imágenes que provee el satélite, son tomadas cada 16 días aproximadamente dependiendo de la localización geográfica y tienen diferentes bandas del espectro electromagnético, tanto visibles como no visibles. Las bandas dentro del espectro visible son B4, B3, B2 corresponden a las bandas Rojo, Verde y Azul respectivamente. Estas bandas permiten construir una imágen RGB con una resolución de 30 metros por pixel (m/px). También se puede usar la banda Pancromática (B8) que tiene una resolución de 15 m/px y está ubicada en el espectro intermedio de las ya mencionadas B4, B3 y B2. El objetivo de usar la banda pancromática es mejorar la resolución de la imágen RGB obtenida previamente.

El objetivo de la recolección de estos datos, es obtener información de los cultivos y las fuentes de agua, ambos pertenecientes a la dimensión de disponibilidad física de la FAO. También se puede obtener información de la presencia de carreteras, relacionado a la dimensión de acceso físico y económico de la FAO y la información de escala temporal se asocia con la dimensión relacionada a la sostenibilidad de las dimensiones en el tiempo.

3.3.2. Descripción de los datos: Datos de nutrición.

Los datos de nutrición provienen de la encuesta realizada por la ENSIN para el 2015, estos datos se encuentran estructurados en una base de datos con sus correspondientes llaves y relaciones entre cada una. Para entender la estructura y los datos que contiene la encuesta, se cuenta con un diccionario, llamado Dic_Ensin_Pública.xlsx. Dentro del diccionario se encuentra la hoja de cálculo llamada Detalle de Tablas, la cual explica cómo se dividen los formularios y el

contenido de cada tabla junto con sus variables. Mediante el “Detalle de Tablas” se seleccionó las tablas relacionadas con la seguridad alimentaria las cuales se presentan en la siguiente Tabla 8:

Formulario	Tabla	Capítulo, variables contenidas
Hogar	PTS	Condiciones Habitacionales
	PTS2	Actividad económica, ingresos y gastos
	SA_1	Experiencia inseguridad alimentaria
	SA_2	Datos seguridad alimentaria (alimentos)
	SA_3	Datos seguridad alimentaria (animales)
	SA_4	Datos seguridad alimentaria (otras fuentes de obtención de alimentos)
R24 y Prácticas de Alimentación	PISNSP	Frecuencia de consumo, prácticas entorno a seguridad alimentaria, Inocuidad y calidad de los alimentos
Vitaminas y Minerales	VIT	Preguntas vitaminas y minerales por personas
Recordatorio de 24 Horas	R24	Información de control e identificación

Tabla 8.Contenido Formularios ENSIN.

Los archivos entregados por el ENSIN se encuentran en formato STATA, estos fueron utilizados para extraer la información necesaria para cumplir con las dimensiones de la FAO y establecer el índice GFSI. Cada archivo lleva por nombre las tablas mencionadas anteriormente. Para ver el contenido de cada tabla y las variables que almacena, la ENSIN brinda una hoja de cálculo llamada “Ensin Pública” dentro del archivo Dic_Ensin_Pública.xlsx que contiene la información de las variables que almacena cada tabla, la descripción de cada variable y los valores respuesta que contiene cada variable. Este archivo fue utilizado para extraer los datos o variables importantes para la construcción del dataset.

Cada archivo cuenta con un identificador o llave primaria que permite establecer la relación entre los demás archivos, es decir, cada encuesta fue realizada en algún hogar de cada municipio, cada hogar encuestado cuenta con un código de referencia único llamado LLAVE_HOGAR, y cada persona encuestada cuenta con un código de referencia único llamado LLAVE_PERSONA. Cada tabla cuenta con la LLAVE_HOGAR o LLAVE_PERSONA, es por esto que las tablas SA_1, SA_2,

SA_3, SA_4, PTS, PTS_2 cuentan con la LLAVE_HOGAR, mientras que VIT, PISNSP cuentan con la LLAVE_PERSONA. El archivo R24 cuenta con ambas llaves y también contiene el código de municipio, por lo que es el archivo clave para unir todos los archivos en uno y así clasificarlos por municipio. En el **anexo D** se encuentra el diccionario de variables seleccionadas de la encuesta ENSIN.

Las variables más importantes que se encuentran por tablas están a continuación (Tabla 9):

Tabla	Variable	Descripción
SA_1	inseguridad	Es el resultado de la respuesta a 4 preguntas anteriores, y mide la presencia de inseguridad alimentaria.
SA_2	cultivo_0.0 - cuultivo_7.0	Corresponde a varias respuestas relacionadas al consumo de alimentos provenientes de la agricultura.
SA_3	Alimento_0.0 - Alimento_3.0	Corresponde a varias respuestas relacionadas al consumo de alimentos provenientes de la carne, huevo, leche.
PTS	noacueducto	Corresponde a la presencia de acueducto en el hogar encuestado.
	noenergia	Variable que contiene información relacionada a la presencia del servicio de energía público en el hogar en cuestión.
PTS_2	Ing_dest_alim	Variable que almacena información del consumo de alimentos en el hogar.
VIT	FERRITINA	Información relacionada al nivel de hierro en la sangre, tomado para cada persona.
	VITAMINAA	Información relacionada al nivel de vitamina , tomado para cada persona.
PISNSP_	Huevos	Corresponde a la frecuencia de consumo de huevos por cada persona.
	Verduras	Corresponde a la frecuencia de consumo de verduras por cada persona.
	Carnes	Corresponde a la frecuencia de consumo de carnes por cada persona.
R24	c_mpio	Contiene el código de municipio.

Tabla 9. Variables Dataset ENSIN.

El archivo final que contiene el dataset creado a partir de los datos del ENSIN, se llama ENSIN.ipynb (https://github.com/dsrestrepo/FoodSecurity/blob/main/ENSIN/ENSIN.ipynb).

3.3.3. Descripción de los datos: Datos del Censo Agropecuario.

La información del censo agropecuario proviene del Departamento Administrativo Nacional de Estadística (DANE) y corresponde al censo del 2014. Los datos que se encuentran en la encuesta corresponden a la respuesta de agricultores y granjeros, respecto a las condiciones de los cultivos, acceso a financiamiento para cultivos o agricultura, infraestructura, áreas plantadas, entre otros. Estos datos se encuentran en formato .csv, y al igual que para el ENSIN, el DANE brinda un diccionario de datos llamado TEMATICA_DISENO_DE_REGISTRO_CNA2014.xlsx, el cual contiene la información de las variables que contiene cada tabla de datos, además de la descripción de cada variable y los valores que puede tomar cada una.

Al igual que en el caso del ENSIN, este diccionario fue útil para extraer la información necesaria para cumplir con las dimensiones de la FAO. El archivo CNA2014_ENCABEZADO_19.csv contiene todas las variables utilizadas para obtener los datos del censo agropecuario, contiene la variable P_MUNIC la cual contiene el código de municipio de la persona a la que se realizó la encuesta. Esta variable es muy importante porque es la usada para establecer relaciones con los demás datasets. Algunas variables utilizadas se encuentran en la Tabla 10. En el **anexo D** se encuentra el diccionario de variables seleccionadas del censo nacional agropecuario.

Nombre	Variable	Descripción	Valores
Credito_banco	P_S11P136B_SP1	¿Cuales fueron las fuentes de los créditos o financiaciones aprobados?Banco Agrario	1:Si 0:No
	P_S11P136B_SP2	¿Cuales fueron las fuentes de los créditos o financiaciones aprobados?Otros bancos	1:Si 0:No
	P_S11P136B_SP3	¿Cuales fueron las fuentes de los créditos o financiaciones aprobados?Cooperativa	1:Si 0:No

	P_S11P136B_SP4	¿Cuales fueron las fuentes de los créditos o financiaciones aprobados?Particulares o prestamistas	1:Si 0:No
inversion_tecnologia	P_S11P137_SP3	Los recursos de financiación fueron destinados para:Compra de maquinaria de uso agrícola	1:Si 0:No
	P_S11P137_SP4	Los recursos de financiación fueron destinados para:Compra de maquinaria de uso pecuario (incluye pesca)	1:Si 0:No
	P_S11P137_SP10	Los recursos de financiación fueron destinados para:Procesos poscosecha	1:Si 0:No
area_cultivo	P_S12P142	Hoy; ¿Cuánta es el área con cultivos presentes transitorios; cultivos permanentes; plantaciones forestales; pastos sembrados y pastos o sabanas naturales? (ÁREA CON USO AGROPECUARIO)	Metros
agua_acueducto	P_S11P124_SP8	Las fuentes del agua que utiliza para las act. agropecuarias son: Acueducto	1:Si 0:No
no_agua_contaminacion	P_S11P126_SP1	Durante 2013; ha tenido dificultades en el uso del agua para el desarrollo de las act. agropecuarias por: Contaminación	1:Si 0:No
P_MUNIC		Código de Municipio	Numérico

Tabla 10. Variables Dataset Censo Agropecuario.

Los datos obtenidos del censo agropecuario se encuentran en el archivo CensoAgro.ipynb (https://github.com/dsrestrepo/FoodSecurity/blob/main/Censo%20Nacional%20Agropecuario/CensoAgropecuario.ipynb).

3.3.4. Descripción de los datos: Registros de Salud.

Los registros de salud contienen datos semana a semana de cada uno de los reportes en emergencias en los hospitales de Colombia. Este conjunto de datos nos entrega el número de casos en un municipio en base a la semana epidemiológica. Actualmente en la página de sivigila los datos se encuentran desde el año 2009, sin embargo se vió que los datos de interés (Desnutrición aguda en menores de 5 años y muertes por desnutrición) solo se encuentran

disponibles a partir de 2013 en el caso de mortalidad por desnutrición y 2016 en el caso de desnutrición aguda en menores de 5 años.

3.3.5. Descripción de los datos: Datos Metereológicos.

Los datos meteorológicos de worldclim se encuentran en archivos en formato raster. Los raster son unos mapas que contienen información específica de una variable en un punto basado en su localización geográfica. En este caso se tenían raster para precipitación, temperatura máxima promedio, temperatura mínima promedio y elevación entre los años 1960 y 2018.

3.3.6. Descripción de los datos: Otras fuentes varias de datos.

Los datos de otras fuentes contenían información necesaria para calcular el GFSI para el departamento del Cauca, pero que no se encontraban en los otros conjuntos de datos previamente explicados. Estos contenían información como conflicto armado, población de los municipios, medidas de pobreza, estratificación, valor del producto interno bruto (PIB) en el Cauca, entre otros valores.

3.4. Exploración de los datos y caracterización del departamento del Cauca en base a los datos recolectados.

En esta sección se analizará los datos recolectados en el departamento del Cauca en los diferentes conjuntos de datos encontrados.

3.4.1. Exploración de los datos: Registros de salud.

Los registros de salud se dividieron en 2 tipos de registros que pueden ser usados como una estimación para medir el estado nutricional en el Cauca. Por un lado la desnutrición aguda en menores de 5 años y por otro lado la mortalidad por desnutrición. A continuación un análisis de los 2 tipos de datos:

3.4.1.1. Exploración de registros de desnutrición aguda en menores de 5 años en el Cauca.

La desnutrición aguda en menores de 5 años en el Cauca es una medida que si bien no incluyen a toda la población, se pueden usar como referencia para saber qué está pasando en cuanto a la situación nutricional en el departamento, debido a que esta población (menores a 5 años) son los más afectados por la desnutrición y a los que mayores secuelas les puede producir este problema a

largo plazo. Además que si estos problemas se presentan en menores a 5 años, también se presenta en otras edades pero de forma menos evidente.

Con los datos obtenidos de sivigila entre los años 2016 y 2019 los casos de desnutrición en el Cauca tienen una tendencia a aumentar como se ve en la figura 4 al igual que en país (figura 5).

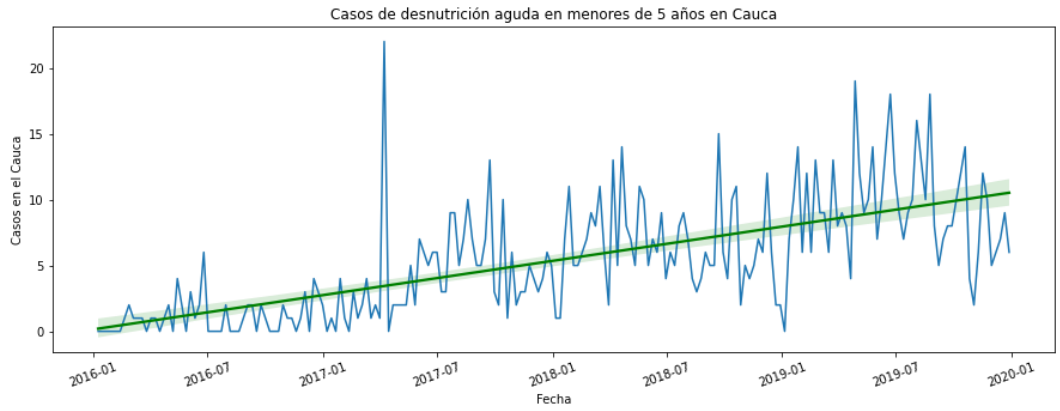


Figura 4. Series de tiempo de desnutrición aguda en menores de 5 años en el Cauca.

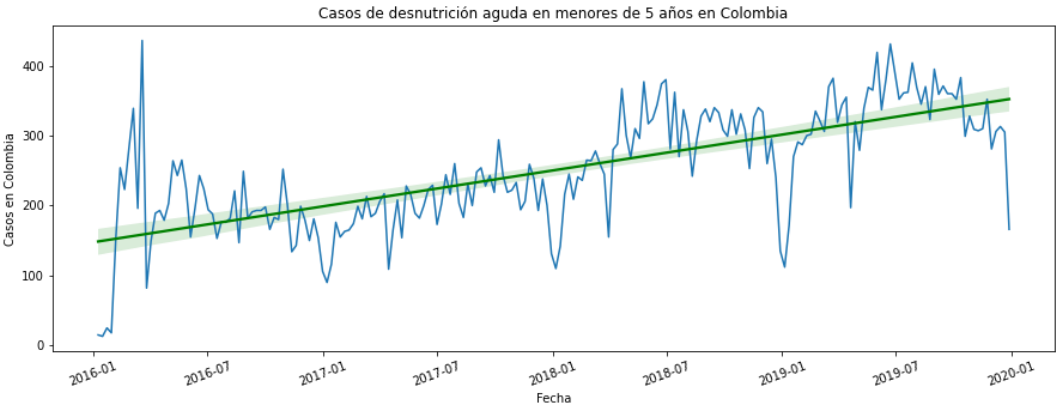


Figura 5. Series de tiempo de desnutrición aguda en menores de 5 años en Colombia.

Además se ve en la tabla 11 que de los municipios el más afectado es el Popayán con número de casos promedio semanales de 0.6875, mientras que el municipio con menor cantidad de casos en promedio es Padilla con un promedio de casos semanales de 0.0048. En total en el Cauca se presentan cada semana 5,36 casos

de desnutrición aguda en menores de 5 años de los 250,4 en promedio cada semana a nivel nacional. Es decir un 2.14% del total del país.

	Promedio	Desviación estándar	mínimo	máximo
POPAYÁN	0,6875	1,22905151757371	0	12
ALMAGUER	0,0913461538461538	0,305065313961186	0	2
ARGELIA	0,0528846153846154	0,224343047690795	0	1
BALBOA	0,0432692307692308	0,203953401458655	0	1
BOLÍVAR	0,0144230769230769	0,119514532684749	0	1
BUENOS AIRES	0,0625	0,702609009542189	0	10
CAJIBÍO	0,490384615384615	1,1163708654956	0	7
CALDONO	0,139423076923077	0,386716206029623	0	2
CALOTO	0,0625	0,261798868108698	0	2
CORINTO	0,0336538461538462	0,180771602726439	0	1
EL TAMBO	0,00961538461538462	0,0978209255833176	0	1
FLORENCIA	0,0144230769230769	0,119514532684749	0	1
GUACHENÉ	0,0144230769230769	0,119514532684749	0	1
GUAPI	0,158653846153846	0,480360286870727	0	3
INZÁ	0,346153846153846	0,758452922810903	0	5
JAMBALÓ	0,129807692307692	0,447727446356303	0	3
LA SIERRA	0,0432692307692308	0,203953401458655	0	1
LA VEGA	0,0336538461538462	0,205767364032972	0	2
LÓPEZ DE MICAY	0,100961538461539	0,346687622640768	0	2
MERCADERES	0,0721153846153846	0,294213675931567	0	2
MIRANDA	0,0528846153846154	0,298290372024247	0	3
MORALES	0,302884615384615	0,715319830601686	0	5
PADILLA	0,00480769230769231	0,0693375245281536	0	1
PÁEZ	0,350961538461538	0,843788923431586	0	7
PATÍA	0,0769230769230769	0,301118933997421	0	2
PIAMONTE	0,139423076923077	0,422533940947142	0	2
PIENDAMÓ - TUNÍA	0,317307692307692	0,882075101438774	0	5
PUERTO TEJADA	0,0913461538461538	0,388633323680549	0	3
PURACÉ	0,177884615384615	0,522065154237875	0	4
ROSAS	0,0192307692307692	0,169156431090029	0	2
SAN SEBASTIÁN	0,0432692307692308	0,203953401458655	0	1
SANTANDER DE QUILICHAO	0,197115384615385	0,560013238416543	0	4
SANTA ROSA	0,0144230769230769	0,119514532684749	0	1

SILVIA	0,326923076923077	0,563671670840713	0	2
SOTARÁ PAISPAMBA	0,00961538461538462	0,0978209255833176	0	1
SUÁREZ	0,105769230769231	0,365631456452028	0	2
SUCRE	0,0336538461538462	0,205767364032972	0	2
TIMBÍO	0,0288461538461539	0,167777790082713	0	1
TIMBIQUÍ	0,307692307692308	0,607154177061437	0	4
TORIBÍO	0,0625	0,242645444206568	0	1
TOTORÓ	0,0817307692307692	0,274615030701759	0	1
VILLA RICA	0,0144230769230769	0,119514532684749	0	1
CAUCA	5,36057692307692	4,45479534860121	0	22
COLOMBIA	250,403846	83.962842	13.0	436.0

Tabla 11. Estadísticas de desnutrición aguda en menores de 5 años.

3.4.1.2. Exploración de registros de mortalidad por desnutrición.

Para lo datos de mortalidad por desnutrición en comparación con los datos de desnutrición aguda en menores de 5 años, hay mucho menos muestras y puede que no sean suficientes para entender la gravedad de los problemas de seguridad alimentaria y nutricional, esto debido a que son casos extremos que se presentan en muy pocas ocasiones y que hay que tener en cuenta que los problemas nutricionales como la desnutrición, no necesariamente son la causa principal de muerte, sino que también puede desencadenar otras enfermedades. Como se puede observar en la gráfica 1, los datos de muerte por desnutrición en el Cauca son muy pocos y fluctúan entre 0 y 1 caso por semana como se puede confirmar en la tabla 12. Por otro lado como se ve en la figura 7, los casos de mortalidad por desnutrición en Colombia en general si son mayores, pero muy poco siendo el valor máximo 15 casos en una semana en toda Colombia.

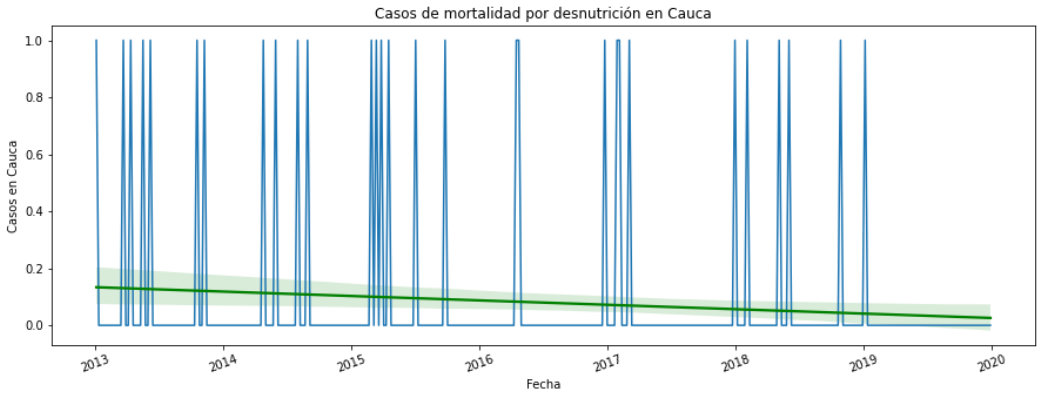


Figura 6. Series de tiempo de mortalidad por desnutrición en el Cauca.

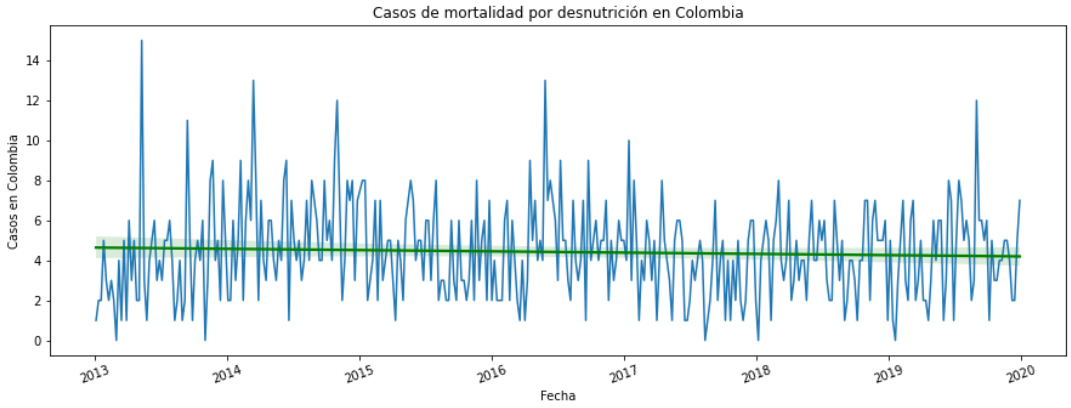


Figura 7. Series de tiempo de mortalidad por desnutrición en Colombia.

	promedio	desviación estándar	mínimo	máximo
POPAYÁN	0,00274725274725275	0,0524142418360959	0	1
ALMAGUER	0,00274725274725275	0,0524142418360959	0	1
BALBOA	0,00274725274725275	0,0524142418360959	0	1
BUENOS AIRES	0,00274725274725275	0,0524142418360959	0	1
CALDONO	0,00274725274725275	0,0524142418360959	0	1
CALOTO	0,00274725274725275	0,0524142418360959	0	1
GUAPI	0,00549450549450549	0,0740227607951281	0	1
INZÁ	0,00274725274725275	0,0524142418360959	0	1
JAMBALÓ	0,00549450549450549	0,0740227607951281	0	1
MORALES	0,00274725274725275	0,0524142418360959	0	1
PÁEZ	0,010989010989011	0,104394409411493	0	1
PURACÉ	0,00274725274725275	0,0524142418360959	0	1
SAN SEBASTIÁN	0,00274725274725275	0,0524142418360959	0	1
SANTANDER DE QUILICHAO	0,00549450549450549	0,0740227607951281	0	1
SILVIA	0,010989010989011	0,104394409411493	0	1
SUÁREZ	0,00274725274725275	0,0524142418360959	0	1
SUCRE	0,00274725274725275	0,0524142418360959	0	1
TIMBIQUÍ	0,00549450549450549	0,0740227607951281	0	1
TOTORÓ	0,00274725274725275	0,0524142418360959	0	1
CAUCA	0,0796703296703297	0,271154863295122	0	1
COLOMBIA	4,42032967032967	2,37450273781683	0	15

Tabla 12. Estadísticas de mortalidad por desnutrición.

3.4.2. Exploración de los datos: Censo Nacional Agropecuario.

Los datos del Censo Nacional Agropecuario se dividieron y agruparon en unas variables más generales. La exploración de los datos del censo se hará con base a estas variables agrupadas para hacer la explicación más concisa.

3.4.2.1. Datos de financiamiento a agricultores.

Los datos de financiamiento consisten en los datos de las encuestas hechas a los agricultores sobre los créditos realizados, créditos aprobados y entidades que aprobaron estos créditos. Las estadísticas en general calculadas de estas variables se pueden ver en la tabla 13 donde se observa que en el Cauca de los agricultores encuestados, el 11.06% solicitaron créditos y de los cuales el 86,97% fueron aprobados.

	Créditos solicitados (%)	créditos aprobados (%)	Créditos solicitados a bancos (%)	Créditos solicitados a ONG (%)	Créditos solicitados a gobierno (%)	Créditos solicitados a onu (%)
PADILLA	31,0097719	96,8487394	17,8571428	0,210084033	78,78151260	0
PUERTO TEJADA	27,6178010	96,6824644	27,4881516	1,895734597	68,24644549	0
MIRANDA	25,2747252	85,6521739	42,6086956	0,652173913	42,17391304	0,2173913043
TIMBÍO	23,5595828	90,3483309	93,2510885	0,870827285	0	0,0725689404
PURACÉ	20,7463884	87,6208897	87,4274661	0,580270793	0	0
SANTANDER DE QUILICHAO	20,3981042	90,7063197	85,8736059	0,557620817	6,319702602	0
TOTORÓ	17,7788484	88,4823848	88,21138211	1,626016260	0,135501355	0
LA SIERRA	17,6789168	82,7133479	83,8074398	0,656455142	0,437636761	0
PÁEZ	17,4960296	90,4689863	90,7715582	0,605143721	0	0
ROSAS	16,2288382	94,2446043	96,4028776	0,359712230	0,179856115	0
CAJIBÍO	15,9443660	89,2541087	91,9089759	0,379266750	0,126422250	0
MORALES	15,4947106	78,3132530	77,6104417	1,706827309	0,100401606	0
PIENDAMÓ - TUNÍA	15,3523238	91,015625	95,8007812	0,29296875	0,1953125	0,1953125
GUAPI	15,0661299	88,3587786	88,5496183	0	0	0
INZÁ	14,8126911	90,9677419	85,5483870	7,483870967	1,161290322	0,1290322580
CALDONO	14,3427567	92,3162583	95,7683741	0,445434298	0,222717149	0,1113585746
POPAYÁN	14,2988929	88,2580645	90,5806451	0,258064516	0,774193548	0,1290322580
CALOTO	13,9698736	89,5652173	84	1,217391304	5,739130434	0,3478260869

EL TAMBO	13,5459425	85,6060606	89,2773892	0,757575757	0,641025641	0
TORIBÍO	13,0978130	84,9541284	69,9082568	12,11009174	1,834862385	0
PIAMONTE	12,6564673	86,8131868	84,6153846	0	0	0
BALBOA	12,6499207	88,9087656	88,9087656	0,357781753	0,357781753	0
SUÁREZ	12,4006359	86,5384615	87,6068376	0,427350427	0	0
JAMBALÓ	12,3229461	88,2183908	85,6321839	1,149425287	0,574712643	0
FLORENCIA	11,7866004	91,5789473	92,1052631	2,631578947	0	0
SOTARÁ PAISPAMBA	11,5639222	84,2576028	87,2987477	1,431127012	1,073345259	0,1788908765
ARGELIA	11,1349036	72,8365384	73,7980769	0	0,240384615	0,2403846153
SILVIA	10,4712041	87,5581395	89,0697674	0,348837209	0,697674418	0,1162790697
SUCRE	9,17827967	88,4816753	89,0052356	0,523560209	0,523560209	0
BUENOS AIRES	9,03614457	77,6068376	75,3846153	1,025641025	2,222222222	0
PATÍA	8,97274633	80,3738317	82,0093457	0,467289719	0	0,2336448598
GUACHENÉ	8,72363246	89,7540983	47,5409836	1,639344262	43,44262295	0
MERCADERES	7,46210649	91,6666666	93,75	0,78125	0,260416666	0
TIMBIQUÍ	6,02006688	65,0793650	63,4920634	0,793650793	0	0
LA VEGA	5,59427387	86,5319865	83,5016835	0,673400673	0,336700336	0,3367003367
SANTA ROSA	5,41686292	56,5217391	60	0	0,869565217	0
VILLA RICA	5,21172638	91,25	25	1,25	66,25	0
CORINTO	4,41527446	87,3873873	31,5315315	0,900900900	55,85585585	0
SAN SEBASTIÁN	3,13185092	74	70	0,5	0	0
LÓPEZ DE MICAY	2,49650023	66,3551401	65,4205607	0,934579439	0	0
BOLÍVAR	2,31415643	86,7783985	90,1303538	0,372439478	0,186219739	0,1862197392
ALMAGUER	1,98888563	77,9411764	76,4705882	0,490196078	0,490196078	0
CAUCA	11,0619469	86,9675977	83,0122905	1,224581005	4,992178770	0,0670391061

Tabla 13. Estadísticas de créditos del censo nacional agropecuario en Cauca.

También se puede ver tanto en la Tabla 13 como en la figura 8 como los municipios que más créditos solicitan son Padilla y Puerto Tejada, mientras que los que menos realizan son Almaguer y Bolívar (Figura 9).

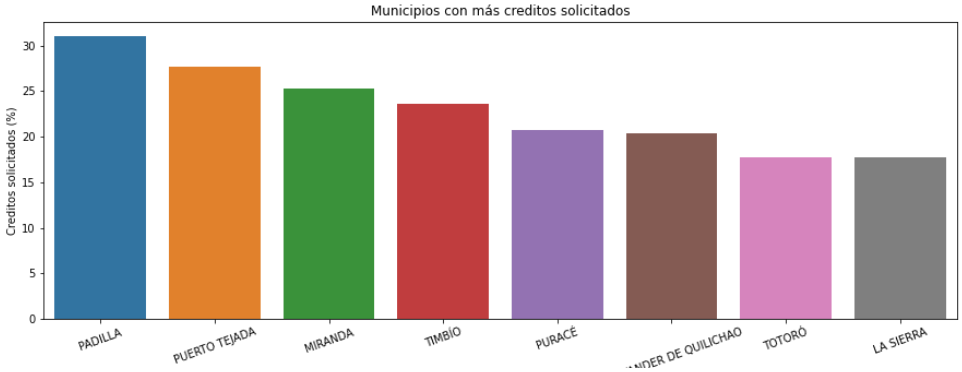


Figura 8. Municipios que más créditos realizan para actividades agropecuarias en Cauca.

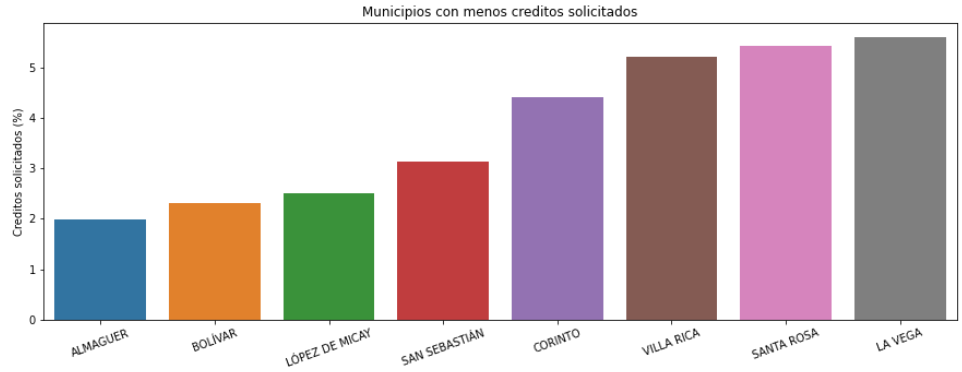


Figura 9. Municipios que menos créditos realizan para actividades agropecuarias en Cauca.

Si por un lado, la cantidad de créditos solicitados puede dar una idea de las necesidades y la inversión de cada municipio en agricultura; el saber si estos créditos se aprobaron o no también es un factor importante a analizar. En este caso la figura 10 muestra cómo Padilla y Puerto Tejada fueron los que mayor porcentaje de créditos aprobados tuvieron, mientras que Santa Rosa y Timbiquí los que menos créditos aprobados tuvieron.

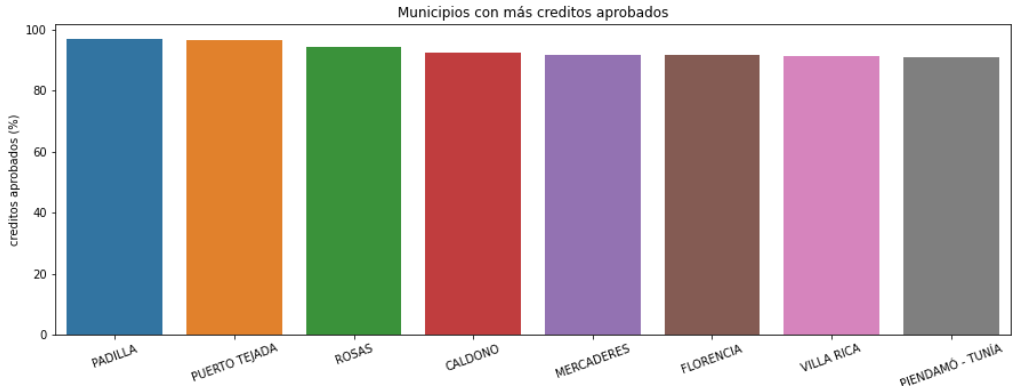


Figura 10. Municipios con más créditos aprobados para actividades agropecuarias en Cauca.

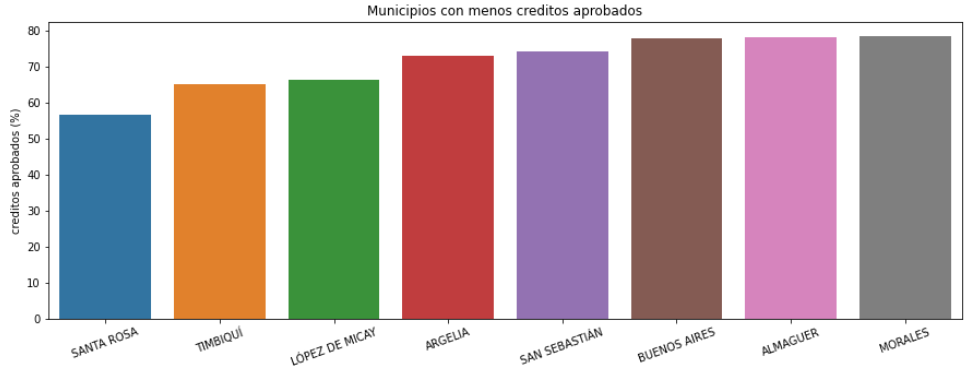


Figura 11. Municipios con menos créditos aprobados para actividades agropecuarias en Cauca.

3.4.2.2. Datos de inversión de los agricultores.

De los datos del censo nacional agropecuario los datos de inversión se dividieron en 2 grandes clases: Inversión en tecnología, donde se agrupó la inversión en maquinaria e inversión en nuevas técnicas de cultivo, y por otro lado la inversión en otros, que se refiere a inversiones varias como nuevos terrenos, ganado, entre otros. La distribución de estas clases se puede ver en la tabla 14 y 15 donde se vé como en el Cauca en promedio la inversión en tecnología es del 9.8% y los municipios en los que menos se invierte en tecnología son San Sebastián y Sucre, mientras que el que más invierte en tecnología es Guapi.

Municipality	Inversión tecnología (%)	Inversión otros (%)
GUAPI	52,9304029304029	47,0695970695971
TIMBIQUÍ	44,1489361702128	55,8510638297872
VILLA RICA	36	64
GUACHENÉ	28,1081081081081	71,8918918918919

PUERTO TEJADA	24,5762711864407	75,4237288135593
PADILLA	24,1976893453145	75,8023106546855
CORINTO	22,4264705882353	77,5735294117647
MIRANDA	21,5086646279307	78,4913353720693
LÓPEZ DE MICAY	17,0212765957447	82,9787234042553
TOTORÓ	8,69565217391304	91,304347826087
PATÍA	8,42105263157895	91,5789473684211
SANTANDER DE QUILICHAO	7,31707317073171	92,6829268292683
CALOTO	7,26392251815981	92,7360774818402
CAJIBÍO	7,07317073170732	92,9268292682927
POPAYÁN	6,59630606860158	93,4036939313984
TORIBÍO	6,36550308008214	93,6344969199179
PÁEZ	6,28571428571429	93,7142857142857
JAMBALÓ	5,52325581395349	94,4767441860465
SUÁREZ	5,09090909090909	94,9090909090909
EL TAMBO	4,73273942093541	95,2672605790646
PIENDAMÓ - TUNÍA	4,70914127423823	95,2908587257618
TIMBÍO	4,65416936005171	95,3458306399483
BOLÍVAR	4,3956043956044	95,6043956043956
SILVIA	4,39146800501882	95,6085319949812
CALDONO	4,38212094653812	95,6178790534619
LA SIERRA	3,9832285115304	96,0167714884696
BUENOS AIRES	3,97727272727273	96,0227272727273
MORALES	3,9485766758494	96,0514233241506
INZÁ	3,55781448538755	96,4421855146125
FLORENCIA	3,55329949238579	96,4467005076142
ALMAGUER	3,55329949238579	96,4467005076142
LA VEGA	3,14685314685315	96,8531468531469
SOTARÁ PAISPAMBA	2,97805642633229	97,0219435736677
ARGELIA	2,710027100271	97,289972899729
MERCADERES	2,64423076923077	97,3557692307692
ROSAS	2,63565891472868	97,3643410852713
PIAMONTE	2,5	97,5
BALBOA	2,4232633279483	97,5767366720517
PURACÉ	2,37154150197628	97,6284584980237
SANTA ROSA	1,58730158730159	98,4126984126984
SAN SEBASTIÁN	0	100
SUCRE	0	100

Tabla 14. Inversión en agricultura según censo nacional agropecuario en Cauca.

	promedio	desviación estándar	mínimo	máximo
Inversión tecnología (%)	9,80284308	11,93770429968	0	52,9304029304029
Inversión otros (%)	90,1971569	11,93770429968	47,0695970695971	100

Tabla 15. Estadísticas de Inversión en agricultura según censo nacional agropecuario en Cauca.

3.4.2.3. Datos del área de los agricultores.

Las áreas, al igual que las otras variables del Censo Nacional Agropecuario, se definieron con base en el porcentaje en vez de usar el valor total. La razón de usar el porcentaje es tener una composición más equilibrada entre los diferentes municipios.

Como resultado del análisis se obtiene la tabla 16 que contiene las áreas usadas para cada municipio y la tabla 17 que contiene las estadísticas de distribución de área en el Cauca que también se pueden ver reflejadas en la figura 12. En promedio en el Departamento del Cauca el área de cultivo que usan los agricultores es de 48.9%, aunque tiene una desviación estándar muy grande lo cual se evidencia con el valor mínimo (Santa Rosa) con 10.2% del área cultivada, ó el valor máximo (Puerto Tejada) con un 95,2% de tierra cultivada.

Situación similar pasa con las otras estadísticas como el área cultivada con promedio 15.9% y desviación estándar 14,3%, área de bosques con promedio 34,2% y desviación estándar 26,8% e infraestructura con un área promedio de 0,093% y una desviación estándar de 1,52%.

Lo anterior hace del Cauca una región con muchas variaciones entre municipios por lo cual si se desea hacer el análisis de un municipio en particular, este debe ser muy especializado y enfocado solo en los valores de ese municipio.

Municipality	área cultivada (%)	área no cultivada (%)	área bosques (%)	area infraestructura (%)
PUERTO TEJADA	95,2124487029666	0,144554225391936	0,683019351783472	3,95997771985803
VILLA RICA	92,7001329654401	3,15298097773123	0,0214426277632976	4,12544342906541
PADILLA	90,2031132088641	0,381027895383366	5,26398249628173	4,15187639947076
GUACHENÉ	89,2906344924503	2,57799765605738	1,71942470853171	6,41194314296061

TIMBÍO	82,9468482054508	10,7431392684784	5,72331194698129	0,586700579089463
FLORENCIA	81,4924948322496	12,2026355011038	6,2352288570538	0,0696408095928079
SANTANDER DE QUILICHAO	78,787495147976	6,82729213051154	10,5294315219605	3,85578119955196
PIENDAMÓ - TUNÍA	77,1892576092948	15,7365404430329	6,83817792553288	0,236024022139438
POPAYÁN	72,4674582347531	10,8023079724439	15,6851285442161	1,04510524858688
MIRANDA	71,5555369232305	1,22595191265209	23,986459835371	3,23205132874633
ROSAS	69,5553742853415	26,7063805064394	3,62215560303032	0,116089605188795
CALOTO	61,8264178723458	4,41829548796789	31,0668903245376	2,68839631514873
CORINTO	57,0046633618272	5,92129334668344	34,9701614004451	2,10388189104423
SOTARÁ PAISPAMBA	55,2123439011245	14,1559783088592	29,813479587817	0,818198202199297
TOTORÓ	54,2746649867361	6,22022175091291	39,4510646465534	0,0540486157976156
BALBOA	53,885036988697	22,2828856252257	23,7633024800891	0,0687749059881518
BUENOS AIRES	50,9971887101675	27,8379884940602	19,9235496559478	1,24127313982456
CAJIBÍO	50,1583796197383	31,0277935730383	17,7497461478476	1,06408065937567
CALDONO	48,6968553177364	19,4698114228118	31,6369836130591	0,196349646392744
BOLÍVAR	48,6784010602049	42,6117669216134	8,69146220979853	0,0183698083830959
SUCRE	48,5856891799176	47,0370842076185	4,37162949668253	0,0055971157813137
PATÍA	47,7388744016248	39,7097073529727	11,9922794127921	0,559138832610401
MERCADERES	47,2676129604035	41,9951846736453	10,6659200929153	0,0712822730358817
MORALES	45,82705149992	19,3477117949856	34,6195419265787	0,205694778515745
LA SIERRA	44,0253587023329	31,4201942695094	23,8527609303859	0,701686097771744
TORIBÍO	43,9313420208723	2,30409202973049	53,7296980347744	0,034867914622814
SUÁREZ	40,8663564978273	50,324326484562	8,54656145641014	0,262755561200523
JAMBALÓ	39,2072650621232	5,30556218507084	55,4525642418216	0,0346085109843133
EL TAMBO	36,4163032212725	18,679535495586	44,4532720648334	0,450889218308094
INZÁ	34,9939826060681	8,20926874510396	56,726281648095	0,0704670007328904
LA VEGA	33,8891529424404	44,3619860340148	21,6289810941118	0,119879929432962
SAN SEBASTIÁN	30,0776706856802	13,5930560526563	56,317094021889	0,0121792397744693
ALMAGUER	29,1625009319034	15,9658664739154	54,1121805413578	0,759452052823461
PURACÉ	29,1350626132412	4,75479805376644	65,9882665621313	0,121872770861123
PÁEZ	25,4243965705502	10,9294044695415	63,6241703297909	0,0220286301173406
TIMBIQUÍ	24,0048604966759	0,237584617571434	75,7517373928227	0,00581749292999524
ARGELIA	22,2439981083203	19,5068140962542	58,2114933381439	0,0376944572815298
PIAMONTE	18,2239116796444	11,4760842010033	70,2889011893033	0,0111029300489122

LÓPEZ DE MICAY	14,3343771213442	0,7911906690092	84,8331186900575	0,0413135195891698
SILVIA	14,2412175593487	2,65550268495226	83,0624711988972	0,0408085568018848
GUAPI	11,1328612946142	3,81027397295332	85,0436928360715	0,0131718963609579
SANTA ROSA	10,1937259079953	13,8770139286752	75,9275964810247	0,00166368230473413

Tabla 16. Áreas de agricultores según censo nacional agropecuario en Cauca.

	promedio	desviación estándar	mínimo	máximo
área cultivada (%)	48,925789191019	23,6770527042451	10,1937259079953	95,212448702966
área no cultivada (%)	15,911680985658	14,3074139783248	0,14455422539193	50,324326484562
área bosques (%)	34,233525654654	26,8011640714801	0,02144262776329	85,043692836071
área infraestructura (%)	0,9290041686669	1,52481847857698	0,00166368230473	6,4119431429606

Tabla 17. Estadísticas de áreas de agricultores según censo nacional agropecuario en Cauca.

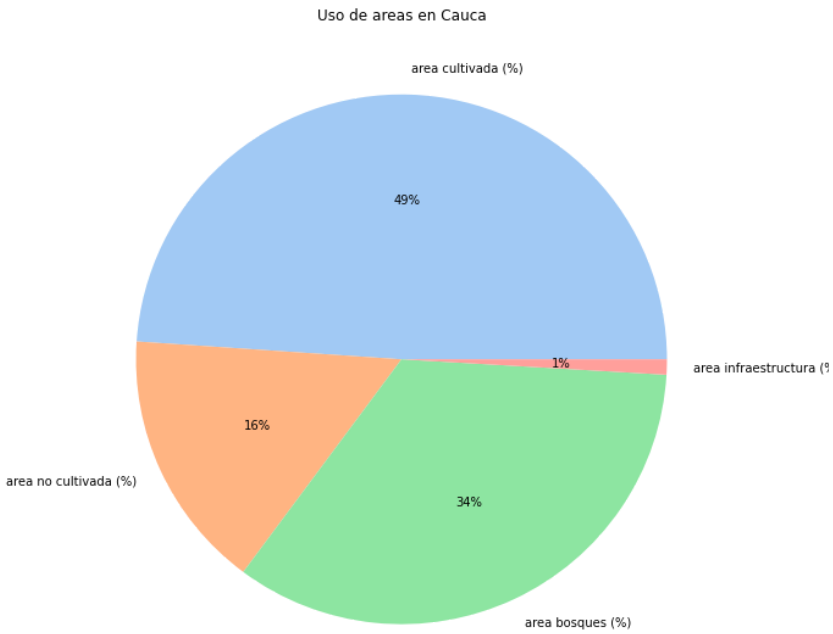


Figura 12. Uso de áreas por agricultores en el Cauca.

3.4.3. Exploración de los datos: ENSIN.

Los datos de ENSIN contienen muchas variables, aún después del filtrado y pre-procesamiento final, por esta razón en esta sección solo se explicarán las más representativas. También se debe recalcar que las estadísticas y gráficas presentadas en esta sección sólo estarán para los 9 municipios del Cauca incluidos en ENSIN.

3.4.3.1. Datos de seguridad alimentaria según ENSIN.

ENSIN ofrece un cálculo de la seguridad alimentaria en los hogares, sin embargo estos datos no pueden ser comparados con el GFSI porque se calculan por hogar con base en 4 preguntas. Esta encuesta da cuenta principalmente de la dimensión de acceso económico y físico a los alimentos de la FAO. Sin embargo nos permite obtener un porcentaje de la población que ha tenido problemas para acceder a comida.

De las tablas 18 y 19 se puede ver como en el departamento del Cauca en promedio el 75% de los hogares encuestados tienen seguridad alimentaria, con Santander y Timbiquí siendo los municipios con más hogares con seguridad alimentaria (100%) y Paez el municipio con menos hogares con seguridad alimentaria (32,7%). Además en el Cauca el 33.6% de los hogares encuestados han tenido que disminuir la calidad de alimentos siendo Santander y Timbiquí los que menos (0%) y San Sebastián el que más. Finalmente el 20,2% de los hogares debieron disminuir las porciones de alimentos en el Cauca con Santander, Timbiquí y Paez los que menos han tenido que recurrir a esto (0%) y Balboa el municipio que más (60,9%).

	Hogares con seguridad alimentaria (%)	disminuir calidad alimentos (%)	disminuir porciones alimentos (%)
SANTANDER DE QUILICHAO	100	0	0
TIMBIQUÍ	100	0	0
PATÍA	94,7772657450077	71,1384195255163	46,2877624167947
SAN SEBASTIÁN	87,9900787607154	76,1324572472912	46,7516644184326
BOLÍVAR	82,8852574240028	18,1338956246705	17,5364610788965
BALBOA	68,10063167066	63,3957743411021	60,9126551949466
POPAYÁN	68,0758017492711	22,1574344023324	9,91253644314869
TIMBÍO	40,0979233496539	40,0979233496539	0,590916765152794
PÁEZ	32,7172470131727	11,2937812723374	0

Tabla 18. Seguridad alimentaría en Cauca según ENSIN.

	promedio	desviación estándar	mínimo	máximo
Hogares con seguridad alimentaria (%)	74,9604673013871	24,9338564132779	32,71724701317	100
disminuir calidad alimentos (%)	33,5944095292116	30,1533038012872	0	76,13245724729
disminuir porciones alimentos (%)	20,2213329241524	24,4045101359689	0	60,91265519494

Tabla 19. Estadísticas seguridad alimentaría en Cauca según ENSIN.

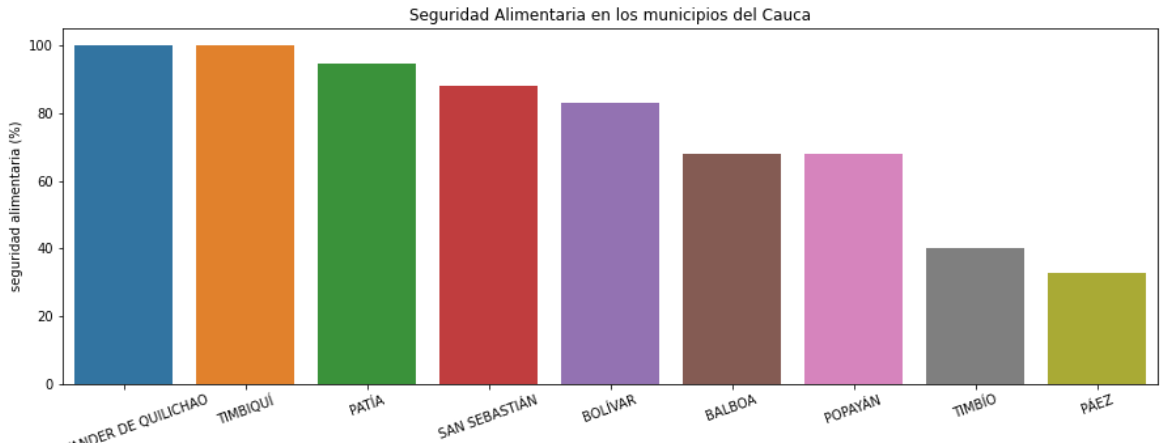


Figura 13. Porcentaje de seguridad alimentaria en los municipios del Cauca según ENSIN.

3.4.3.2. Datos socioeconómicos.

Los datos socioeconómicos de ENSIN se han englobado en 3 variables relacionadas a: si los ingresos de los hogares son suficientes para cubrir los gastos mínimos, si los hogares cuentan con acueducto y si los hogares cuentan con electricidad. En cuestión de ingresos como se ve en la figura 14 y tablas 20 y 21, en promedio en el Cauca el 50,2% de las viviendas dicen tener ingresos insuficientes y el municipio con mayor proporción de hogares con ingresos insuficientes es Timbiquí, mientras que el que menor porcentaje presenta es Santander.

Sobre el acceso a servicios públicos en el Cauca los hogares con acceso a electricidad son el 94%, con San Sebastián, Popayán, Balboa, Timbío y Santander con el máximo valor cubriendo el 100%, mientras que Timbiquí el que menor

porcentaje cubre con el 70,2%. Por otro lado el acceso a acueducto en el Cauca es de 87,66%, dónde el máximo de hogares lo cubre Popayán, Balboa, Bolívar, Timbío y Santander con el 100% de los hogares, y el menor porcentaje lo tiene nuevamente Timbiquí con el 4,2%.

	ingresos insuficientes (%)	hogares con energía (%)	hogares con acueducto (%)
TIMBIQUÍ	99,101315930245	70,2150422595485	4,21525623194608
SAN SEBASTIÁN	96,4492406770811	100	96,4492406770811
PATÍA	88,2744495647722	99,2831541218638	99,2831541218638
BALBOA	62,6769766935308	100	100
POPAYÁN	55,8309037900875	100	100
PÁEZ	30,695394669662	89,829470029613	88,9717144899418
BOLÍVAR	14,9358636443507	88,929889298893	100
TIMBÍO	4,08576734762789	100	100
SANTANDER DE QUILICHAO	0	100	100

Tabla 20. Datos socioeconómicos en los municipios del Cauca.

	promedio	desviación estándar	mínimo	máximo
ingresos insuficientes (%)	50,227768035	39,4301933057426	0	99,10131593024
hogares con energía (%)	94,250839523	10,1005506703079	70,215042259548	100
hogares con acueducto (%)	87,657707280	31,5009370367186	4,2152562319460	100

Tabla 21. Estadísticas de datos socioeconómicos en los municipios del Cauca.

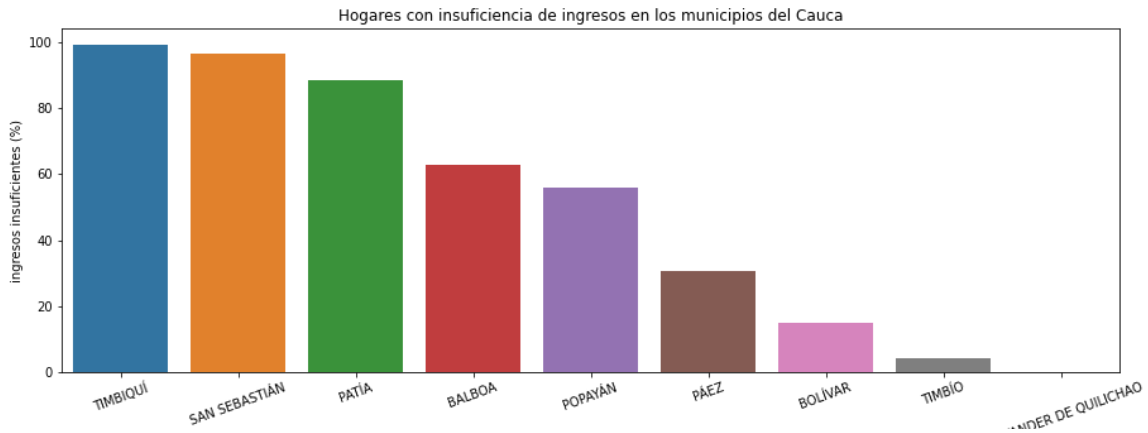


Figura 14. Porcentaje de hogares con ingresos insuficientes en los municipios del Cauca según ENSIN.

3.5. Verificación de la calidad de los datos

Este procedimiento consiste en evaluar la calidad de los datos obtenidos provenientes de las diferentes fuentes anteriormente mencionadas. Permite evaluar si los datos se encuentran completos, si todos los datos son correctos, contienen errores o datos nulos. En caso de que se presente algún inconveniente con los datos, se procede a realizar las acciones necesarias para garantizar un dataset completo con los menores errores posibles. Los datos se preparan en el siguiente capítulo (fase de preparación de los datos).

Debido a que los datos se obtuvieron desde diferentes fuentes abiertas, se identificaron problemas relacionados con la calidad de los datos. Algunas soluciones para estos problemas consisten en imputar datos, eliminar las variables con muchos datos faltantes, asignar valores a los datos vacíos. La explicación de los métodos anteriores se realiza en la siguiente fase del CRISP-DM (Capítulo 4, Preparación de los datos).

3.5.1. Valores nulos o vacíos

En los datasets de la encuesta ENSIN, censo agropecuario y registros de salud se encontraron valores nulos o vacíos, generalmente se representan con NaN. Estos errores son provenientes desde la fuente que origina los datos, y se debe principalmente a la cantidad de información que se recolecta en cada encuesta. Para no generar inconvenientes en la construcción del dataset, se tomaron varias opciones como solución para lidiar con los datos nulos o vacíos. Estas acciones se mencionan en el capítulo que se presenta a continuación y tiene relación con la preparación de los datos.

3.5.2. Valores faltantes

Debido a la alta complejidad para realizar encuestas y al arduo trabajo de recolección de información, hay algunos municipios que no son tomados en cuenta en la encuesta del ENSIN. Es decir, de los 42 municipios del departamento del Cauca, la encuesta solo se realizó en 9 municipios, por lo que afecta a la generalización de la información para este departamento.

3.5.3. Temporalidad de los datos

Es importante mencionar un problema que involucra a todos los datos, este consiste en la temporalidad, es decir, las fuentes de datos usadas para la construcción del dataset no presentan la misma temporalidad, como es el caso del dataset obtenido de la encuesta ENSIN, el cual corresponde al año 2015, mientras que el dataset del censo agropecuario corresponde al año 2014. También hay que considerar los datos usados para la construcción del dataset provenientes de fuentes abiertas diferentes a las mencionadas anteriormente, algunos datos se obtienen de reportes, censos, encuestas. Aunque estos datos no siempre son del mismo año, se seleccionaron los datos más cercanos a las fechas mencionadas con anterioridad (2014 - 2015).

3.6. Conclusiones del capítulo 3.

En este capítulo se abordó la exploración de los datos y se presentó una caracterización de la situación del departamento del Cauca con relación a aspectos nutricionales, agropecuarios, sociodemográficos y socioeconómicos, por lo cual se da por culminado el primer objetivo específico. En este punto se tiene un entendimiento general del problema y de las distintas fuentes de información disponibles. A partir de este proceso se puede reconocer qué variables y fuentes de datos pueden ser usadas para la creación del dataset, además de tener un entendimiento general de la situación del Cauca y de sus municipios en aspectos de seguridad alimentaria y nutricionales.

Con esta información en mente en el siguiente capítulo se abordará la construcción de un dataset en el cual se integren todas estas fuentes. El análisis realizado en este punto será muy importante en etapas de pre-procesamiento posteriores para entender qué datos pueden ser usados y qué procesos se deben hacer a los datos para su limpieza, procesos como imputación de datos o eliminación de datos nulos y/o anómalos.

El entender los datos además es vital, sobre todo en casos como este donde se trabajan con varios conjuntos de datos de diferentes fuentes y recolectados en diferentes temporalidades, ya que ayuda a entender cómo se debe hacer el proceso de alineación, de datos de forma que se conozcan las características espaciales y temporales de cada conjunto de datos y se pueda así buscar una variable en común como puede ser el código de municipio.

Capítulo 4

4. Preparación de los datos

4.1. Limpieza y pre-procesamiento de los datos

Una vez explorados los datos y los problemas existentes en los datasets seleccionados, se debe hacer un pre-procesamiento y una limpieza de datos con el fin de evitar errores futuros por valores nulos, diferencias de formato, o mal funcionamiento del modelo debido a valores anómalos. En este caso debido a las 2 modalidades de datos existentes (Imágenes satelitales y metadatos), se dividió esta sección en 2 partes. Primero se explicarán las imágenes satelitales, y luego los metadatos.

4.1.1. Preparación de los datos para imágenes satelitales

La preparación de las imágenes satelitales es un proceso muy largo que incluye desde la captura de imágenes en la región de interés, hasta procesamiento para aumento de resolución, eliminar impurezas de la imágen como nubes o sombras, y aplicar modelos de machine learning para la extracción de características. A continuación se explicará este proceso paso a paso.

4.1.1.1. Selección de la fuente de datos.

La selección de las imágenes es un proceso de vital importancia. Hay que tener en cuenta que GEE ofrece imágenes de todo el mundo casi en tiempo real, esto significa muchas fuentes diferentes de datos, de muchos tipos diferentes, formatos, resoluciones y temporalidades.

Para facilitar la visualización de las imágenes y la simplicidad del análisis, se buscaron satélites que ofrecieran imágenes RGB y que además fueran imágenes públicas y que posean cierta temporalidad para permitir hacer el análisis en el tiempo. Teniendo en cuenta las características mencionadas los satélites que cumplen con esto son Sentinel 2 y Landsat 8. De los satélites mencionados, Landsat 8, al ser una misión más vieja tiene menor resolución espacial (30 m/px) y menor tiempo de revisita (16 días), pero también al ser el satélite que tiene más tiempo es el que tiene las imágenes más acordes al análisis que se está realizando, esto teniendo en cuenta las fechas de recolección de los otros datos.

Así mismo las imágenes al tener más tiempo tienen un mejor pre-procesamiento por parte del proveedor de los datos.

4.1.1.2. Filtrado de imágenes.

Una vez localizada la fuente de las imágenes, se puede usar GEE para acceder a la colección de imágenes, pero se debe filtrar para evitar sobrecargar el almacenamiento. Debido a que las imágenes satelitales se pueden obtener desde la fecha en que el satélite comenzó a capturar imágenes, hasta el presente, y debido a que las imágenes se pueden obtener para cualquier localización del mundo. Se hace necesario filtrar los datos en tiempo y espacio para así quedarse solo con las imágenes en la región de interés y en el tiempo deseado para el análisis.

Para filtrar por localización se hace uso del archivo shapefile con los municipios de Colombia. Este archivo se puede leer como una instancia de la clase FeatureCollection de GEE. Como se mencionó anteriormente, estos archivos shapefile tienen en los metadatos información de cada municipio. En este caso se puede usar esa información para filtrar por código de municipio y obtener la región solo del municipio de interés. Para este ejemplo el municipio de interés fue Popayán.

Una vez se se tiene la región de interés se filtra por esta región y además por temporalidad de forma que se obtenga una colección de imágenes en un tiempo deseado. En este caso todas las imágenes de Landsat 8 de Popayán fueron tomadas entre abril del 2013 y diciembre de 2020. Un ejemplo de una de las imágenes resultantes se puede ver en la figura 15.

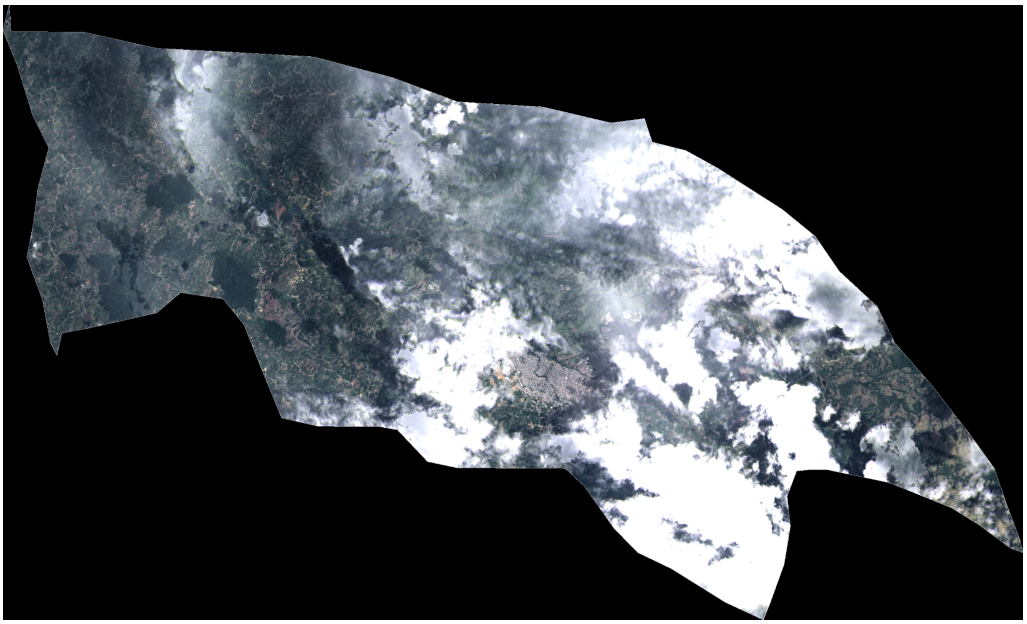


Figura 15. Imágen RGB de Popayán capturada por el satélite Landsat 8.

4.1.1.3. Eliminación de nubes y sombras.

Para eliminar la mayor cantidad posible de ruido de la imágen y evitar que estos factores afecten las predicciones de los modelos en el momento de la extracción de características, se generaron composiciones de imágenes a partir de múltiples imágenes. Para lograr hacer una composición de imágenes con la menor cantidad de nubes se tomaron imágenes en un intervalo de tres mesas al futuro y un mes en el pasado sobre el mes que se esté analizando. La razón de tomar un espectro tan amplio de imágenes es tener una cantidad lo suficientemente grande de muestras para cubrir menos pixeles sin sombras o nubes como se puede ver en la figura 16, pero tratar de no tener tantas imágenes en un rango muy grande de tiempo para no perder la temporalidad. Además, si bien idealmente Landsat 8 dice tener un periodo de revisita de 16 días, en algunos casos esto puede extenderse por problemas técnicos y se llegan a presentar casos de meses sin ninguna imágen.

Una vez seleccionado el intervalo de tiempo en el que se tomarán las imágenes para hacer la composición, se toman de estas imágenes los píxeles con la tasa más baja de nubes. De este grupo de píxeles filtrados, se toma el valor en el percentil 75. La razón de tomar el valor en el percentil 75 es que se han eliminado los píxeles con mayor cantidad de nubes densas los cuales son los valores más altos, cercanos al blanco, sin embargo los valores más bajos de los píxeles restantes serán los más oscuros cercanos al negro, es decir las sombras. Por lo tanto se desea tomar un valor alto teniendo en cuenta que se eliminaron los

píxeles con nubes, pero aún así hay píxeles con una nubosidad leve como neblina. Un ejemplo entre la diferencia de tomar el pixel en el percentil 50 y el percentil 75 se puede ver en la imágen 17.

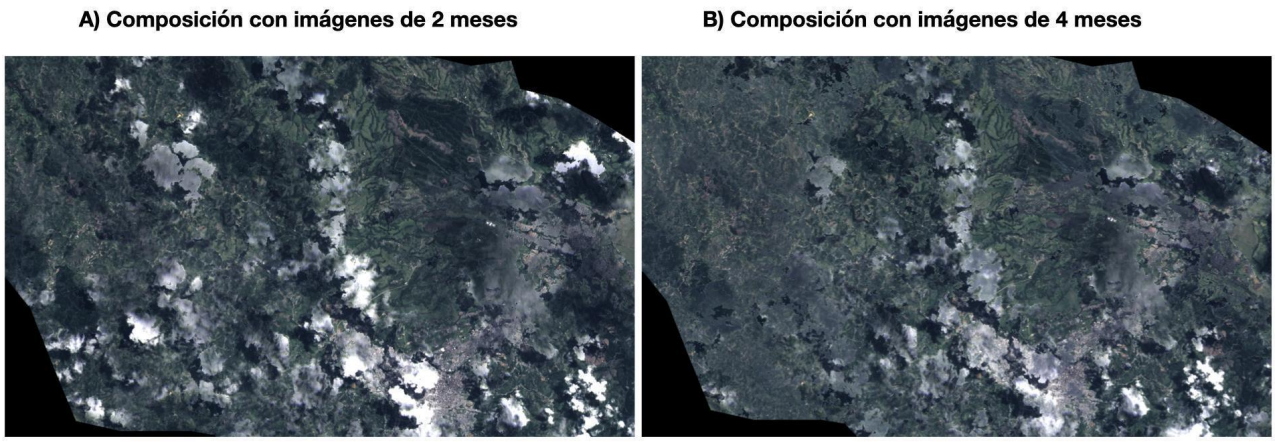


Figura 16. Composición RGB de Popayán usando imágenes de 2 meses y 4 meses.

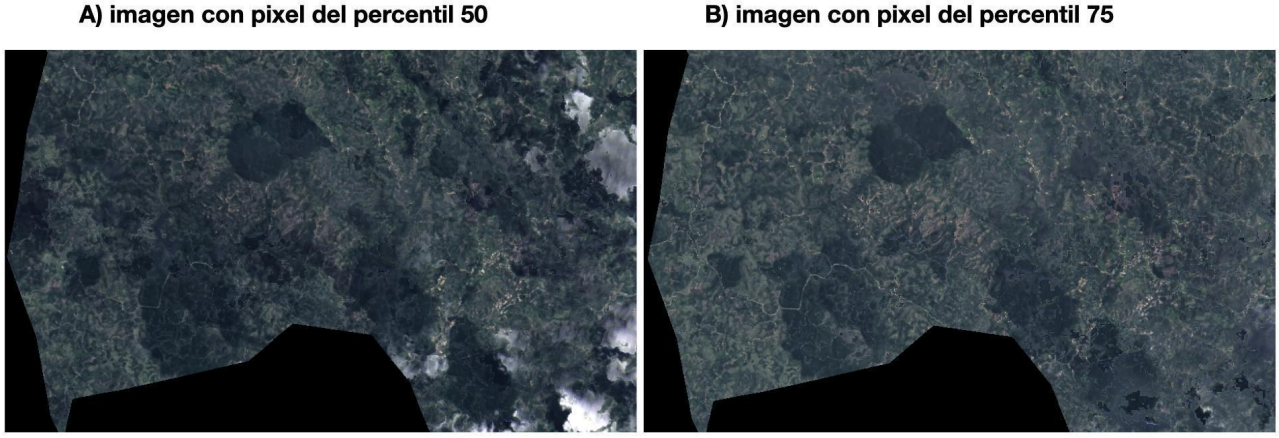


Figura 17. Composición RGB de Popayán usando el píxel con el valor del percentil 50 y percentil 75.

4.1.1.4. Aumento de resolución.

Se realizó un aumento de resolución con el fin de mejorar la calidad de las imágenes y poder distinguir las características presentes en las mismas. Para aumentar la resolución de la imágen RGB se hizo uso de la banda pancromática B8. Por defecto las imágenes extraídas del satélite Landsat 8 en las bandas B4, B3 Y B2 correspondientes a Rojo, Verde y Azul (RGB), se encuentran a una resolución de 30 m/px, sin embargo la banda B8 que se encuentra en una

frecuencia que abarca el mismo espectro electromagnético que las bandas RGB, se encuentra a una resolución de 15 m/px.

Para realizar el aumento de resolución se hizo una transformación de la escala RGB a la escala Matiz, Saturación, Valor (HSV) de las imágenes. Una vez hecha esta transformación se puede multiplicar los valores por la banda B8 (pancromática). El resultado es la imágen HSV pero ahora con una resolución de 15 m/px, finalmente se hace nuevamente una conversión de escala para regresar la imágen a RGB. Como resultado de este proceso se tendrá una imágen a 15 m/px. Un ejemplo de la imágen antes y después de este proceso se puede ver en la figura 18.

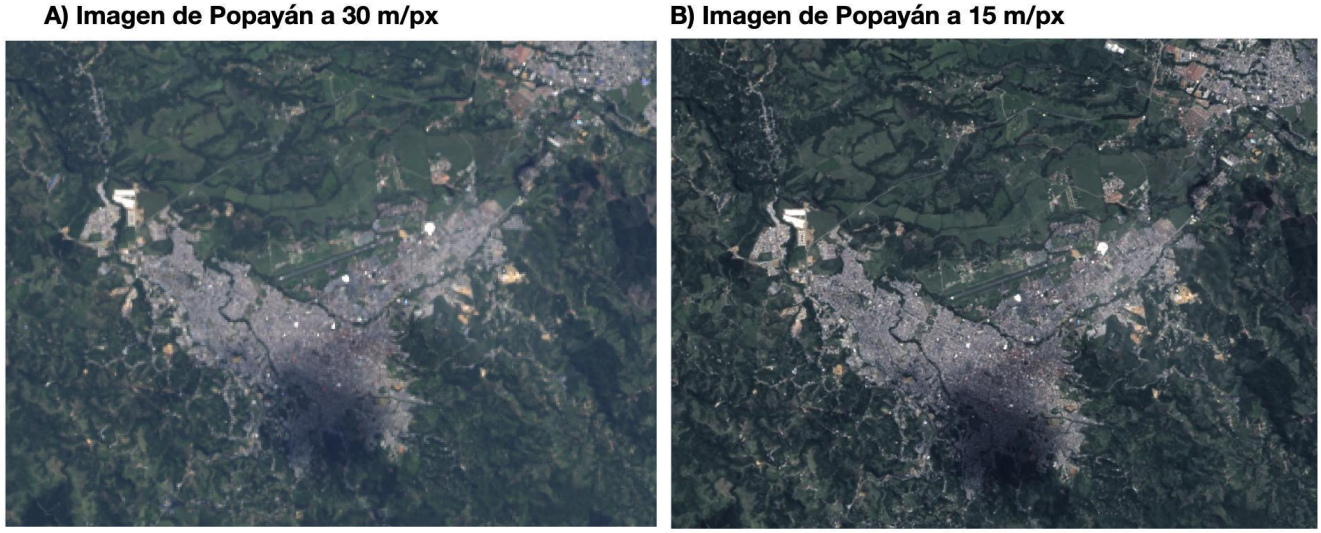


Figura 18. Imagen de Popayán antes y después de aplicar el aumento de resolución.

4.1.1.5. Extracción de Características.

La extracción de características de imágenes satelitales se presenta como un alternativa económica a la falta de datos, ya que son datos que se encuentran disponibles a nivel mundial de forma abierta y con una resolución tanto temporal como espacial lo suficientemente buenas.

Este ejemplo se hizo solo para las imágenes de Popayán por cuestiones de almacenamiento y recursos computacionales, sin embargo la metodología puede ser usada como referencia para los diferentes municipios del departamento o para otras regiones del mundo.

Para extraer las características se probaron 2 métodos diferentes. Un método basado en una clasificación para encontrar las características a nivel de píxel

usando segmentación semántica, y otro método que busca obtener una clasificación de características a nivel de escena. Los dos métodos se seleccionaron teniendo en cuenta que se debe generar un resultado lo suficientemente simple como para poder ser interpretado fácilmente. Los métodos usados se explicarán a continuación:

4.1.1.5.1. Extracción de Características usando clasificación multi-etiqueta.
Para este método se debe tener en cuenta que la clasificación multi-etiqueta es una tarea de aprendizaje supervisado, por lo cual se debe tener un dataset etiquetado con una cantidad suficiente de imágenes como para que un modelo de aprendizaje profundo pueda generalizar a nuevos datos. El clasificar imágenes satelitales para identificar características es una tarea compleja ya que requiere conocimientos del campo además de que es una tarea que requiere mucho tiempo. Sorprendentemente pese a la fuerza que está tomando el uso de imágenes satelitales en el campo de la investigación en aprendizaje automático, es muy difícil encontrar un conjunto de datos de imágenes satelitales para clasificación, ya que en su mayoría están diseñados para segmentación semántica de imágenes o detección de objetos. Afortunadamente se localizó un dataset de clasificación de imágenes satelitales disponible en la plataforma Kaggle. Este es el dataset "Planet: Understanding the Amazon from Space". El dataset “Planet” está compuesto por 40479 imágenes RGB de 256 x 256 píxeles y un archivo csv que mapea el nombre de la imágen con las características presentes en cada una de las imágenes. Una vez se tienen las imágenes y etiquetas descargadas, se procede a hacer la partición de los datos para entrenar el modelo y otra para evaluar el correcto funcionamiento del modelo de forma que pueda generalizar sobre nuevos datos. Los datos se dividen entonces aleatoriamente en 80% para entrenamiento y el 20% restante para evaluación.

Al ser modelos de deep learning, el entrenamiento se hace durante iteraciones ó épocas. Cada vez que se le pasa todo el dataset de entrenamiento por el modelo se cumple una época. Determinar un número de épocas no es una tarea fácil, ya que si se define un número de épocas muy bajo, entonces el modelo no ajustará los pesos lo suficientemente bien para realizar las predicciones correctamente. Por otro lado, un número de épocas muy alto producirá, además de un tiempo más grande de entrenamiento, que el modelo memorice los datos con los que se está entrenando y pierda la capacidad de generalizar a nuevos datos. Para evitar este problema y que los pesos del modelo se ajusten lo mejor posible a nuevos datos, el 10% del dataset de entrenamiento se usó para validación, de forma que en cada época se podrá verificar que tan bien está generando las predicciones el modelo sobre datos no entrenados. En caso de que las pérdidas en el conjunto de datos

de validación comiencen a subir, mientras que en el entrenamiento continúan bajando, significa que el modelo está comenzando a memorizar los datos. En ese caso se debe detener el entrenamiento y quedarse con los mejores pesos.

Una vez entendido el dataset, se puede pasar a los modelos. Los modelos entrenados son modelos de deep learning que fueron diseñados para realizar la tarea de clasificación de imágenes y cuyas arquitecturas se han modificado de las originales para adaptarlos al dataset usado y a la tarea. En general las arquitecturas entrenadas fueron: ResNet 50, VGG 19 y Vision Transformers. Los modelos ResNet 50 y VGG 19 fueron entrenados desde cero y usando también modelos con pesos pre-entrenados de ImageNet. Una vez entrenado cada modelo, se generaron predicciones sobre el conjunto de datos de evaluación y se calcularon las métricas de error, las cuales fueron: precision, recall, accuracy y el F1-Score. Los modelos evaluados y sus desempeños se pueden ver en la tabla 22-25. Una vez se tiene el mejor modelo se guardan sus pesos para su posterior uso.

Agricultura				
Modelo	**Accuracy**	**Precision**	**Recall**	**F1 score**
ResNet50	0.892652	0.844029	0.813934	0.828709
ResNet50 (ImageNet)	0.674294	0.431267	0.065574	0.113838
Vision Transformer	0.890919	0.858929	0.787234	0.821520
VGG16	0.875915	0.770992	0.869262	0.817184
VGG16 (ImageNet)	0.857218	0.778282	0.772541	0.775401

Tabla 22. Métricas calculadas sobre los datos de evaluación para los modelos entrenados para detección de agricultura.

Habitantes				
Modelo	**Accuracy**	**Precision**	**Recall**	**F1 score**
ResNet50	0.946914	0.829167	0.551247	0.662230
ResNet50 (ImageNet)	0.910303	0.673077	0.096953	0.169492

Vision Transformer	0.916362	0.910000	0.125864	0.221142
VGG16	0.937108	0.703204	0.577562	0.634221
VGG16 (ImageNet)	0.938415	0.688722	0.634349	0.660418

Tabla 23. Métricas calculadas sobre los datos de evaluación para los modelos entrenados para detección de habitantes.

Carreteras				
Modelo	**Accuracy**	**Precision**	**Recall**	**F1 score**
ResNet50	0.921810	0.891725	0.715705	0.794077
ResNet50 (ImageNet)	0.798902	0.638783	0.104283	0.179296
Vision Transformer	0.877610	0.789564	0.571517	0.663075
VGG16	0.916972	0.819372	0.777157	0.797706
VGG16 (ImageNet)	0.911480	0.818120	0.745500	0.780123

Tabla 24. Métricas calculadas sobre los datos de evaluación para los modelos entrenados para detección de carreteras.

Agua				
Modelo	**Accuracy**	**Precision**	**Recall**	**F1 score**
ResNet50	0.906642	0.884654	0.595544	0.711864
ResNet50 (ImageNet)	0.805178	0.423729	0.016880	0.032468
Vision Transformer	0.874478	0.856164	0.421727	0.565099
VGG16	0.887552	0.792648	0.567860	0.661684
VGG16 (ImageNet)	0.884806	0.816456	0.522620	0.637299

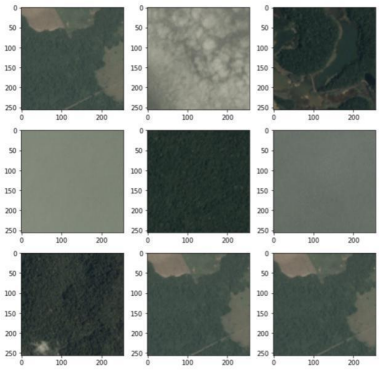
Tabla 25. Métricas calculadas sobre los datos de evaluación para los modelos entrenados para detección de agua.

La métrica usada para elegir el mejor modelo fue la Macro F1-Score, la cual es un promedio de las F1-Score para cada característica. En este caso el modelo que mejor desempeño tuvo es el ResNet 50 entrenado desde cero, sin embargo las características de la arquitectura original de ResNet 50 se modificaron para la tarea y el dataset.

La capa de entrada de la arquitectura original de ResNet 50 se modificó para tener las dimensiones de las imágenes del dataset, esto quiere decir que la entrada del modelo era de 256 x 256 x 3. Por otro lado, la salida del modelo también se editó para acoplarse a la tarea y el dataset. Normalmente el modelo ResNet 50, está diseñado para extraer características con las capas convolucionales y luego entrenar un modelo de red neuronal de clasificación. En este caso lo que se hizo fue entrenar el modelo que hiciera una clasificación multi-etiqueta, por esta razón las neuronas a la salida se definieron en base a el número de características a extraer, debido a que puede que haya más de una característica o ninguna en la imágen, la función de activación usada en estas neuronas fue softmax.

Una vez definida la arquitectura del modelo, se procede a diseñar el “dataloader”, el cual será el encargado de cargar los datos al modelo usando pequeños grupos de datos (batches), esto evita que se sobrecargue la memoria, en este caso GPU. Para este caso se cargaron los datos en batches de 32, lo cual significa que en cada batch en el entrenamiento el modelo tendría como entrada 32 imágenes RGB de 256 x 256 píxeles (un ejemplo de 9 de estas imágenes se puede ver en la parte izquierda de la figura 19) y las características a predecir en la salida del modelo, se cargaron en una matriz de 32 filas (por el número de imágenes en un batch) x 4 columnas (definidas por el número de características), esta matriz indicaba si había presencia (1) o no (0) de alguna de las características en la imágen. Un ejemplo de esta matriz para 5 imágenes se puede ver en la parte derecha de la figura 19 (en el ejemplo real la matriz tendría 32 imágenes en vez de 5).

- Entrada:
  - Imágenes : 256 x 256 x 3



Ejemplos de Imágenes aleatorias del dataset “Planet”

- Salida:
  - Agricultura
  - Habitantes
  - Carreteras
  - Agua

Image_name	Agriculture	Habitation	Road	Water
train_0.jpg	0	0	0	0
train_1.jpg	1	0	0	1
train_2.jpg	0	0	0	0
train_3.jpg	0	0	0	0
train_4.jpg	1	1	1	0

Ejemplos de matriz de características para algunas imágenes del dataset “Planet”

Figura 19. Ejemplos de algunas muestras de las entradas y salidas del modelo de machine learning.

Adicionalmente para evitar el sobreentrenamiento y que el modelo pudiese luego generalizar a imágenes obtenidas de otro satélite en una localización diferente, se añadió un aumento de datos (data augmentation). En el aumento de datos se realizaron las siguientes modificaciones a las imágenes aleatoriamente:

- Volteado aleatorio de la imágen
- Corrimiento de la imágen en cualquier dirección entre 0% - 15%
- Zoom entre 0% - 40%
- Rotación entre 0% - 30%

Con los datos listos para entrenar el modelo y el modelos se procede a el entrenamiento. Para el entrenamiento se usó como función de pérdida la entropía cruzada binaria (binary cross entropy) debido a que es una clasificación multi etiqueta, además como método para optimizar los pesos se usó el optimizador Adam con tasa de aprendizaje de 0,001, beta 1 de 0,9, beta 2 de 0,999 y épsilon de 1e-07.

Como se mencionó anteriormente, el número de épocas es un valor complicado de definir, por lo tanto se usó un método de parado temprano (early stop). A medida que transcurría una época del entrenamiento, se calculan las pérdidas tanto en entrenamiento como en validación y se comparan con las de la época anterior, de esta manera se tendría una idea de si las pérdidas en el conjunto de datos de validación estarían aumentando o disminuyendo. Finalmente en caso de

que las pérdidas de validación aumentaron durante 5 épocas seguidas con una diferencia superior a 1e-3, se detendría el entrenamiento y se restaurarán los pesos a los mejores encontrados.

Una vez finalizado el entrenamiento, se guardan los pesos óptimos de los modelos y se realizan las predicciones para el departamento del Cauca, en este caso el municipio de Popayán. Para realizar las predicciones de cada municipio se debió primero tomar una imágen de popayán y realizar el respectivo pre-procesamiento para que la imágen tuviera el formato de los datos de entrada del modelo. Para esto se dividió la imágen en parches de 256 x 256 como se observa en la figura 20. Como se ve hay parches que son totalmente negros o en otros casos pueden tener muchas nubes, estos parches se eliminaron contando el numero de valores en 0 (negro) en la imágen y en blanco (255). Una vez se tiene el número de píxeles blancos y negro se suma y se verifica que estos sean el 30% de la imágen o menos, de esta forma se asegura que la imágen tenga 70% de la información como mínimo.

Figura 20. Imagen de Popayán en parches.

Con los parches de imágenes de 256 x 256 ya se puede extraer información de las imágenes. Para extraer información de las imágenes se carga el modelo con los pesos pre-entrenados y se pasa cada uno de los parches. El resultado de las predicciones para cada una de las características de cada uno de los parches se puede guardar en un arreglo y sumar. Al dividir los valores en ese vector sobre el

número de parches se puede tener un indicador de en cuántas de las imágenes del municipio hay agua, carreteras, población o agricultura. Un ejemplo de predicciones se ve en la figura 21.

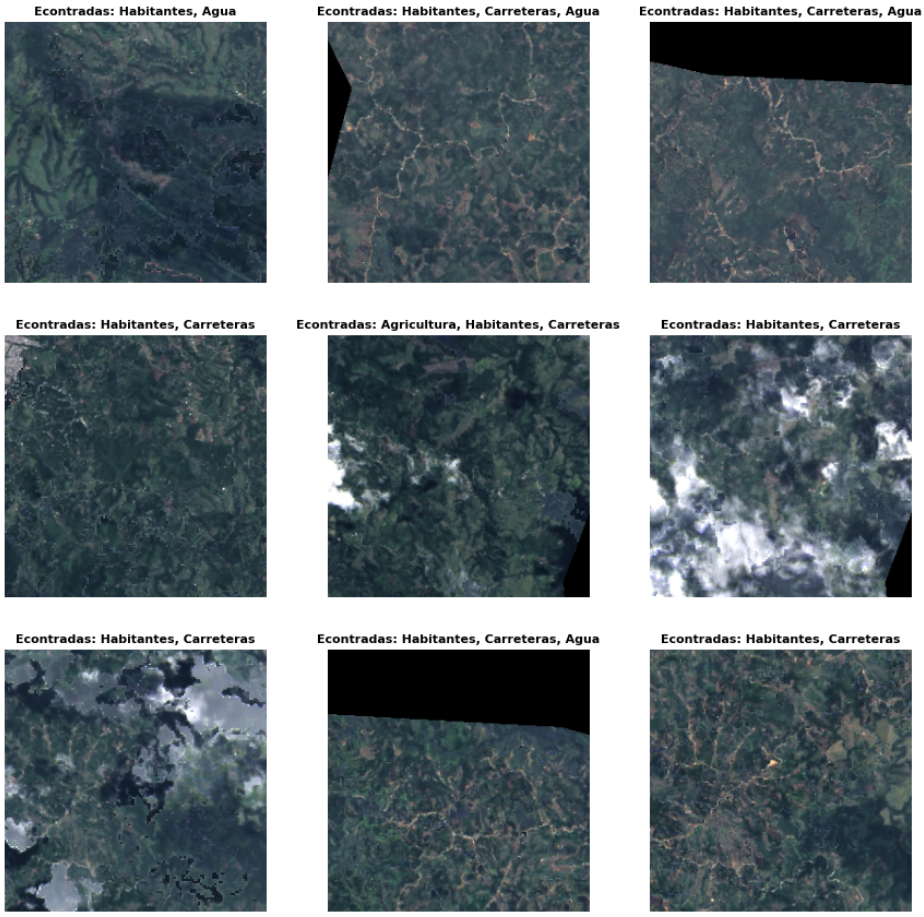


Figura 21. Ejemplos de predicciones sobre parches.

4.1.1.5.2. Extracción de Características usando segmentación semántica no supervisada.

Para entender este método se parte de la premisa de que las imágenes satelitales se pueden interpretar como un conjunto píxeles y cada píxel tiene un valor dependiendo de la presencia de una característica. De esta forma un pixel azul puede representar agua, un pixel verde puede representar vegetación, uno gris ciudad, etc. De forma que una imágen satelital es un conjunto de datos que expresan una característica de un territorio en particular. Entendiendo esto, se puede forzar una imágen mediante un algoritmo no supervisado a asignar una clase a cada píxel, de forma que se pueda simplificar y entender mejor las características de la imágen.

Los métodos más comunes de segmentación no supervisada en imágenes son 2: Gaussian mixture y k-means. Los dos métodos se testean en este trabajo haciendo uso de 3 bandas (bandas 2, 3 y 4, para imágen RGB) y 6 bandas (bandas 1-6) para comparar los resultados.

La metodología a seguir en ambos casos es igual. Primero se toma la imágen y se convierte en un arreglo de una dimensión de forma que la librería usada (sklearn) lo pueda interpretar, luego se define el número de clusters a generar de 5, de forma que se pueda mantener cierta simplicidad a la hora de generar los mapas de segmentación y se puedan comparar mejor las características encontradas. Los resultados se pueden ver a continuación.

- **Gaussian Mixture:** Las imágenes segmentadas haciendo uso de Gaussian mixture se pueden ver en la figura 22 para 3 bandas, y en la figura 23 para 6 bandas. En la figura 22, se ve como la segmentación basada en Gaussian mixture con 3 bandas en comparación a la imágen original, si bien puede detectar bien las ciudades, no se ve que distinga otras características como ríos o zonas de bosques y pese a tener 5 clusters las zonas no están muy bien definidas habiendo clasificado la mayoría de la imágen en un cluster posiblemente relacionado al color verde. Por otro lado la imágen de la figura 23 que usó Gaussian mixture con 6 bandas si puede distinguir otras clases en los clusters y de una forma más balanceada. En la figura 23 se ve como las zonas de ciudad han sido detectadas y se ven en una tonalidad verde, mientras que por otro lado en una tonalidad azúl más claro se detectan zonas de bosque y en un azul un poco más oscuro zonas de ríos, además las zonas verdes se puede ver otras zonas de vegetación como césped o cultivos.

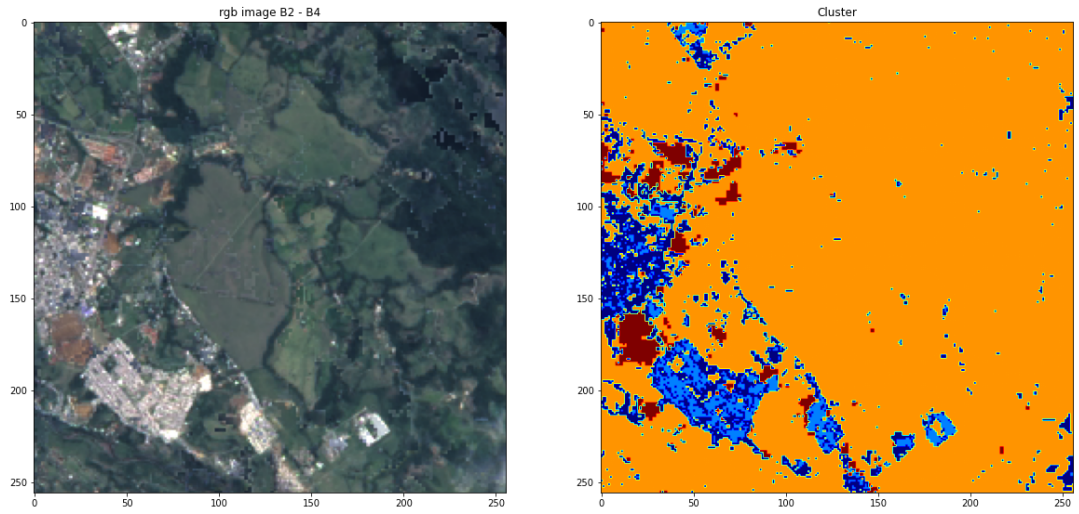


Figura 22. Ejemplos de segmentación no supervisada con imágen RGB usando Gaussian mixture.

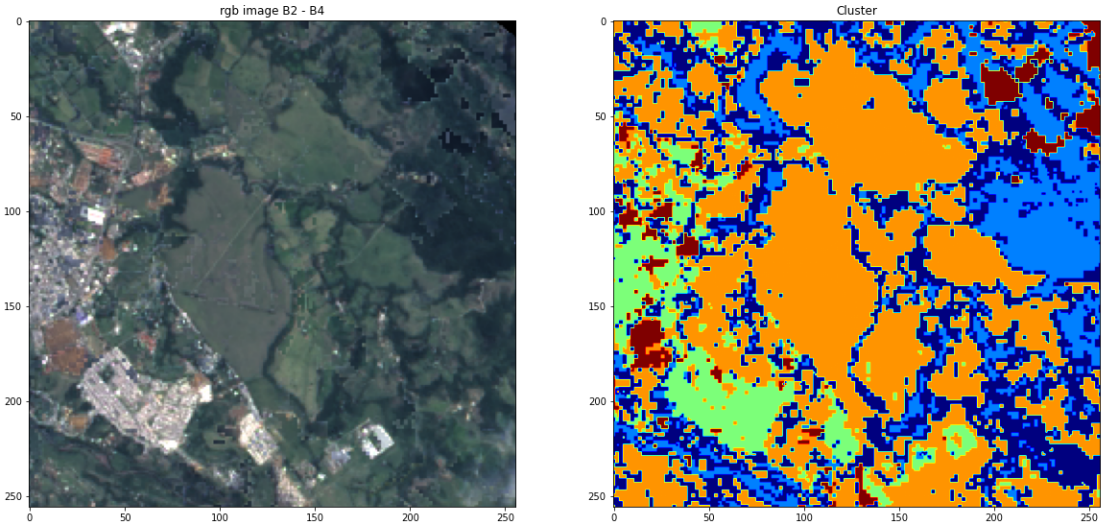


Figura 23. Ejemplos de segmentación no supervisada con imágen con 6 bandas usando Gaussian mixture.

- **k-Means:** En el caso de k-means los resultados son mejores a primera vista si se compara directamente la imágen segmentada RGB usando K-means (figura 24) con la imágen segmentada usando Gaussian mixture, esto debido a que en la imágen RGB de k-means se puede ver una mejor distribución de las clases, sin embargo se ve que en este caso la clase correspondiente la la ciudad (en color azul claro) no se muestra tan clara en

la imágen segmentada, por otro lado la clase que muestra los ríos y zonas boscosas no es tan precisa en comparación a la imágen usando Gaussian Mixture con 6 bandas, por lo tanto se procederá a analizar la imágen segmentada usando k-means con 6 bandas (figura 25). En la imagen segmentada con k-means haciendo uso de 6 bandas se puede distinguir mejor las zonas de ciudad en color verde, de igual forma otras características como ríos o zonas con árboles se marcan con un tono azul claro, mientras que las zonas de pastos o cultivos se marcan con un azúl más oscuro.

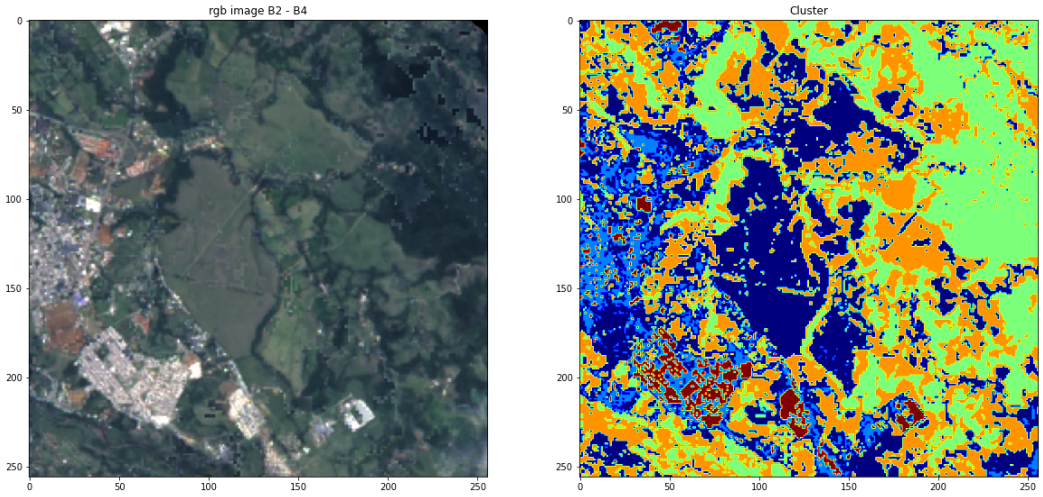


Figura 24. Ejemplos de segmentación no supervisada con imágen RGB usando k-means.

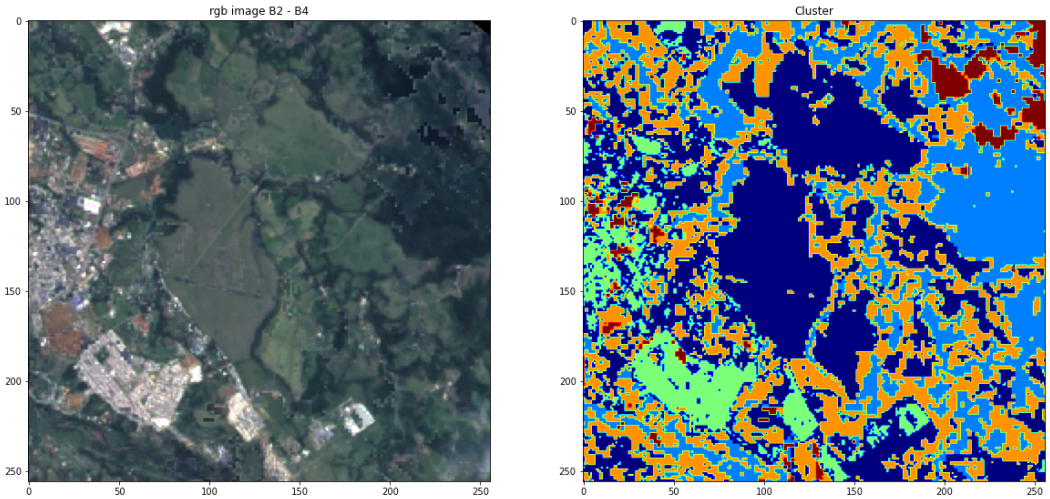


Figura 25. Ejemplos de segmentación no supervisada con imágen con 6 bandas usando k-means.

De los anteriores experimentos se tiene que el uso de algoritmos no supervisados para generar mapas de segmentación de imágenes satelitales puede ser un método efectivo para extraer características tanto con Gaussian Mixture como con K-means, sin embargo se debe también tener en cuenta que para el caso de imágenes satelitales el uso de 3 bandas no es tan efectivo como el uso de imágenes con más bandas que pueden tener más información. También es importante tener en cuenta que las imágenes satelitales son imágenes que pueden tener mucho ruido como nubes, sombras o problemas de calibración de imágen, que se deben tratar de mitigar lo mejor posible debido a que afectarían de gran manera este tipo de modelos confundiendo sombras con bosques o zonas de agua, o nubes con ciudades.

4.1.2. Preparación de los datos para metadatos

Al igual que para las imágenes satelitales, el procedimiento para la preparación de los metadatos es muy importante, debido a que es necesario tener los datos ordenados, limpios, con la menor cantidad de errores. Debido a lo anterior, en esta fase se desarrollaron procedimientos de selección de datos, cambios de formatos de las variables, creación de nuevas variables y unión de los datos en un solo dataset.

4.1.2.1. Selección de variables.

La selección de variables es importante para reducir la dimensionalidad de los datos, filtrar variables innecesarias y así disminuir el tamaño del dataset, hacerlo más entendible y facilitar el trabajo de pre-procesamiento. En este caso para la selección de variables, se tiene en cuenta principalmente todos los datos que estén relacionados con alguna de las cuatro dimensiones que dicta la FAO para garantizar la seguridad alimentaria y la variable código de municipio, la cual será usada para la integración de las diferentes fuentes de datos en un único dataset. A continuación se explica el procedimiento de selección de variables para el censo agropecuario, encuesta ENSIN, registros de salud, y datos meteorológicos.

4.1.2.1.1. Selección de variables: Censo Nacional Agropecuario y ENSIN.

La selección de variables del Censo Nacional Agropecuario y de la encuesta ENSIN se facilitó mediante el uso de los diccionarios de datos provistos por el DANE, y por ENSIN. Mediante los diccionarios, se seleccionaron las variables que tuvieran relación con las 4 dimensiones de la FAO, de esta forma se redujo el

número de variables disponibles facilitando el pre-procesamiento de datos, asimismo se seleccionó la variable de código de municipio, con el fin de facilitar la agrupación más adelante.

Para seleccionar las variables se tuvo en cuenta que la información que almacenaban fuera numérica o respuestas binarias (Si/ No), además de variables fácilmente agrupables, y variables cuyo formato se pudiera cambiar sin complicaciones, es decir variables del estilo de preguntas abiertas fueron eliminadas. Lo anterior se realizó con el fin de facilitar el procesamiento y la limpieza de datos.

Otras variables seleccionadas de la encuesta ENSIN fueron las “LLAVE_PERSONA” y “LLAVE_HOGAR”, esto debido a que ENSIN se encuentra dividido en múltiples conjuntos de datos y estas variables permiten relacionar los diferentes archivos o tablas, de forma que se puedan usar estas variables en un futuro para agrupar las tablas de ENSIN en una única.

Después de tener los datos procesados y limpios, se realizó otra selección de variables, pero enfocada principalmente en el índice GFSI. Haciendo uso de los indicadores que define este organismo y la relación con las variables del censo nacional agropecuario y la encuesta ENSIN se filtraron solo las variables que tuvieran relación con el GFSI.

4.1.2.1.2. Selección de variables: Registros de Salud.

La información obtenida en los registros de SIVIGILA tiene que ver con registros de salud de pacientes. En este caso, la información relevante corresponde a los registros de desnutrición aguda en niños menores de 5 años, la mortalidad por desnutrición y el código de municipio. La información restante es irrelevante para este estudio, ya que no tiene relación con las dimensiones definidas por la FAO ni por el índice GFSI. En el caso de SIVIGILA los datos son datos semana a semana y corresponden a valores enteros (número de casos registrados en la semana) almacenados en archivos de excel diferentes, un archivo de excel se publica cada año, por lo que este proceso de selección de variables se hizo para todos los archivos.

4.1.2.1.3. Selección de variables: Datos Metereológicos.

La selección de datos meteorológicos se llevó a cabo a partir de archivos de imágenes que contenían valores por píxel correspondientes a temperatura o precipitación de un mes específico en todo el mundo. Para extraer los valores de cada municipio en Colombia, se hizo uso de un archivo csv que contenía los valores de código de municipio, junto con la posición geográfica (latitud y longitud)

de cada uno. Estas variables de precipitación y temperatura se relacionan con factores que pueden afectar la seguridad alimentaria, es decir, están relacionadas con alguna dimensión definida por la FAO como la disponibilidad de los alimentos y la estabilidad en el tiempo de las dimensiones, así mismo tiene concordancia con el GFSI.

Como resultado se generó un dataset con el código de municipio y la precipitación y temperatura promedio de cada municipio más a mes.

4.1.2.2. Limpieza de datos.

El procedimiento de limpieza se realiza para depurar o eliminar datos erróneos del dataset y así garantizar un dataset de calidad para las siguientes fases. En esta acción se identifican datos faltantes, duplicados, erróneos, de formatos diferentes. Este procedimiento se realizó para las variables seleccionadas de la encuesta ENSIN, el censo nacional agropecuario y los registros de salud de SIVIGILA.

4.1.2.2.1. Limpieza de datos: ENSIN.

En el caso de la encuesta ENSIN, con las variables seleccionadas en la fase anterior se verificó que todos los datos correspondiesen con a un valor numérico o de caracteres binarios (Por ejemplo si y no). Si las variables no estaban en este formato, se procedía a cambiar el formato a alguno de los anteriores, en caso de no ser posible la variable se descarta. Se asignó a las variables categóricas binarias valores numéricos para facilitar el análisis, es decir, a las variables que almacenan valores binarios Sí o No se asignaron los números 1 y 0 respectivamente. Dando como resultado un conjunto de datos con valores sólo numéricos.

Seguidamente se verificó los datos nulos o vacíos por cada variable. En este caso se procedió de tres formas dependiendo de la importancia de la variable.

- Eliminar la variable: Se eliminarán las variables que presentan una cantidad considerablemente grande de datos faltantes, esto debido a que la información que contiene puede no ser de suma importancia o los datos no son los suficientes para realizar otro procedimiento.
- Imputación de datos usando el promedio: Si la cantidad de datos faltantes no es tan grande se puede proceder a imputar datos usando el promedio de datos disponibles para reemplazar los valores faltantes. Este método es útil sobre todo en casos de valores numéricos no binarios. En este caso fue de utilidad en variables como FERRITINA, VITAMINA A, ZINC.
- Imputación de datos con ceros: La tercera se aplica para las variables con valores binarios (0/1), en este caso se reemplazaron los valores vacíos con

0 ya que muchas veces lo que se hizo fue tomar en cuenta los valores positivos (1) con respecto al total, de esta forma los valores imputados no tendrían contribución en el valor final.

Finalmente con un conjunto de formularios con las variables filtradas, en el formato deseado y sin valores nulos, se procede a la agrupación de formularios de la encuesta ENSIN en un único dataset. Haciendo uso de la variable "LLAVE_HOGAR" como variable común entre algunos formularios (SA_1, SA_2, SA_3, PTS, PTS_2) se hizo la integración de estos en uno solo y haciendo uso de la variable común "LLAVE_PERSONA", se integraron otros formularios (VITAMINAS y PISNSP). Finalmente se tiene un formulario (R24) el cual se usa para integrar todos los formularios de ENSIN en unos solo. Como resultado se tiene un dataset de ENSIN integrado con las variables que pueden ser usadas para el cálculo del GFSI y la variable código de municipio.

4.1.2.2.2. Limpieza de datos: Censo Nacional Agropecuario.

El procedimiento de limpieza de datos del censo nacional agropecuario es similar al que se aplicó para la encuesta ENSIN. Primero se verificó el formato de los datos para que fueran de carácter binario o valores numéricos y se asignaron valores numéricos (1/0) a los datos de caracteres binarios es decir Si y No, de esta forma se tendría un dataset totalmente con valores numéricos. Se verificaron la cantidad de datos nulos por variables y se siguió el procedimiento mencionado anteriormente en la encuesta de ENSIN para valores faltantes con la excepción de que no se realizó imputación de datos usando el promedio para ninguna variable del censo nacional agropecuario.

Debido a que habían variables que podían ser agrupadas en una sola variable y con el objetivo de simplificar el conjunto de datos, variables tenían relación entre sí se agruparon formando variables como por ejemplo crédito_bancario, inversion_otros, agua_fuente_natural, los cuales están compuestos por la suma de varias variables más específicas del dataset original.

4.1.2.2.3. Limpieza de datos: Registros de Salud.

Los registros de salud proporcionados por SIVIGILA se encuentran en archivos excel de los casos durante un año en concreto. Como el objetivo final es obtener un solo archivo que contenga los casos disponibles juntos, se debe concatenar los archivos de cada año. Para concatenar los archivos se debe mantener todos en un mismo formato, el problema en este caso es que no todos los archivos tenían el mismo formato ya que habían algunos cambios año a año, por lo cual se debió tomar las columnas en común entre datasets y asegurarse de mantener un mismo formato en cuanto a nombres de las variables y el orden. Con los archivos de cada

año con el mismo formato y orden se procedería a concatenarlos en orden cronológico, es decir, desde el año más viejo hasta el más reciente.

Entre los años había municipios que no tenían casos o que en alguna semana en concreto no tenían casos, por lo cual al juntar los archivos se generarían valores nulos, los cuales se reemplazaron por 0 casos. Finalmente se puso el dataset resultante en un formato que fuera una fila por cada municipio usando el código de municipio como referencia, y una columna por cada semana con el número de casos correspondientes.

4.2. Integración de los datasets en un único dataset

Para la integración de los datos se debe pensar en la característica común para realizar la alineación de los conjuntos de datos, en este caso una alineación espacial, es decir, por municipio. La integración de los datos de la encuesta ENSIN, el dataset de censo nacional agropecuario, dataset de registros de salud y dataset de datos metereológicos, se realizó mediante la agrupación de la variable común código de municipio, presente en cada dataset.

El dataset resultante contiene 926 columnas que representan diferentes variables y características multidimensionales, junto con 42 filas representan cada municipio del departamento del Cauca, aunque en solo 9 filas, que correspondientes a los municipios de Popayán, Balboa, Bolívar, Paez, Patía, San Sebastián, Santander de Quilichao, Timbío, y Timbiquí, se tiene los valores de la encuest ENSIN.

Es importante mencionar que no todas las columnas son usadas para el cálculo del GFSI, pero estas pueden ser utilizadas para investigaciones futuras. Por último, el dataset obtenido a partir de la recolección, preparación y unión de los datos se encuentra en el **Anexo B**.

4.3. Conclusiones del capítulo 4.

En esta sección se da por concluida la creación de un dataset multidimensional del Cauca que integre las principales variables que influyen en la seguridad alimentaria con fuentes provenientes de repositorios abiertos, dando por concluido el segundo objetivo específico. Como resultado de este capítulo se tienen 2 conjuntos de datos, un conjunto de datos con la respectiva limpieza realizada con 926 columnas y 9 filas que representan 9 municipios del Cauca de los cuales se disponía de todos los datos, y por otro lado se tiene un conjunto de datos con 926 columnas y 42 filas, el cual representa las mismas variables para los 42 municipios del Cauca. Sin embargo, en el segundo conjunto de datos hay que tener en cuenta que se tienen algunos valores vacíos en las columnas correspondientes a los

valores extraídos de ENSIN que no están disponibles para todos los municipios del Cauca.

Además en este capítulo se exploró la integración de imágenes satelitales como una alternativa para la falta de información. En este caso se vieron resultados interesantes en los dos métodos evaluados. En el caso de la segmentación semántica, si bien se tiene más granularidad en los resultados, hay que tener en cuenta también que el método usado fue no supervisado, por lo que se deben interpretar los resultados para entender a qué clase pertenece cada conjunto de datos. En el caso de la clasificación de las imágenes satelitales usando el dataset “Planet”, los resultados para las características extraídas siguen siendo muy buenos, se ve que se pueden extraer efectivamente características de las imágenes satelitales, además de que los modelos usados y sus respectivos pesos se pueden guardar para ser usados en un futuro.

En el siguiente capítulo se hará uso del conjunto de datos generado en esta sección para realizar el cálculo del GFSI para el departamento del Cauca haciendo uso de modelos de machine learning.

Capítulo 5

Modelado y Evaluación

5.1. Adquisición de datos de entrenamiento y test para el modelo encargado del cálculo del GFSI.

Debido a que el GFSI es realizado a escala de país, no existe un índice de seguridad alimentaria que cumpla con todas las dimensiones dispuestas por la FAO para el departamento del Cauca. Con el fin de poder realizar el cálculo del GFSI en el Cauca, se decide entrenar un modelo de machine learning que aprenda los pesos de cada variable y puede dar como resultado el índice GFSI. Para el entrenamiento del modelo, es necesario utilizar un conjunto de datos que ya tenga el índice GFSI calculado así como sus características, el cual no existe a nivel del Cauca. La solución que se da, es usar los datos del GFSI publicados por la EIU en el 2020 a escala de país para entrenar el modelo. El archivo que contiene los datos se llama 2020_GFSI_Data_Table_Data.xlsx, disponible en el

enlace: https://github.com/dsrestrepo/SeguridadAlimentaria/blob/main/FoodSecurityIndex/Data_Countries/2020_GFSI_Data_Table_Data.xlsx.

El dataset del GFSI a nivel de países se compone de 60 columnas (59 variables que son las utilizadas para el cálculo del GFSI para cada país y el valor del GFSI calculado) y 113 filas, cada fila representando a un país. Es importante mencionar que estos datos no están normalizados, es decir, se debe hacer un pre-procesamiento para cada variable y luego se calculan los índices correspondientes.

Debido a lo anterior se realizó un proceso de normalización para cada variable, según lo indicado por el GFSI, con el fin de reajustar las variables a un rango de 0 a 100. El procedimiento de normalización se realizó de diferentes formas.

1. Normalización usando ecuación lineal: El primer método de normalización fue calculando una ecuación lineal mediante los valores que brinda el GFSI en el archivo Final_GFSI_model_2020.xlsm en la sección SCORE CALCULATION, disponible en el enlace: (https://impact.economist.com/sustainability/project/food-security-index/Downloads). Usando esta ecuación lineal se hace la proyección de la variable en un rango de 0 a 100. En caso de que el valor de la proyección en la línea recta fuese superior o inferior a ciertos umbrales se le asignaría 0 o 100 directamente.
2. Normalización de valores cualitativos: El segundo método consiste en convertir los valores cualitativos en valores en un rango de 0 a 100, es decir, valores de 0, 1, 2 se convierten a puntajes de 0, 50 o 100 dependiendo el caso.

En ambos casos tanto los interceptos con los ejes en las ecuaciones lineales como los valores de umbral o los valores de las conversiones de valores cualitativos, están dados por el GFSI y varían para cada una de las 59 variables.

Para validar y con el fin de comparar el desempeño de los modelos antes y después de la normalización, se crean dos datasets, uno con los datos de las variables sin normalizar, y otro con los datos de las variables normalizadas. En ambos casos la variable a predecir consiste del puntaje del GFSI para cada país, mientras que las otras 59 variables son usadas para entrenar el modelo y predecir el GFSI.

5.2. Entrenamiento y validación de modelos con los datos del GFSI a escala de país.

Como se dijo anteriormente, se entrenaron modelos de machine learning con los datasets del GFSI a nivel de país con variables normalizadas y sin normalizar. Para la evaluación se dividió el conjunto de datos aleatoriamente en 80% para entrenamiento y 20% de evaluación, de forma que el modelo fuese entrenado solo con el grupo de datos de entrenamiento y evaluado usando el 20% de datos de prueba, con el fin de evitar que el modelo memorice los datos.

Se entrenaron múltiples modelos para una tarea de regresión haciendo uso sólo con el conjunto de datos que corresponden a entrenamiento (80% del total). Una vez entrenados los modelos se procedió a evaluar cada uno con el porcentaje restante de los datos que corresponden a el dataset de evaluación.

Para evaluar los modelos, se utiliza la métrica de Error Absoluto Medio (MAE), la cual nos dice la diferencia absoluta promedio entre el valor predicho por el modelo y el valor real promedio, esta métrica no es tan sensible como el error cuadrático medio, pero brinda un valor más real de qué tan preciso es el modelo. Los resultados de los modelos entrenados con datasets de variables normalizadas y no normalizadas se encuentran en la tabla 26.

Modelo	**MAE con datos normalizados**	**MAE con datos sin normalizar**
Regresión lineal	0.39	2.71
Regresión de crestas	0.39	3.34
Regresión de lazo	0.41	3.44
Regresión de la red elástica	0.39	3.42
Perceptrón multicapa	0.82	4.38
Máquinas de vector de soporte	0.38	3.54
Bosque aleatorio	3.29	3.32

Tabla 26. Puntaje GFSI para los modelos entrenados con datos normalizados y sin normalizar.

Como se puede observar en la tabla anterior, el modelo máquinas de vector de soporte tuvo el mejor puntaje MAE para datos normalizados, en cambio el modelo regresión lineal tuvo el mejor puntaje MAE para datos sin normalizar.

Considerando los puntajes para datos normalizados y sin normalizar, el modelo de regresión lineal es el que se comporta mejor en general. Considerando que la diferencia entre el puntaje MAE para datos normalizados de la regresión lineal y el modelo de máquinas de vector soporte es de 0.01 puntos lo cual no es muy significativo, el modelo elegido para el cálculo del GFSI a nivel del Cauca es el de regresión lineal.

5.3. Selección de las variables usadas para generar las predicciones del GFSI en Cauca.

Para el cálculo del GFSI a nivel de país como se vió en el anterior punto la EIU hace uso de 59 características, las mismas con las que se entrenó el modelo en el punto anterior y que son necesarias para generar nuevas predicciones, por lo tanto para obtener un índice de seguridad alimentaria a nivel del departamento del Cauca, es necesario hacer uso de las mismas variables pero utilizando los datos del dataset generado a partir del censo agropecuario, encuesta del ENSIN, formulario de sivigila, y datos meteorológicos; o datos relacionados al departamento del Cauca.

Debido a lo anterior, es necesario un procedimiento de selección de variables, de forma que se filtren de las 926 columnas del dataset del Cauca solo los 59 indicadores utilizados por el GFSI. En el **anexo D** en la pestaña "**Variables GFSI**", donde se encuentran explicados los indicadores usados para el cálculo del GFSI a escala de país en las columnas "**Variable**" y "**Descripción**", junto con la normalización de las variables en la columna "**Normalización (Score Calculation)**" y con las variables del departamento del Cauca utilizadas para reemplazar los indicadores del GFSI a escala de país en la columna "**Obtención Variable**".

La mayoría de variables utilizadas para cubrir los indicadores del GFSI se obtuvieron del dataset generado en el capítulo 4, sin embargo otras variables se obtuvieron a partir de la imputación de datos provenientes de informes a nivel departamental o nacional, o también a partir de los datos usados para calcular el GFSI de Colombia para el año 2014. Es importante mencionar que se realizó el mismo proceso de normalización de las variables que se realizó para la obtención del dataset de entrenamiento, además de realizar cálculos adicionales para transformar las variables provenientes del dataset del capítulo 4 en variables relacionadas con los indicadores del GFSI original a nivel de país.

Cabe resaltar que los datos de nutrición obtenidos corresponden a Popayán, Balboa, Bolívar, Paez, Patía, San Sebastián, Santander de Quilichao, Timbío, Timbiquí, lo anterior debido a que la encuesta ENSIN solo se realizó en estos municipios. Las variables de nutrición utilizadas para calcular el GFSI de los 33 municipios restantes se imputaron a partir del promedio de los datos del ENSIN de los 9 municipios disponibles.

5.4. Resultados de las predicciones del modelo GFSI en el Cauca.

Una vez seleccionadas las variables del departamento del Cauca según los indicadores del GFSI, se procede a calcular el índice de seguridad alimentaria GFSI a partir del modelo entrenado previamente con los datos normalizados y sin normalizar a nivel de país. Los resultados obtenidos del índice GFSI para el departamento del Cauca se encuentran en la Tabla 27. Se puede observar que el índice de seguridad alimentaria para el departamento del Cauca es de 57,44 puntos calculado a partir de datos normalizados. En comparación del índice para el departamento del Cauca con el índice GFSI de Colombia para el año 2014 calculado por el EIU, el departamento del Cauca cuenta con un déficit de -5.76 puntos por debajo del promedio nacional, el cual corresponde a 63.2 puntos. El índice del departamento del Cauca corresponde al promedio de los índices de los 42 municipios obtenidos a partir de datos normalizados.

El municipio con el índice GFSI más bajo es Timbiquí con 53,637 puntos, obtenido a partir de datos normalizados. En comparación con el índice del departamento del Cauca, el municipio de Timbiquí se encuentra 3.803 puntos por debajo del promedio del departamento (57.44 puntos). Adicionalmente, el municipio de Timbiquí se encuentra a 9.563 puntos por debajo del promedio nacional correspondiente al valor calculado por el EIU para el 2014.

De los resultados obtenidos, el municipio de Popayán obtuvo el puntaje más alto, correspondiente a 60.569 puntos, calculado a partir de datos normalizados. Al compararlo con el promedio del departamento del Cauca (57.44 puntos), Popayán se encuentra 3.129 puntos por encima del promedio departamental. Este resultado es de esperar, debido a que Popayán es la capital del departamento del Cauca, por lo que recibe mucha más atención del gobierno nacional y departamental. Sin embargo, al comparar el índice de Popayán con el promedio nacional, el municipio de Popayán se encuentra 2.631 puntos por debajo del índice de seguridad alimentaria GFSI de Colombia. El archivo que contiene los procedimientos para el

cálculo del índice de seguridad alimentaria, junto con los resultados se encuentra en el **anexo C**.

Municipios	GFSI sin normalizar	GFSI Normalizado
Popayán	59,5	60,569
Almaguer	54,33	56,445
Argelia	54,67	55,945
Balboa	54,9	57,019
Bolívar	55,583	56,787
Buenos aires	55,9	57,992
Cajibío	55,789	57,643
Caldono	54,841	56,525
Caloto	55,416	57,913
Corinto	56,27	58,6
El tambo	54,455	56,737
Florencia	56,461	57,975
Guachené	58,486	59,764
Guapi	53,274	55,952
Inzá	55,883	57,001
Jambaló	56,193	57,418
La sierra	54,945	56,643
La vega	55,55	56,997
López de micay	53,105	54,285
Mercaderes	54,212	56,504
Miranda	58,37	59,921
Morales	55,257	56,885
Padilla	59,055	59,724
Páez	56,112	57,142
Patía	55,052	56,723
Piamonte	54,548	56,924
Piendamó - tunía	57,976	58,801
Puerto tejada	58,022	59,847
Puracé	56,606	58,124
Rosas	56,343	57,593
San sebastián	55,857	56,821
Santander de quilichao	57,812	59,857
Santa rosa	55,78	55,877
Silvia	56,611	58,066
Sotará paispamba	58,285	58,775

Suárez	54,772	56,663
Sucre	53,156	55,095
Timbío	57,373	58,444
Timbiquí	53,219	53,637
Toribío	55,362	56,806
Totoró	54,935	56,538
Villa rica	58,027	59,345

Tabla 27. GFSI para los municipios del Cauca.

Capítulo 6

Conclusiones y Trabajos Futuros

6.1. Conclusiones.

Calcular un índice como el GFSI se hace difícil para un territorio con pocos recursos o problemas de facilidad de acceso a ciertas regiones como es el caso del Cauca. En este trabajo se demuestra como el uso de datos abiertos surge como una solución a la construcción de una base de datos que pueda integrar múltiples dimensiones de múltiples fuentes. Esta es una solución que si bien puede llegar a requerir más trabajo de pre-procesamiento y limpieza de los datos, así como presentar algunas debilidades como el desfase temporal de los datos o faltas de datos, es una solución que sigue siendo más económica que el realizar la recolección de los datos necesarios con métodos tradicionales como encuestas o censos.

Además de la integración de múltiples fuentes como encuestas y censos, surge la posibilidad de usar nuevas fuentes de datos de diferente modalidad como las imágenes satelitales. El uso de imágenes satelitales presenta una solución a la falta de datos, esto debido a que son imágenes que se encuentran disponibles a nivel mundial y con una resolución temporal muy buena. El hecho de poder tener una imágen cada aproximadamente 16 días como es el caso de Landsat-8 ó una imágen cada aproximadamente 5 días como es el caso de Sentinel-2, ofrece una resolución temporal mejor de la que ofrecen la mayoría de las encuestas o censos y a un costo muy bajo. Sin embargo hay que tener en cuenta que el uso de este

tipo de datos aumenta considerablemente el tamaño de los datos y por lo tanto los recursos computacionales necesarios para el procesamiento.

Métodos para extraer características de imágenes satelitales son presentados en este trabajo, destacando el método de extracción de características mediante una clasificación multi-etiqueta. El método se presenta como una solución óptima debido a la interpretabilidad y facilidad del método para integrar los datos con otras modalidades de datos como son los metadatos. Además los modelos ya entrenados con el dataset “Planet” han tenido buenos resultados a la hora de clasificar las imágenes satelitales de forma que se puede usar estas arquitecturas entrenadas en este trabajo como Resnet 50, VGG 16 o Vision Transformers, disponibles en el enlace: https://github.com/dsrestrepo/FoodSecurity/tree/main/Satelital%20Data/AmazonModel/architectures, junto con los pesos guardados (pesos: https://drive.google.com/drive/folders/1av_YPGx3xWK4m7Aefr_9OEV1v9YFEGDn?usp=sharing) para clasificar nuevas imágenes, presentando así una solución fácil de interpretar ya que no requiere de expertos para entender el resultado del modelo y es fácil de implementar, ya que no hay que re-entrenar el modelo, solo cargar los pesos.

Este trabajo demuestra la necesidad de apoyar la legislación y protocolos sobre datos abiertos de forma eficiente y oportuna, de forma que los datos recolectados por diferentes organizaciones puedan ser usados con fines investigativos. Problemas comunes con la liberación de datos se presentan habitualmente como tardanza en la liberación de datos o liberación de datos en formatos inconsistentes como como pdf. Algunos casos se vieron en el desarrollo de este proyecto, razón por la cual algunos datos necesarios para calcular el GFSI se debieron de imputar manualmente añadiendo las columnas al conjunto de datos.

El uso de técnicas de machine learning se presenta como una alternativa para el procesamiento y comprensión de grandes cantidades de datos. En este caso se usaron modelos de aprendizaje profundo para aprender a detectar características presentes en imágenes satelitales mostrando buenos resultados. Así mismo modelos de machine learning clásicos como regresión lineal se usaron para el cálculo del GFSI aprendiendo los pesos de cada variable a partir del conjunto de datos del GFSI a escala de país, dando una facilidad de uso e implementación para la solución de problemas que requieren de aprender de grandes cantidades de datos y extrapolar información.

6.2. Trabajos futuros.

Múltiples trabajos futuros se pueden derivar de la implementación actual. Un ejemplo es el extender el trabajo actual a diferentes regiones del país añadiendo más departamentos y municipios al dataset o incluso haciendo el modelo para todos los departamentos del país.

Por el lado de las imágenes satelitales también hay muchos trabajos pendientes. La exploración de nuevas fuentes de datos como Sentinel-2, el cual fue mencionado, pero cuyas imágenes no fueron analizadas ni exploradas durante este trabajo por problemas de capacidad de almacenamiento puede ser una buena fuente de información para mejorar el desempeño de los modelos actuales tanto de clasificación como de segmentación. Otra opción es el uso de imágenes provenientes del mismo satélite usado por Kaggle, el satélite Planet, el cual cuenta con mayor resolución, sin embargo no es de uso abierto.

Se pueden seguir implementando mejoras en el modelo encargado de la extracción de características. Como bien se vió en los ejemplos de segmentación, el uso de más bandas además de las comunes RGB, permite extraer más información de las imágenes. En este sentido Kaggle no solo provee imágenes RGB en el conjunto de entrenamiento, sino que también provee imágenes de múltiples bandas, por lo que un modelo como el Resnet 50 o el VGG 16 usados en este trabajo entrenados con estos conjuntos de datos podría tener un mejor desempeño. Cabe destacar que para realizar el entrenamiento de modelos con más bandas se debe tener en cuenta que el espacio de almacenamiento requerido aumenta considerablemente, así como el tiempo que llevaría entrenar el modelo, por estas razones y otras como que las imágenes extraídas fueron de Landsat-8 y que las bandas espectrales del dataset “Planet” podrían ser diferentes, se descartó el uso de un dataset de entrenamiento con imágenes multiespectrales.

Se puede hacer un análisis individual de cada una de las dimensiones de la FAO haciendo uso del dataset para entender mejor las falencias y fortalezas de cada municipio en cuanto a seguridad alimentaria haciendo uso del dataset. Además extender el uso del dataset a otras áreas más allá del campo de la seguridad alimentaria con tareas como el análisis de nutrición en series temporales, análisis de características sociodemográficas o incluso de cambio climático a través de series de tiempo.

Las técnicas tratadas en este trabajo para la fusión de imágenes satelitales fueron las que se consideraron más propicias debido a los recursos tanto de personal como de hardware, sin embargo se puede buscar otras alternativas para

integración de las imágenes satelitales con el dataset, algunas de las alternativas pueden ser las mencionadas como posibilidades en este mismo documento como el uso de técnicas como co-learning, redes neuronales de grafos o embeddings.

Si bien el dataset “Planet” es un dataset que ha sido de grán ayuda para el desarrollo de este trabajo, se ha visto que se tiene un grán potencial en el campo de los censos remotos a través de imágenes satelitales. Sin embargo en algunos casos los conjuntos de datos disponibles para el entrenamiento de las imágenes satelitales, puede que no sean suficientes para poder realizar diversas tareas. Ante el problema de la falta de conjuntos de datos, realizar un conjunto de datos de imágenes satelitales segmentadas para entrenamiento de modelos de segmentación semántica, así como la creación de datasets de clasificación de imágenes satelitales como el de “Planet” pero con otras características podría ser de utilidad para la comunidad científica.

Referencias

[1] Conceptos Básicos | Programa Especial para la Seguridad Alimentaria (PESA) Centroamérica | Food and Agriculture Organization of the United Nations [Online], Available: http://www.fao.org/in-action/pesa-centroamerica/temas/conceptos-basicos/en/, [Accessed: April 10, 2021].

[2] Supervivencia y desarrollo infantil [Online], Available: https://www.unicef.org/colombia/supervivencia-y-desarrollo-infantil, [Accessed: April 10, 2021].

[3] ENSIN: Encuesta Nacional de Situación Nutricional [Online], Available: https://www.icbf.gov.co/bienestar/nutricion/encuesta-nacional-situacion-nutricional#ensin3, [Accessed: April 10, 2021].

[4] Ubicación Geográfica - Consejo Regional Indígena del Cauca [Online], Available: https://www.cric-colombia.org/portal/estructura-organizativa/ubicacion-geogr

afica/, [Accessed: April 10, 2021].

[5] The Economist Intelligence Unit (EIU). (2021). Global Food Security Index 2020 Addressing structural inequalities to build strong and sustainable food systems. The Economist Intelligence Unit Limited.

[6] Sahoo, K., Sahoo, B., Choudhury, A. K., Sofi, N. Y., Kumar, R., & Bhadoria, A. S. (2015). Childhood obesity: causes and consequences. Journal of family medicine and primary care, 4(2), 187–192. https://doi.org/10.4103/2249-4863.154628

[7] Sharifi, A. (2020). Yield prediction with machine learning algorithms and satellite images. Journal of the Science of Food and Agriculture, 101(3), 891–896. https://doi.org/10.1002/jsfa.10696

[8] Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., & Xie, J. (2021). Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. Agricultural and Forest Meteorology, 297, 108275. https://doi.org/10.1016/j.agrformet.2020.108275

[9] Gómez, D., Salvador, P., Sanz, J., & Casanova, J. L. (2021). Modelling wheat yield with antecedent information, satellite and climate data using machine learning methods in Mexico. Agricultural and Forest Meteorology, 300, 108317. https://doi.org/10.1016/j.agrformet.2020.108317

[10] Chen, J., Chen, J., Zhang, D., Nanehkaran, Y. A., & Sun, Y. (2021). A cognitive vision method for the detection of plant disease images. Machine Vision and Applications, 32(1), 0. https://doi.org/10.1007/s00138-020-01150-w

[11] Ramadhani, F., Pullanagari, R., Kereszturi, G., & Procter, J. (2020). Automatic Mapping of Rice Growth Stages Using the Integration of SENTINEL-2, MOD13Q1, and SENTINEL-1. Remote Sensing, 12(21), 3613. https://doi.org/10.3390/rs12213613

[12] Salvador, P., Gómez, D., Sanz, J., & Casanova, J. L. (2020). Estimation of Potato Yield Using Satellite Data at a Municipal Level: A Machine Learning Approach. ISPRS International Journal of Geo-Information, 9(6), 343. https://doi.org/10.3390/ijgi9060343

[13] Yang, R., Ahmed, Z. U., Schulthess, U. C., Kamal, M., & Rai, R. (2020). Detecting functional field units from satellite images in smallholder farming systems using a deep learning based computer vision approach: A case study from Bangladesh. Remote Sensing Applications: Society and Environment, 20, 100413. https://doi.org/10.1016/j.rsase.2020.100413

[14] Qu, Y., Zhao, W., Yuan, Z., & Chen, J. (2020). Crop Mapping from Sentinel-1

Polarimetric Time-Series with a Deep Neural Network. Remote Sensing, 12(15), 2493. https://doi.org/10.3390/rs12152493

[15] Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., & Fritschi, F. B. (2020). Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sensing of Environment, 237, 111599. https://doi.org/10.1016/j.rse.2019.111599

[16] Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agricultural and Forest Meteorology, 284, 107886. https://doi.org/10.1016/j.agrformet.2019.107886

[17] Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y., & Guanter, L. (2020). Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. Environmental Research Letters, 15(2), 024019. https://doi.org/10.1088/1748-9326/ab68ac

[18] Maimaitijiang, M., Sagan, V., Sidike, P., Daloye, A. M., Erkbol, H., & Fritschi, F. B. (2020). Crop Monitoring Using Satellite/UAV Data Fusion and Machine Learning. Remote Sensing, 12(9), 1357. https://doi.org/10.3390/rs12091357

[19] Wang, Y., Zhang, Z., Feng, L., Du, Q., & Runge, T. (2020). Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States. Remote Sensing, 12(8), 1232. https://doi.org/10.3390/rs12081232

[20] Wang, X., Huang, J., Feng, Q., & Yin, D. (2020). Winter Wheat Yield Prediction at County Level and Uncertainty Analysis in Main Wheat-Producing Regions of China with Deep Learning Approaches. Remote Sensing, 12(11), 1744. https://doi.org/10.3390/rs12111744

[21] Zhang, L., Zhang, Z., Luo, Y., Cao, J., & Tao, F. (2019). Combining Optical, Fluorescence, Thermal Satellite, and Environmental Data to Predict County-Level Maize Yield in China Using Machine Learning Approaches. Remote Sensing, 12(1), 21. https://doi.org/10.3390/rs12010021

[22] Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Han, J., & Li, Z. (2020). Identifying the Contributions of Multi-Source Data for Winter Wheat Yield Prediction in China. Remote Sensing, 12(5), 750. https://doi.org/10.3390/rs12050750

[23] Yang, Q., Shi, L., Han, J., Zha, Y., & Zhu, P. (2019). Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. Field Crops Research, 235, 142–153.

https://doi.org/10.1016/j.fcr.2019.02.022

[24] Ma, J. W., Nguyen, C. H., Lee, K., & Heo, J. (2018). Regional-scale rice-yield estimation using stacked auto-encoder with climatic and MODIS data: a case study of South Korea. International Journal of Remote Sensing, 40(1), 51–71. https://doi.org/10.1080/01431161.2018.1488291

[25] Zhao, H., Chen, Z., Jiang, H., Jing, W., Sun, L., & Feng, M. (2019). Evaluation of Three Deep Learning Models for Early Crop Classification Using Sentinel-1A Imagery Time Series—A Case Study in Zhanjiang, China. Remote Sensing, 11(22), 2673. https://doi.org/10.3390/rs11222673

[26] Terliksiz, A. S., & Altylar, D. T. (2019). Use Of Deep Neural Networks For Crop Yield Prediction: A Case Study Of Soybean Yield in Lauderdale County, Alabama, USA. 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), https://doi.org/10.1109/agro-geoinformatics.2019.8820257

[27] Van Tricht, K., Gobin, A., Gilliams, S., & Piccard, I. (2018). Synergistic Use of Radar Sentinel-1 and Optical Sentinel-2 Imagery for Crop Mapping: A Case Study for Belgium. Remote Sensing, 10(10), 1642. https://doi.org/10.3390/rs10101642

[28] Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018). Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data. Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. https://doi.org/10.1145/3209811.3212707

[29] Xie, Michael & Jean, Neal & Burke, Marshall & Lobell, David & Ermon, Stefano. (2015). Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping.

[30] Una introducción a los conceptos básicos de la seguridad alimentaria [Online], Available: http://www.fao.org/3/al936s/al936s00.pdf , [Accessed: April 18, 2021].

[31] Von Grebmer, Klaus; Bernstein, Jill; Nabarro, David; Prasai, Nilam; Amin, Shazia; Yohannes, Yisehac; Sonntag, Andrea; Patterson, Fraser; Towey, Olive; and Thompson, Jennifer. 2016. The Concept of the Global Hunger Index. In 2016 Global hunger index: Getting to zero hunger. Chapter 1 Pp. 6-9. Bonn Washington, DC and Dublin: Welthungerhilfe, International Food Policy Research Institute, and Concern Worldwide. http://dx.doi.org/10.2499/9780896292260_01.

[32] Ballard T, Kepple A, Cafiero C. The food insecurity experience scale: development of a global standard for monitoring hunger worldwide. Rome: FAO. Retrieved from http://www.fao.org/fileadmin/templates/ess/voh/FIES_Technical_Paper_v1.1.

pdf. [Accessed August 14, 2021].

[33] What Is Data Mining: Definition, Examples, Tools, and Techniques (For Beginners) [Online], Available: https://bootcamp.pe.gatech.edu/blog/what-is-data-mining/, [Accessed: Februaryl 06, 2022].

[34] Conceptos de minería de datos [Online], Available: https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions, [Accessed: Februaryl 06, 2022]

ANEXOS
Tesis de Trabajo de Grado

David Santiago Restrepo Rodriguez
100614021014
Luis Enrique Pérez Maldonado
100615020647

Director: PhD. Diego Mauricio López Gutiérrez
Co-Director. PhD. Rubiel Vargas Cañas.

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones Departamento de Telemática
Línea de Investigación en e-Salud
Popayán, agosto de 2021

Capítulo 7

Anexo A

Durante la realización del trabajo de grado, se elaboró un artículo el cual se presenta en el anexo A. El artículo pasó la fase de revisión y se encuentra en fase de publicación en la revista Frontiers in Nutrition (de la editorial Frontiers Media SA), la cual se encuentra en el cuartil Q1 según el Scimago Journal & Country Ran (SJR del año 2020), en la categoría A1 según consulta en el listado de homologación de Publindex (Min Ciencias, 2021).

Autores del artículo: David S. Restrepo, Luis E. Perez, Diego M. Lopez, Rubiel Vargas-Cañas. Juan S. Osorio.

Título: Multi-dimensional Dataset of Open Data and Satellite Images for Characterization of Food Security and Nutrition

Revista: Frontiers in Nutrition (de la editorial Frontiers Media SA)

Estado: Publicado (Provisionally Accepted)

El artículo se encuentra en el siguiente enlace: https://www.frontiersin.org/articles/10.3389/fnut.2021.796082/full?&utm_source=Email_to_authors_&utm_medium=Email&utm_content=T1_11.5e1_author&utm_campaign=Email_publication&field=&journalName=Frontiers_in_Nutrition&id=796082

Capítulo 8

Anexo B

En este anexo se presenta el dataset creado a partir de la fusión de los diferentes datos provenientes de las diferentes fuentes abiertas (encuesta ENSIN, censo nacional agropecuario, formulario SIVIGILA, datos metereológicos, medida de pobreza multidimensional), la limpieza y procesamiento de los mismos.

Restrepo, david; Lopez, Diego; Perez, Luis; Vargas-Canas, Rubiel; Osorio, Juan (2021), “Multidimensional Dataset Of Food Security And Nutrition In Cauca.”, Mendeley Data, V1, doi: 10.17632/wsss65c885.1

Disponible en el enlace: https://data.mendeley.com/datasets/wsss65c885/1

Capítulo 9

Anexo C

En el anexo C se presenta el archivo que contiene el los procedimientos para el cálculo del índice GFSI para el departamento del Cauca, junto con el modelo, y los resultados correspondientes a los índices para cada municipio del departamento del Cauca. El archivo se encuentra en el lenguaje de programación python, utiliza las librerías pandas, scikit-learn, numpy.

Disponible en el siguiente enlace: https://github.com/dsrestrepo/SeguridadAlimentaria/blob/25f7273c2ef9e510a189420de12c2631e7993ffc/food_security_index_42Municipalities.ipynb

Capítulo 10

Anexo D

En el anexo D se encuentra un archivo en formato excel que contiene los diccionarios de las variables seleccionadas de la encuesta ENSIN, y del censo nacional agropecuario. Además contiene el diccionario de indicadores del GFSI, junto con los valores de normalización de las variables y las variables seleccionadas para reemplazar los indicadores del GFSI.

El archivo se encuentra en el siguiente enlace: Formularios.xlsx

Capítulo 11

Anexo E

En el presente anexo se incluyen todos los archivos utilizados, obtenidos, durante la elaboración del proyecto de grado. El enlace corresponde al github que contiene los dataset de donde se obtuvo la información, archivos de imágenes satelitales, los resultados del índice de seguridad alimentaria GFSI para el departamento del Cauca, el modelo para el cálculo del índice de seguridad alimentaria, entre otros.

El enlace al directorio de github es el siguiente: https://github.com/dsrestrepo/SeguridadAlimentaria

Capítulo 12

Anexo F

En el presente anexo se encuentra el enlace que direcciona a la página web desarrollada que contiene toda la información relacionada al índice de seguridad alimentaria calculado para el departamento del Cauca, junto con información del artículo publicado y el proyecto de investigación relacionado.

La página web se encuentra en el siguiente enlace: https://sites.google.com/view/seguridad-alimentaria-y-nutric/inicio