UNIVERSIDAD TÉCNICA DE MACHALA

FACULTAD DE INGENIERÍA CIVIL

MAESTRÍA EN SOFTWARE

CLASIFICACIÓN DE LA COBERTURA DEL SUELO EN LA PROVINCIA DEL CHIMBORAZO MEDIANTE IMÁGENES SATELITALES Y APRENDIZAJE AUTOMÁTICO

ING. LUIS EFRAIN QUINGUE GUAMINGA

TUTOR: ING. EDUARDO ALEJANDRO TUSA JUMBO, PHD.

MACHALA - ECUADOR

2022

DEDICATORIA

A Dios por ser el guía de mis metas y proyectos, a mi familia que son el soporte a lo largo de mi vida enseñando principios y valores que me ha llevado a la persona que soy, a mi madre que ha sido el principal ejemplo de valentia y sacrificio para sus hijos. De manera especial a mi hermana Maritza por acompañar en esta travesia y estar cuando mas lo necesitaba; para todos ustedes con infinito amor.

Efraín

AGRADECIMIENTO

Agradezco a Dios por proveer salud, bienestar y sabiduría para encaminar y cumplir mis sueños, a mi familia por su apoyo incondicional y mis amigos por estar siempre en las buenas y más en las malas. A mis compañeros y al programa de Maestría en Software por abrir las puertas de la Universidad Técnica de Machala y acompañar en este proceso académico. De manera particular mi agradecimiento al Ing. Eduardo Tusa por instruir y compartir su conocimiento en la consecución del presente objetivo.

Efraín

RESPONSABILIDAD DE AUTORÍA

Yo, LUIS EFRAIN QUINGUE GUAMINGA con C.I. 0604508390, declaro que el trabajo "CLASIFICACIÓN DE LA COBERTURA DEL SUELO EN LA PROVINCIA DEL CHIMBORAZO MEDIANTE IMÁGENES SATELITALES Y APRENDIZAJE AUTOMÁTICO", en opción al título de Magister en Software, es original y auténtico; cuyo contenido: conceptos, definiciones, datos empíricos, criterios, comentarios y resultados son de mi exclusiva responsabilidad.

LUIS EFRAIN QUINGUE GUAMINGA

C.I.: 0604508390

Machala, 2022/11/25

REPORTE DE SIMILITUD

CERTIFICACIÓN DEL TUTOR

Yo EDUARDO ALEJANDRO TUSA JUMBO con C.I. 0704323427; tutor del trabajo de "CLASIFICACIÓN DE LA COBERTURA DEL SUELO EN LA PROVINCIA DEL CHIMBORAZO MEDIANTE IMÁGENES SATELITALES Y APRENDIZAJE AUTOMÁTICO", en opción al título de Magister en Software, ha sido revisado, enmarcado en los procedimientos científicos, técnicos, metodológicos y administrativos establecidos por el Centro de Posgrado de la Universidad Técnica de Machala (UTMACH), razón por la cual doy fe de los méritos suficientes para que sea presentado a evaluación.

Firmado electrónicamente por:
EDUARDO ALEJANDRO TUSA JUMBO

EDUARDO ALEJANDRO TUSA JUMBO

C.I. 0704323427

Machala, 2022/11/25

CESIÓN DE DERECHOS

Yo, LUIS EFRAIN QUINGUE GUAMINGA, Declaro que estoy de acuerdo con ceder los derechos de autoría del presente trabajo investigativo a la Universidad Técnica de Machala. Cualquier uso ya sea total o parcial debe ser realizado con la autorización previa de la institución previamente mencionada.

Firmado electrónicamente por:
LUIS EFRAIN QUINGUE GUAMINGA

LUIS EFRAIN QUINGUE GUAMINGA

C.I.: 0604508390

Machala, 2022/11/25

RESUMEN

En la provincia de Chimborazo la falta de información y estudios geográficos realizados es una de las causas de la escaza información respecto a las clases de coberturas de suelo. En el presente trabajo se ha desarrollado los algoritmos de clasificación CART, RF, SVM basado en aprendizaje supervisado de ML e imágenes satelitales S2. El soporte de las TIC ha tenido un rol importante como proveedor de datos y procesamientos en la nube mediante la plataforma GEE. Se ha definido 7 clase de coberturas del suelo: agua, urbano, forestal, cultivo, suelo desnudo, arbustivo y nieve; que se clasifica utilizando 7 de 13 bandas discriminantes de S2, además de los índices NDVI, BSI y RGB para composición del dataset. Con 73% de datos de entrenamiento (1581) y 27% para prueba (598) etiquetados en la plataforma GEE se ha construido los dataset de entrada. Para obtener los clasificadores se aplica las fases de la metodología CRISP-DM que ha permitido desarrollar los tres modelos ML validadas ejecutando tres casos de pruebas para cada uno, donde se determina que los algoritmos RF y SVM alcanzan mejores métricas por encima del ≈80%. Validado los modelos ML aplicando la inferencia estadística se ha realizado el análisis de correlación y prueba de hipótesis, con índice kappa diferente de cero para los tres modelos ML se establece la existencia de concordancia; por lo tanto, se afirma que las imágenes satelitales proporcionan suficiente información discriminatoria. Finalmente, con el análisis y discusión de resultados se determina que el algoritmo RF tiene mejor desempeño en base a los resultados de precisión general ≈88,80% e índice kappa de 0.867. El trabajo finaliza dando pauta a estudios futuros como: comparación de cambios de coberturas e implementación de los modelos ML en sistemas para soporte de decisión, tomando como referencia el presente estudio.

Palabras claves: <APRENDIZAJE AUTOMÁTICO>, <GEE>, <CART>, <RF>, <SVM>, <SENTINEL 2>, <CRISP-DM>

ABSTRACT

In the province of Chimborazo, the lack of information and geographic studies carried out is one of the causes of the lack of information regarding the types of land cover. In the present work, the CART, RF, SVM classification algorithms based on supervised learning of ML and S2 satellite images have been developed. TIC support has played an important role as a provider of data and cloud processing through the GEE platform. 7 classes of land cover have been defined: water, urban, forest, cultivation, bare soil, bushy and snow which is classified using 7 of 13 discriminant bands of S2, in addition to the NDVI, BSI and RGB indices for dataset composition. With 73% of training data (1581) and 27% for testing (598) labeled in the GEE platform, the input datasets have been built. To obtain the classifiers, the phases of the CRISP-DM methodology are applied, which has allowed the development of the three validated ML models, executing three test cases for each one, where it is determined that the RF and SVM algorithms achieve better metrics above ≈80%. Once the ML models have been validated by applying statistical inference, the correlation analysis and hypothesis test have been carried out, with a kappa index different from zero for the three ML models, the existence of concordance is established; therefore, it is claimed that satellite images provide sufficient discriminatory information. Finally, with the analysis and discussion of the results, it is determined that the RF algorithm has a better performance based on the results of general precision ≈88.80% and kappa index of 0.867. The work ends by giving guidelines for future studies such as: comparison of coverage changes and implementation of ML models in decision support systems, taking the present study as a reference.

Palabras claves: <MACHINE LEARNING>, <GEE>, <CART>, <RF>, <SVM>, <SENTINEL 2>, <CRISP-DM>

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ABREVIATURAS

TIC	Tecnología de la Información y Comunicación
SLR	Revisión Sistemática de Literatura (por sus siglas en ingles)
LCLU	Uso y Cobertura del Suelo (por sus siglas en ingles)
LC	Cobertura del Suelo (por sus siglas en ingles).
MAGAP	Ministerio de Agricultura Ganadería Acuacultura y Pesca
MAE	Ministerio del Ambiente
ML	Aprendizaje Automático (por sus siglas en ingles)
CRISP-DM	Proceso Estándar Intersectorial Para Minería De Datos
GEE	Google Earth Engine
S2	Satélite Sentinel 2
RGB	Rojo, Verde, Azul (Red, Green, Blue)
NDVI	Representación del Índice de Vegetación
BSI	Índice de Suelo Desnudo (por sus siglas en ingles)
CART	Arboles de Clasificación y Regresión (por sus siglas en ingles)
RF	Bosques Aleatorios (por sus siglas en ingles)
SVM	Soporte de Máquina Vectorial (por sus siglas en inglés)

INTRODUCCIÓN

La transformación de las superficies terrestres por fenómenos naturales o acciones del hombre, son causas que determinan la evolución del suelo que tienen incidencia en las actividades como la agricultura, pesca o la conservación ambiental. En este contexto, la clasificación y uso del suelo permite la representación de tipos de coberturas terrestres como: forestación [1], deforestación, vegetación, cambios climáticos; así también la identificación de zonas y propiedades del suelo para la agricultura [2] y análisis de elevaciones (DEM) terrestres [3].

La clasificación de cobertura del suelo una de las herramientas para obtener información de una geografía es aplicable en varios contextos. Mediante la revisión sistemática de literatura se ha obtenido estudios sobre la temática, por ejemplo: evaluación de mapas para representar diferencias en la variación de bosques según el continente utilizando imágenes del satélite Landsat [1]; análisis de precisión de las zonas montañosas de 4 estados de EE. UU realizado por Ravanelli, Nascetti y Crespi, 2020 [3] utilizando la computación en la nube Google Earth Engine (GEE).

Así mismo en el área de la agricultura, el estudio realizado por Yu et al, 2020 [4], analiza la clasificación de coberturas del suelo de zonas de cultivo y bosques de un área superficie de 87 kilómetros cuadrados de Habei, China fusionando imágenes satelitales y sus resoluciones espaciales. Wuyun et al [5], en el mismo país estudia las causas que origina cambios en las superficies agrícolas mediante imágenes satelitales, utilizando la plataforma GEE y datos del satélite Landsat, entre los años 1990 a 2000.

En el contexto latinoamericano estudio realizado en Costa Rica, relaciona las coberturas de suelo con variación de temperaturas, describe que la presencia de vegetación influye en la existencia de la humedad, por lo contrario, los materiales de las coberturas urbanas inciden en el aumento de la temperatura, concluyendo que la cobertura del suelo incide en el cambio climático [6].

Los estudios mencionados permiten obtener información técnica, lo que permite definir áreas geográficas según los tipos, a la vez contribuye en la toma de decisión de organizaciones y entidades gubernamentales como los Gobiernos Autónomos Descentralizados (GAD) en Ecuador [7], apoyados mediante la tecnología.

El análisis y clasificación del suelo de manera convencional in situ está sujeto a la intervención humana mediante la recopilación de información, debido que se requiere un tratamiento según los orígenes de datos. Por lo tanto, según Ríos et al [8] este proceso es un paso previo para otro análisis que permite diferenciar los tipos de clasificaciones a mayor o menor detalle.

El avance de la tecnología ha incorporado satélites para la captura de imágenes de alta resolución, esto contribuye el tratamiento y análisis de información geográfica y clasificación del suelo, plataformas como GEE con grandes volúmenes de información histórica de imágenes de distintos satélites [9]. El aprovechamiento de datos se lo hace empleando el aprendizaje automático (ML, por sus siglas en inglés) mediante la aplicación de técnicas y métodos de clasificación de coberturas terrestres [10].

La clasificación de cobertura y uso del suelo (LULC, por sus siglas en inglés) mediante ML requiere la selección de fuente de datos, selección de área geográfica y etiquetado, este último mediante el dispositivo GPS o por el basto conocimiento del área de estudio. La big data y aprendizaje automático inmersos en el campo de la teledetección satelital requieren grandes infraestructuras computacionales para el procesamiento, siendo esto una limitante para los analistas e investigadores. Para hacer frente las plataformas GEE y Amazon Web Services (AWS, por sus siglas en inglés) proveen arquitecturas basadas en las nubes con data de satélites Landsat y Sentinel [11], reduciendo las limitaciones de procesamiento de datos [12].

Mediante la revisión de literatura se ha explorado estudios e investigaciones realizadas sobre la clasificación de coberturas del suelo en el territorio ecuatoriano. Cartaya et al [8] en 2014, estudia la clasificación de cobertura vegetal y uso de la tierra en la provincia de Manabí, mediante análisis de componentes principales. Castelo et al [7], aplicando las redes neuronales clasifica cinco tipos de cobertura del suelo de la provincia de Chimborazo, sin embargo, el objetivo del estudio es la exploración y explotación de big data y computación en la nube; este último el estudio más aproximado a la presente investigación.

De la misma forma se ha explorado el banco de datos geográficos y coberturas del suelo del área seleccionada en el Sistema Nacional de Información (SNI), obteniendo como resultado datos del año 2014 donde algunos vínculos y recursos no están disponibles. Entonces no se ha obtenido información geográfica actualizada con acceso libre [13].

El presente estudio está motivado por la carencia de información geográfica digitalizada, entre otras, las causas que influyen son: información geográfica desactualizada o no se dispone, carencia de clasificación del suelo; cuyas consecuencias conlleva al desconocimiento de tipos de suelo para su uso y explotación, además de información desactualizada. Considerando como caso de estudio la zona geográfica de la provincia de Chimborazo ubicada en el centro del Ecuador con una extensión jurisdiccional de 6.578,10 Km2, dividido políticamente en 10 cantones y 45 parroquias rurales [14]; la presente investigación tiene como finalidad desarrollar modelos de clasificación de cobertura del suelo utilizando imágenes satelitales y ML.

El procesamiento y explotación de datos geográficos se utiliza en la plataforma GEE, apoyados en los estudios realizados por [7], [10], [15] que emplean herramientas tecnológicas y aprendizaje automático sobre esta plataforma para la obtención de datos de imágenes satelitales y procesamiento en la nube. A demás, se construye una data geográfica etiquetada con los tipos de coberturas basadas en las categorías establecidas por el Ministerio de Desarrollo Urbano y Vivienda (MIDUVI) [15].

Las categorías que se utiliza en la clasificación son las definidas en los niveles I y II que corresponde a categorías de bosques, tierras agropecuarias, vegetación, cuerpos de agua, zona antrópica, otras tierras y subclasificaciones de las categorías citadas. Se emplea técnicas de aprendizaje automático supervisado, arboles de regresión y clasificación [16] (CART, por sus siglas en inglés) [17], máquina de soporte vectorial (SVM, por sus siglas en inglés) [18] y árboles aleatorios (RF, por sus siglas en inglés) [17].

El modelo de clasificación es analizada y evaluada según las cantidades de bandas y pixeles que son dependientes del satélite seleccionado además de la resolución de las imágenes. Por lo tanto, se evalúa si las imágenes satelitales proporcionan suficiente información discriminatoria para la clasificación de coberturas. De la misma forma se calcula la precisión de cada modelo de clasificación, matriz de confusión, índice de kappa y se analiza el clasificador los resultados.

La presente investigación está constituida por cuatro capítulos. En el primer capítulo se fundamenta los temas y herramientas utilizadas en la investigación mediante el estado de arte. El segundo capítulo describe de manera sistemática la aplicación de la metodología utilizada donde se recolecta los datos para el análisis. En el cuarto y capítulo final constituye el análisis de los resultados obtenidos y se contrasta con los objetivos planteados. Además, se proyectan estudios futuros a partir de la presente investigación.

FORMULACIÓN DEL PROBLEMA

¿Cómo desarrollar modelos para clasificación de cobertura del suelo de un territorio para reducir la carencia de información geográfica?

Sistematización del problema

- ¿Qué estudios previos existe para la clasificación de la cobertura del suelo aplicando técnicas de aprendizaje automático con imágenes satelitales?
- ¿Cómo se construye una base de datos de imágenes satelitales que provee la plataforma GEE?
- ¿Como se desempeña los modelos de clasificación de cobertura terrestre según las métricas de evaluación de algoritmos de aprendizaje automático?

OBJETIVOS

Objetivo general

Proponer modelos para clasificación de la cobertura del suelo, mediante la aplicación de aprendizaje automático de imágenes satelitales para la toma de decisiones.

Objetivos específicos

- Investigar técnicas y algoritmos de clasificación de la cobertura del suelo que apliquen aprendizaje automático mediante una revisión sistemática de la literatura.
- Elaborar una base datos de imágenes satelitales etiquetada de la provincia de Chimborazo, utilizando la plataforma GEE.
- Evaluar los modelos de clasificación mediante métricas de rendimiento, interpretación y discusión de los resultados para utilización en la toma de decisiones.

CAPÍTULO I

1. MARCO TEÓRICO

1.1 Antecedentes Históricos

La cobertura del suelo (LC, por sus siglas en inglés) hace referencia a la información material y física de la superficie de la tierra, su uso [10] y los cambios que sufre a través de los años, por efectos naturales o por intervención humana, como la tala de árboles y los incendios forestales [19]. A lo largo del tiempo han surgido investigaciones y estudios para determinar los tipos de superficies de suelo para el análisis y toma de decisiones a entidades públicas y privadas [7]. En la actualidad, herramientas como GEE o AWS [19], [20] proveen gran cantidad de imágenes de alta resolución que ayuda a estudios con enfoque de clasificación de coberturas del suelo.

En el presente apartado se explora los estudios relacionados a través del tiempo, aquellos cuyo objetivo haya sido la clasificación del suelo mediante la aplicación del aprendizaje automático, así también las herramientas o software informático empleado para el procesamiento de imágenes satelitales como fuentes de datos.

En 2011, el estudio de Song et al [1] compara seis mapas representativos de alta resolución de la cobertura terrestre global para un mapeo forestal. La evaluación se realiza mediante la aplicación de dos pasos: armonización de leyendas y agregación espacial con datos obtenidos desde GEE y satélite Landsat mediante un examen visual analítico. El resultado del estudio revela que la precisión de la cobertura forestal varía según el continente, existencia de discrepancias significativas en la representación de bosques.

Así también en 2016, se ha estudiado la relación que tiene el cambio climático con el cambio de la cobertura del suelo. El estudio realizado por Barrientos et al [6] en Costa Rica, utiliza las herramientas ARCGIS y QGIS para la vectorización de fotografías de coberturas del suelo y uso de complemento Semi-Automatic Classification (SCP) Congedo para la estimación de temperaturas de imágenes satelitales Landsat-8 (L8). Mediante el cruce de información y análisis de convergencia de los dos grupos de datos ha llegado a la conclusión que la cobertura del suelo es un factor influyente en los cambios de temperatura; aunque no es el único factor que determina estos cambios.

La plataforma GEE se ha revelado como uno de los desarrollos más prometedores para el acceso y análisis de datos de ciencias de la Tierra. La infraestructura de la nube de Google contiene un catálogo de imágenes satelitales de varios petabytes de satélites conocidos, que incluyen Landsat, Sentinel, MODIS, entre otros. Además, proporciona herramientas de programación para acceder, operar y visualizar dichos datos de una manera fácil y escalable [20].

El estudio de García et al [21] en 2017, utiliza las prestaciones de la plataforma GEE, efectúa la detección de imágenes multitemporales de alta resolución de L8 mediante el modelo de regresión; a continuación, se elimina las nubes mediante el filtrado aplicando algoritmo Fmask durante un año, entre el 2015 y 2016 para la zona de la ciudad de Valencia, España. Los resultados mostraron que, al explotar la riqueza de la información en la dimensión temporal, la precisión de detección de nubes aumenta, especialmente en casos críticos como superficies brillantes.

En 2018, otro estudio que utiliza GEE se puede apreciar en Pimple et al [11], donde cuantifican la presencia de manglares en un área aproximada de 240 Km de Tailandia de los últimos 30 años con imágenes obtenidas del satélite Landsat, aplicando el algoritmo de clasificación RF [22], [23] utilizando 60 muestras de bosques de manglar. Como resultados demuestran cambios significativos en la cobertura de la tierra, particularmente en los bosques de manglares en el área de estudio entre 1987 y 2017. Las áreas camaroneras, agrícolas y las tierras baldías presentaron grandes cambios. Además, el estudio se apoya en la aplicación de la herramienta GEE, para el mapeo y monitoreo de ecosistemas costeros debido a la confiabilidad de datos a largo plazo.

En el mismo año 2018, Wu et al [9] utilizan datos desde el año 2003 a 2013 desde fuentes históricas de estudios similares y GEE con la finalidad de comprender mejor sobre los bosques del mundo. Se aplica el método MOS que es un tipo de red neuronal que entrena la muestra de entrada con aprendizaje competitivo para producir un mapa de agrupamiento. Se obtienen 9 categorías de bosques correlacionadas con zonas climáticas y zonas de bosques.

En el 2019 el estudio realizado en dos áreas de Reino Unido por Zhang et al [24], utiliza Join Deep Learning (JDL) con imágenes LC y uso del suelo (LU, por sus siglas en inglés) tomadas por cámaras aéreas Vexcel UltraCam desde julio 2012 a mayo 2016. Se aplicaron los métodos de aprendizaje automático: SVM [18], campos aleatorios de Markov (MRF) [21] y el análisis basado en objetos con SVM y CNN por píxeles. Se obtuvieron los mejores resultados en la iteración 10 del experimento, ajustando a una distribución de probabilidad conjunta de modo que las predicciones se utilizan para actualizar entre sí de forma iterativa.

GEE es una contribución muy importante en la evolución de estudios de coberturas de suelo, a diferencia de herramientas como ENVI o ARCGIS, GEE proporciona en la nube grandes cantidades de datos a escala global. Esto ha permitido clasificar coberturas de tipos tierras cultivadas, bosques, pastizales, tierras de arbustos, humedales, cuerpos de agua, edificios artificiales, terrenos desnudos, nieve y hielo, como lo menciona en el estudio realizado por Jue et al [25] en 2019 que emplea el clasificador RF por la alta capacidad de determinación de variables, obteniendo la precisión general de 93.0% con coeficiente de kappa por encima de 0.90, en análisis de imágenes de L8 del año 2014.

El aprendizaje automático supervisado también ha jugado un papel importante en las clasificaciones de suelos, la aplicación de los algoritmos como RF o SVM ha permitido agilizar la tarea a los investigadores en la explotación de imágenes satelitales de alta resolución tomadas por Sentinel-2 (S2). En el año 2020, Barbosa et al [26] aprovecha la información satelital para clasificar siete clases para LC y LU: agua, formación de pastizales, formación de sábanas, formación de bosques, áreas planas, expansión urbana y edificios, del Distrito Federal de Brasil. El algoritmo de clasificación RF presentó resultados óptimos con mejores desempeños para coberturas vegetales y objetos urbanos.

Así también este tipo de estudios han contribuido para el análisis de desastres naturales. Como el estudio de Suresh Babu and Vanama [19] en 2020 que hace un mapeo de áreas afectadas por los incendios forestales de Australia desde agosto de 2019 hasta febrero de 2020. Con imágenes obtenidas por L8 y S2 desde la plataforma GEE analiza las bandas pre y post incendio del área de estudio, obteniendo índices del espectro que multiplicados por un factor de escala da como resultado 7 niveles de gravedad de los incendios. Los resultados demostraron que el índice de combustión relativizado (RBR) tiene mayor precisión que el índice de combustión normalizado diferenciado (dNBR) para los conjuntos de datos analizados de S2 y L8.

Ya en el año 2021 el estudio realizado por Mayer et al [12] en Camboya sudeste de Asia con imágenes satelitales de Sentinel-1 utilizando la librería TensorFlow de Google Cloud Storage (GCS) y plataforma GEE importa datos a una máquina virtual con altas prestaciones técnicas para el procesamiento con la integración de la API Google AI Platform. Se estudiaron dos métodos de etiquetado automático de datos implementando U-Net, un tipo de CNN para mapeo de aguas superficiales de Sentinel-1. Al probar 12 modelos se determina que el etiquetado JRC muestra mejor rendimiento respecto al etiquetado Edge-Otsu. Además, GEE como computación en la nube es una tecnología versátil para implementar tecnologías de aprendizaje profundo a gran escala.

El estudio más cercano a los objetivos de la presente investigación es de Castelo et al [7] realizado en el año 2021 sobre la clasificación de cobertura terrestre mediante la aplicación de Red Neuronal Artificial (ANN) del territorio ecuatoriano de manera específica de la superficie territorial de la provincia de Chimborazo. Utilizando las imágenes satelitales de S2 y computación en la nube utilizando la plataforma GEE; con entrenamiento de ANN de 100 épocas con precisión de 92%, se obtuvo la clasificación de cinco coberturas: agua, nieve, árboles, vegetativo y no vegetativo. Sin embargo, el estudio no incursiona de manera detallada en la clasificación del suelo, más bien es la explotación de las bondades de la computación en la nube y el uso de ANN.

La clasificación de la cobertura del suelo, siendo una de las áreas de estudio muy joven, hoy en día se sigue revisando la literatura; la clasificación supervisada y no supervisada desempeña un papel importante en la conversión de la imagen ráster multibanda en ráster de una sola banda para diferenciar las clases de cobertura terrestre. En el ámbito de la teledetección, se tiene acceso a una gran cantidad de información, pero la mayoría de los datos no están etiquetados. Para categorizar la imagen de datos se utiliza el clasificador por lo que la precisión de la clasificación de la cobertura terrestre depende en gran medida del momento y la ubicación de la imagen [2].

También hay que considerar que el procesamiento de imágenes de alta resolución demanda grandes capacidades técnicas de equipos computacionales, siendo una de las limitaciones para implementaciones tecnológicas en entornos operativos o modelos en tiempo real [12]. Por lo que la plataforma GEE [12], [20], [26] y el aprendizaje automático [7], [24], [27] aportan de manera significativa a implementación de métodos y modelos para la clasificación de suelo mediante la utilización de imágenes satelitales [26].

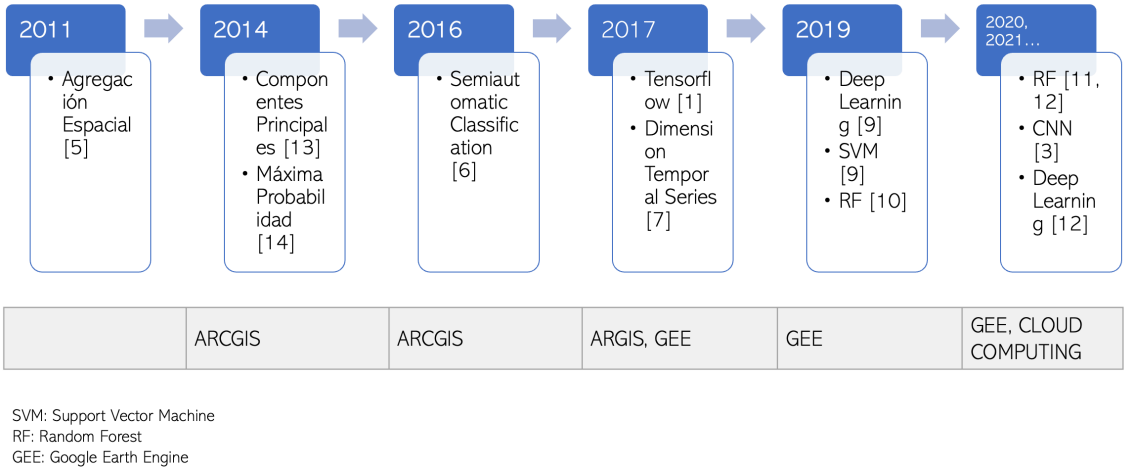


Figura 1. Línea de tiempo antecedentes históricos

A nivel nacional, en Ecuador la investigación realizada por Cartaya Ríos et al [8] en 2014, estudia la distribución de especies cinegéticas en la provincia de Manabí mediante el análisis y composición de color con los primeros componentes principales que permita distinguir coberturas y usos que del número de banda; logrando diferenciar con mayor precisión, coberturas y usos de la tierra entre el porcentaje de concomitancia de 90% y 80% para Pacoche y Flavio Alfaro respectivamente; estos porcentajes así como la nubosidad varía según la zona de la provincia.

De la misma forma en el año 2014 el Ministerio de Agricultura, Ganadería y Pesca (MAGAP) con el Ministerio del Ambiente (MAE), desarrolla un protocolo metodológico para la clasificación de la cobertura del suelo del Ecuador, utilizando imágenes satelitales obtenidas de L8 y RapidEye de una superficie de 248.983.96 km2 [15]. Utiliza las herramientas ENVI y ARCGIS para clasificar mediante el método de máxima probabilidad, por 3 razones: es reproducible por 35 técnicos, resultados similares, resultados en tiempo considerable. MAGAP, concluye la importancia de la metodología por su análisis sistemáticamente de la cobertura del suelo, resaltando que el mapa de cobertura y uso de la tierra, así como la metodología son insumos importantes para el monitoreo para prevención de la deforestación y degradación [15].

1.2 Antecedentes Conceptuales

1.2.1 Uso y cobertura del suelo

El crecimiento de la población tiene impactos significativos en la planificación territorial de gobiernos locales y estatales para satisfacer necesidades del momento y futuras [28]. Esto provoca el incremento y demanda de suelo para el establecimiento de viviendas, construcción de vías para movilidad, disponibilidad del agua, zonas agrícolas, pesca; demandas a las que se incluye la protección ambiental para la sostenibilidad poblacional y ambiental [29].

La superficie territorial puede cambiar debido a muchos factores ya sean fenómenos naturales o por intervención humana. Sucesos como lluvias, inundaciones, incendios, tormentas influye de forma directa en el cambio del suelo; así también afecta a las actividades de agricultura, ganadería, pesca o la movilidad, actividades que desarrolla el hombre [30].

La cobertura de la tierra (LC) indica los elementos físicos que ocupan la superficie de la Tierra, como el agua, bosque o estructuras urbanas [31]. De manera práctica se considera cobertura del suelo al tipo de elemento presente en la superficie de la tierra como, por ejemplo: campos agrícolas, lagos, bosques, ríos, desiertos; estos tipos se puede diferenciar mediante la categorización que existe previamente definida por niveles, en el caso de Ecuador por el MAGAP y MAE [15].

La explotación del suelo está relacionada con el término uso de la tierra (LU) que a su vez está asociado a las actividades humanas que influyen en las generaciones económicas producto del aprovechamiento del suelo [32], se refiere también al manejo al cual somete el hombre [30].

Entre otras cosas según Peña et al [28] define 4 necesidades por lo que es importante la clasificación del suelo:

1. Necesidades materiales y de energía.
2. Necesidades sociales.
3. Necesidades espirituales.
4. Necesidades de información.

En este sentido la cuarta necesidad aparece como las oportunidades que brinda la misma naturaleza para conocer aspectos bióticos, abióticos y los factores influyentes en la evolución de la tierra; mediante algún análisis metódico.

Las encuestas y muestreos eran instrumentos útiles para análisis y conocimiento de los tipos del suelo y sus usos [30] mediante levantamiento de información en campo [8]. Así también los censos e inventarios como instrumentos para evaluar la capacidad productiva del suelo. En la actualidad para precisar los tipos de cobertura y las influencias en el clima [6], desastres naturales como el incendio [19] y los diferentes usos, se lo realiza aplicando clasificadores que utilizan el aprendizaje automático (ML), apoyado de computación en la nube como GEE [3].

1.2.2 Tipos de coberturas del suelo

Como en otras disciplinas, el estudio y análisis de aspectos físicos de la tierra requiere la categorización por grupos reducidos con características homogéneas. Esto permite la

comunicación de conocimiento sobre características específicas de los fenómenos, objetos o superficies, como el uso y cobertura del suelo [28].

En 1974, Sokal [33] define clasificación, o categorización, como "el orden o disposición de objetos en grupos o conjuntos sobre la base de relaciones. Estas relaciones pueden basarse en propiedades observables o inferidas".

Ahlqvist et al [28] representa un sistema universal aplicable para la categorización de la cobertura terrestre desarrollado por la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO por sus siglas en inglés); categorizada y organizada de manera jerárquica.

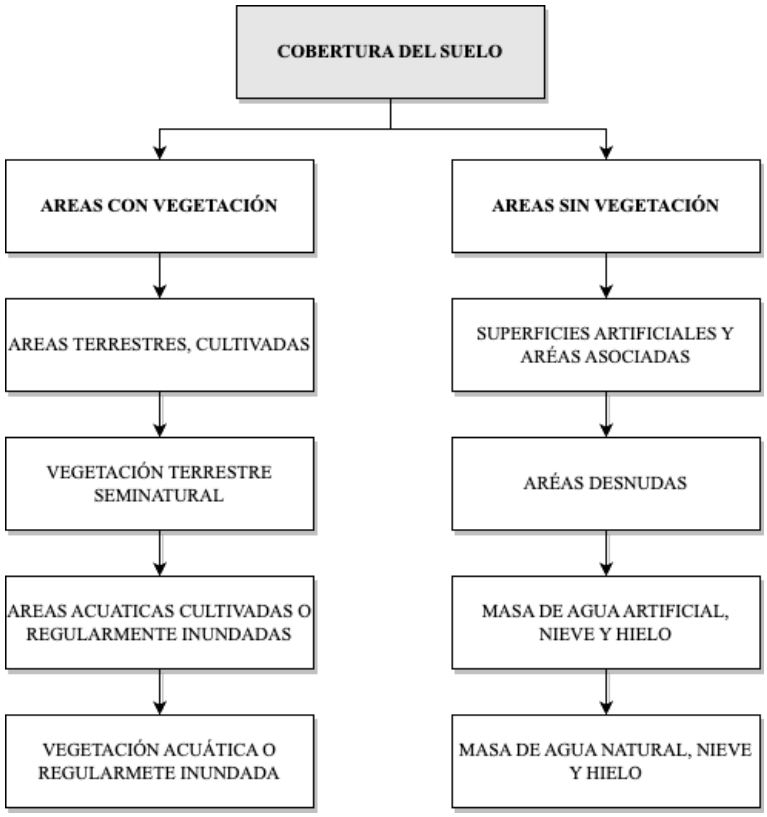


Figura 2. Categorías de cobertura terrestre de LCCS (versión 2.0).
Fuente: Ahlqvist et al [28], 2016.

La categorización presentada en la Figura 2 es un enfoque conceptual de uso de la tierra mediante análisis y recopilación de varios estudios, la agrupación tiene dos enfoques conceptuales básicos, la función: uso de tierra para propósitos comunes; y la actividad: tierras que después de un proceso da como resultado algún tipo de producto [28].

Otra clasificación de tipos de cobertura terrestre presenta Peña et al [30] en 2005, que está basado en los proyectos CORINE Land-Cover y LUCC, está distribuido de manera

jerárquica por medio de divisiones dicotómicas, se presenta 31 categorías en el segundo nivel y 8 en el primer nivel.

La clasificación presentada en la Figura 3, se puede adoptar para los mapas de usos y coberturas generales del suelo, de modo que es compatible con cualquier demanda de extracción estadística de clasificación del suelo [30]. Se distribuye a partir de dos grupos principales; agua y tierra, a partir de ello se agrupan por categorías homogéneas.

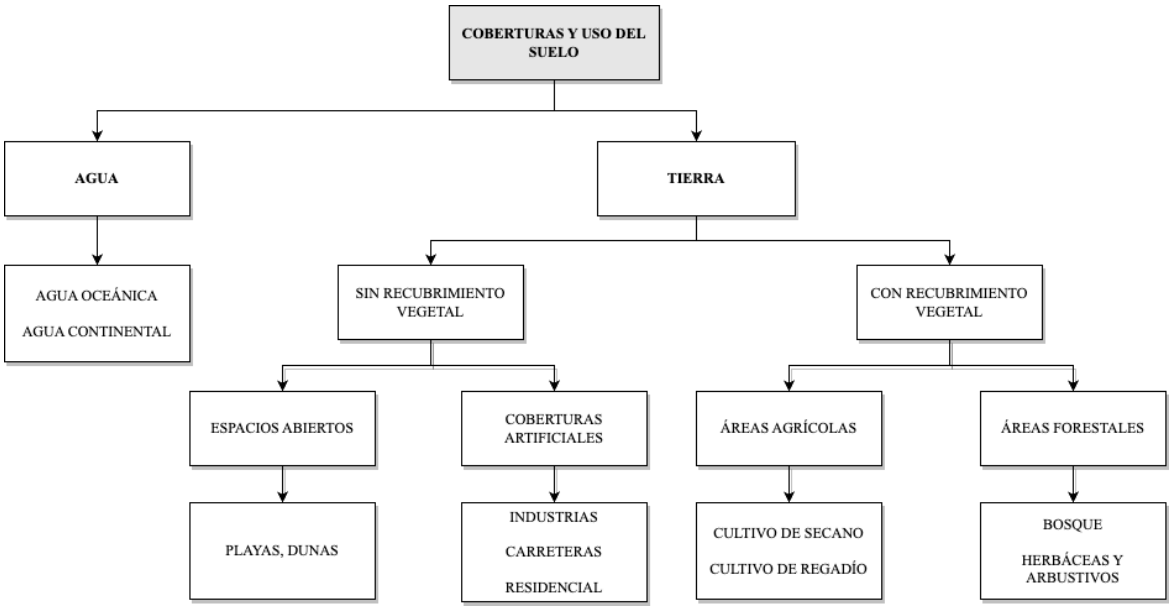


Figura 3. Clasificación jerárquica de las coberturas y usos del suelo CORINE.
Fuente: Peña et al [30], 2005.

El MAE y MAGAP en el proyecto Mapa de Cobertura y Uso de la Tierra del Ecuador Continental, en las consideraciones técnicas elabora un sistema de clasificación (leyendas temáticas) para la generación de geo-información de la cobertura y uso de la tierra, se presenta categorizado desde el nivel I hasta el IV de forma jerárquica.

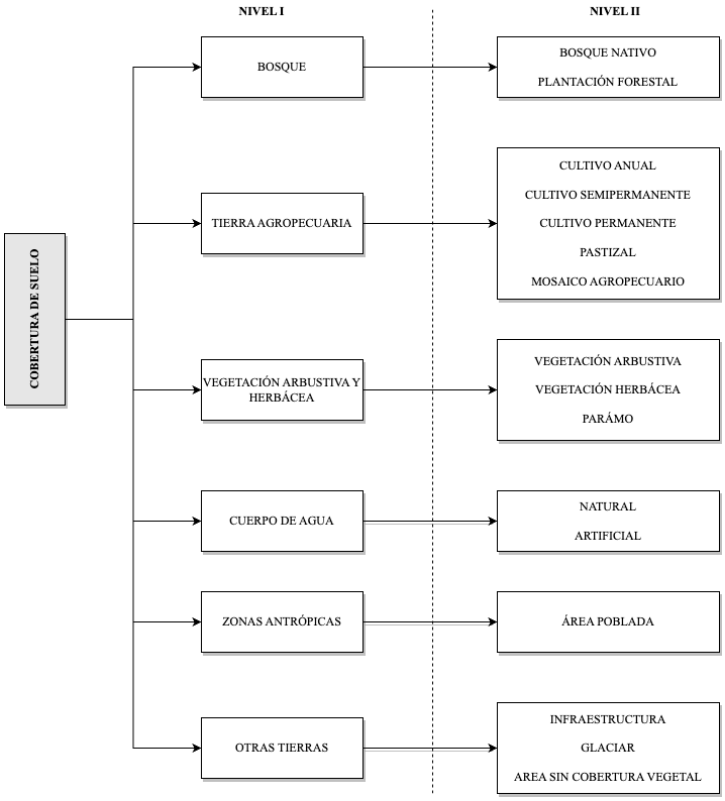


Figura 4. Leyenda temática Nivel I y II – MAGAP
Fuente: MAGAP [15], 2015.

Los niveles I y II presentados en la Figura 4, son elaborados por MAE y MAGAP-CLIRSEN y las leyendas corresponden a las clases de cobertura/uso definido por IPCC adaptado para el Ecuador. Los niveles III y IV de la Tabla 2., fueron generados únicamente por MAGAP-CLIRSEN (actualmente IEE), las leyendas III y IV son generadas por el MAGAP [15].

1.2.3 Procesamiento de imágenes satelitales

En la revisión de la literatura menciona que el tratamiento de imágenes satelitales es reciente, pero existe serie de trabajos que tienden a utilizar este tipo de información para distintos fines, entre ellos el análisis y clasificación de la cobertura terrestre/suelo.

"*Las imágenes digitales de satélite son conjuntos de datos rasterizados, lo que significa sencillamente que la imagen está comprimida en numerosos y diminutos elementos de imagen o píxeles que cubren la totalidad del área de la escena. Los conjuntos de datos vectoriales, por contraste, son mucho más abstractos y están compuestos por puntos, líneas y polígonos*" [34].

El ámbito de procesamiento de digitales se refiere al procesamiento mediante la computadora digital. Hay que tener en cuenta que una imagen está compuesta por miles de píxeles. En este sentido el campo de aplicación es bastante amplio, una de ellas es la aplicación de la visión por computadoras para simular el comportamiento humano [35], aplicando técnicas de la Inteligencia Artificial (IA) o ML.

Gonzales RC y Woods RE [35] propone 3 niveles de procesamientos:

- Nivel bajo, mejoramiento de características como el ruido, contraste o nitidez.
- Nivel medio, implica segmentación por regiones u objetos y su descripción, para reducir de forma óptima para procesamiento informático. Los resultados son bordes, contornos u objetos individuales.
- Nivel alto, implica dar sentido a un conjunto de objetos reconocidos, realizar una visión cognitiva próxima a la visión humana.

1.2.4 Componentes para procesamiento de imágenes

El procesamiento de imágenes digitales demanda altos recursos computacionales [12] y software altamente especializado [35]. Por lo que, es necesario definir componentes necesarios para llevar a cabo estos procesos.

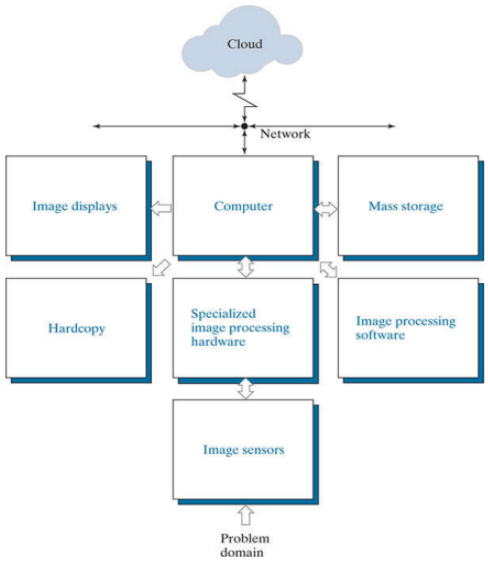


Figura 5. Components of a general-purpose image processing system
Fuente: Gonzales RC y Woods RE [35], 2018.

En el esquema presentado en la Figura 5 corresponde a Gonzáles RC y Woods RE [35]. Se define como fuente o entrada un sensor físico (Image sensors) de imagen como las

cámaras o satélites, estos transfieren hacia un digitalizador (Specialized image processing hardware), este cumple con la función de convertir las salidas del sensor físico a digital.

La información obtenida desde el digitalizador es llevada a un computador como se mencionó antes con altas prestaciones en temas de memorias, procesador y unidad de procesamiento gráfico (GPU) [35]. El tratamiento se lleva a cabo mediante software como Matlab [35], QGIS [34], herramientas de tratamiento de imágenes ARCGIS [6] o en entornos de desarrollo y producción como GEE [3], [19], [20]. Las herramientas mencionadas corresponden al software de procesamiento de imagen (Image processing software) y al almacenamiento masivo (Mass storage).

El componente de copia impresa (Hardcopy) son dispositivos sensibles al calor como los discos ópticos o el CD-ROM. Otro componente es la presentación de la imagen procesada para ello se puede emplear herramientas de proyección digital o algún software especializado.

Por el hecho de requerir grandes demandas de equipos de cómputo se puede procesar en la nube, que en la actualidad ofrece grandes prestaciones a nivel de información, infraestructura y entornos de desarrollo para el tratamiento de imágenes como es el caso de la plataforma GEE [7].

1.2.5 Imágenes satelitales

Las imágenes satelitales proporcionan información más detallada para el análisis y procesamiento. Revisado la literatura, en los estudios realizados a través del tiempo en la actualidad se emplean el histórico de datos que proporcionan los satélites Landsat [1] y Sentinel [26] en sus distintas versiones. Aquí también se involucra la computación en la nube, como AWS o GEE [3], [9], [19] que proporciona data almacenada y plataformas para el análisis.

1.2.6 Imágenes satelitales Landsat

Landsat es un programa conjunto estadounidense entre la NASA y el Servicio Geológico de los Estados Unidos (USGS), ofrece imágenes de alta resolución registro global continuo de la Tierra. Se origina en la década de 1970 [36], iniciando exactamente en 1972 con la primera misión; en la actualidad se encuentra en la novena misión [37], [38].

En la revisión de estudios [6], [19], [39] se ha encontrado la utilización de imágenes obtenidos de Landsat 7 y 8 para diferentes fines, entre ellas la clasificación de coberturas y uso del suelo [6]; por lo que es de interés conocer las características de las imágenes.

El tamaño de la escena del Landsat 8 es de 185 km a lo largo de la pista por 180 km a lo largo de la pista. La altitud nominal de la nave espacial es de 705 km. Los productos de datos Landsat 8 requieren una precisión cartográfica de 12 m o superior [40].

Landsat-7 ETM+ Bands (μm)			Landsat-8 OLI and *TIRS* Bands (μm)		
			30 m Coastal/Aerosol	0.435 - 0.451	Band 1
Band 1	30 m Blue	0.441 - 0.514	30 m Blue	0.452 - 0.512	Band 2
Band 2	30 m Green	0.519 - 0.601	30 m Green	0.533 - 0.590	Band 3
Band 3	30 m Red	0.631 - 0.692	30 m Red	0.636 - 0.673	Band 4
Band 4	30 m NIR	0.772 - 0.898	30 m NIR	0.851 - 0.879	Band 5
Band 5	30 m SWIR-1	1.547 - 1.749	30 m SWIR-1	1.566 - 1.651	Band 6
Band 6	60 m TIR	10.31 - 12.36	*100 m TIR-1*	*10.60 – 11.19*	Band 10
			100 m TIR-2	*11.50 – 12.51*	Band 11
Band 7	30 m SWIR-2	2.064 - 2.345	30 m SWIR-2	2.107 - 2.294	Band 7
Band 8	15 m Pan	0.515 - 0.896	15 m Pan	0.503 - 0.676	Band 8
			30 m Cirrus	1.363 - 1.384	Band 9

Tabla 1. Bandas Landsat-7 y Landsat-8
Fuente: https://landsat.gsfc.nasa.gov/landsat-8

La carga útil del satélite L8 consta de dos instrumentos científicos: el Operational Land Imager (OLI) y el Thermal Infrared Sensor (TIRS). Estos dos sensores proporcionan una cobertura estacional de la masa terrestre global con una resolución espacial de 30 metros (visible, NIR, SWIR); 100 metros (térmica); y 15 metros (pancromático).

El OLI recopila datos para dos nuevas bandas, una banda costera / de aerosoles (banda 1) y una banda de cirros (banda 9) como se representa en la Tabla 3, así como las bandas multiespectrales de Landsat heredadas. Además, el ancho de banda se ha refinado para seis bandas. El instrumento térmico (TIRS) lleva dos bandas infrarrojas térmicas adicionales [40], [41].

1.2.7 Imágenes satelitales Sentinel

Lanzado en 23 de junio de 2015 [42], Sentinel cuenta con seis misiones: Sentinel 1-6 [43] un proyecto de la Comisión Europea (CE) y la Agencia Espacial Europea (ESA). Es un satélite que proporciona observaciones de la Tierra para los servicios operativos del proyecto europeo Capernicus [44].

La misión 1 fue lanzada en 3 de abril de 2014 para captura de imágenes diurna y nocturna para servicios terrestres y oceánicos. La misión 3 fue lanzada en febrero de 2016 para apoyo a los sistemas de pronósticos oceánicos para capturar topografías de la superficie del mar. Las misiones 4 y 5 están enfocados en el monitoreo atmosférico. La misión 6 y última fue lanzada en 21 de noviembre de 2020 para mediciones de altura global de la superficie del mar principalmente para oceanografía y estudios climáticos [39].

SENTINEL 2.

La misión 2 (Sentinel 2 – S2) fue lanzado el 23 de junio de 2015 tiene como objetivo la generación de imágenes multiespectrales de alta resolución para el monitoreo de la tierra, generar imágenes de vegetación, cubierta de suelo y agua, áreas costeras, áreas de navegación [45]. Sentinel 2 incorpora un generador de imágenes multiespectral con resolución de 10m por píxel [42] de 13 bandas cubierta por longitud de onda infrarroja cercana y visible (VNIR, por sus siglas en inglés) y por infrarrojo de onda corta (SWIR). Las resoluciones de estas bandas son de 10 y 60 m con cobertura de -54 y +84 de latitud con franja de 290km de ancho [35].

Sentinel 2 entre sus principales objetivos está el proporcionar datos de observación para la generación de productos operativos, con mapas de cobertura terrestre, mapas de detección de cambios terrestres y variables geofísicas. Por ende contribuye en temas como cambio climático, monitoreo de la tierra, gestión de emergencias y seguridad [46].

A continuación, se presenta la composición de bandas de Sentinel 2.

Sentinel 2 - Banda	Longitud onda central (µm)	Resolución (m)
1: Aerosol costero	0.443	60
2: Azul	0.490	10
3: Verde	0.560	10
4: Rojo	0.665	10
5: Vegetación borde rojo	0.705	20
6: Vegetación borde rojo	0.740	20
7: Vegetación borde rojo	0.783	20
8: NIR	0.842	10
8A: Vegetación borde rojo	0.865	20
9: Vapor de agua	0.945	60
10: SWIR – Cirrio	1.375	60
11: SWIR	1.610	20
12: SWIR	2.190	20

Tabla 2. Sentinel 2 características de las bandas
Fuente: Kaplan G, Avdan U [47]

La constelación de L8 y satélites ópticos S2 poseen un gran potencial para ser utilizados sinérgicamente para una variedad de aplicaciones de observación de la Tierra debido a sus propiedades espectrales y espaciales similares y al acceso libre y abierto a datos [48], ya que la combinación de los sensores de L8 y S2 proporciona una cobertura global multiespectral de 10 a 30 m aproximadamente cada 3 días [49]. Entre otras aplicaciones están las técnicas y análisis de clasificación de la cobertura terrestre [48].

Como se ha encontrado en los estudios previos las imágenes satelitales más explotadas son de Landsat y Sentinel. Por lo que en este estudio se considera la utilización de imágenes satelitales Sentinel 2 debido a la cantidad de bandas que permite el procesamiento de datos para discriminar el tipo de cobertura terrestre.

1.2.8 Aprendizaje Automático

El aprendizaje automático (ML) es una de las ramas de la informática, es la ciencia de programar computadoras con altas capacidades para que puedan aprender, producir y predecir información a partir de datos.

A continuación, se presenta alguna de las definiciones generales:

- El aprendizaje automático es el campo de estudio que brinda a las computadoras la capacidad de aprender sin estar programado explícitamente [50]. - Arthur Samuel, 1959.

- Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna tarea T y alguna medida de rendimiento P, si su desempeño en T, medido por P, mejora con la experiencia E [51]. Tom Mitchell, 1997.

Las primeras aplicaciones de ML aparecen en el tratamiento de correos spam mediante palabras y frases predictoras presentes usualmente en el contenido de los mensajes. Gerón [51] otras aplicaciones como:

- Detectar tumores en escáneres cerebrales.
- Clasificación automática de artículos de noticias.
- Crear un chatbot o un asistente personal.
- Hacer que su aplicación reaccione a los comandos de voz.

De la misma forma como se ha descrito en la revisión histórica de ML, otra de las aplicaciones es el análisis y uso de la tierra [5], estudios de espacios afectados por incendios [19], el impacto de cambio de las coberturas terrestres en el clima [6], clasificación de la cobertura del suelo [1], [3], [7], [10]. Estas aplicaciones básicamente tienen como principales puntos destacables la aplicación de clasificadores de ML y la computación en la nube.

TIPOS

Para la clasificación del sistema de ML se considera criterios como:

- Capacidad para aprender con o sin supervisión humana.
- Si pueden aprender de manera progresiva.
- Capacidad para detectar patrones en los datos de entrenamiento y construcción de modelos predictivos.

En el primer criterio se tiene los tipos de ML supervisado y no supervisado que a su vez se subdivide en 4 categorías según la cantidad y el tipo de supervisión requerida [51], estas son: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje reforzado.

El aprendizaje no supervisado no considera los datos históricos, no posee datos de entrenamiento [51], por lo tanto, no requiere de una muestra de aprendizaje con clasificación previa [49]. Es decir, aprende por sí mismo, consiste en encontrar la partición más adecuada del conjunto de entradas [52]. Considerando lo expuesto por [51], [52] en este estudio no se aborda la clasificación no supervisada, más bien se considera el tipo de ML supervisada.

APRENDIZAJE AUTOMÁTICO SUPERVISADO

El aprendizaje supervisado considera la introducción de datos para el entrenamiento de manera etiquetada [51], que dado una o varias características, mediante el entrenamiento proporciona predicciones y etiquetas [51]. Las aplicaciones más típicas en este tipo de aprendizaje son las clasificaciones [51], técnicas de predicción y comprender nuevas instancias y descubrir información desconocida de interés [53].

- k-Nearest Neighbors (k-Vecinos más cercanos)
- Linear Regression (Regresión lineal)
- Logistic Regression (Regresión logística)
- Support Vector Machines (SVMs) (Máquina de soporte vectorial)
- Decision Trees and Random Forest (Árboles de decisión y bosques aleatorios)
- Neural Networks (Red neuronal)

Para el presente estudio se han seleccionado los algoritmos: Máquina de soporte vectorial [18], Bosques aleatorios [17] y CART [54]. Por otra parte, el estudio se realiza bajo plataforma GEE lo que apoya la computación y procesamiento en nube. Sin embargo, las CNN mencionadas en las secciones anteriores no se consideran debido a requerimientos de infraestructura tecnológica para el procesamiento y almacenamiento de imágenes, ya que se requiere sistemas computacionales con características muy altas.

1.2.9 Máquina de Soporte Vectorial (SVM)

Desarrollado en 1991 por Vladmir Vapnik, Maquina de Soporte Vectorial (SVM, por sus siglas en inglés) es una modelo de aprendizaje potente y versátil capaz de realizar clasificaciones lineales, no lineales, regresión o incluso detección de valores atípicos [51], con mejor desempeño que las modelos tradicionales como las redes neuronales [55].

SVM está basada en algoritmos que aprenden a partir de una serie de datos de muestras estocásticas para clasificar o predecir [56]. Es una herramienta de clasificación que permite encontrar fronteras lineales a partir de no lineales mediante la transformación del espacio al que pertenece los datos a un espacio de dimensión mayor [57].

DEFINICIÓN

Una SVM mapea los puntos de entrada a un espacio de características de una dimensión mayor trazando un plano que divide en fronteras de decisión, los puntos están en R^2 pero el alto grado de clasificación permite extender a un plano R^3 [56], [57], [55].

Durante la tarea de clasificación se tiene dos fases [58]:

Fase 1. Aprendizaje automático: Se selecciona conjuntos de datos de entrenamientos para extraer atributos y características con las que se entrena el clasificador. E resultado del

entrenamiento es un conjunto de parámetros llamados pesos que define el clasificador en función de la discriminante que representa las fronteras de clases o regiones [58].

Fase 2. De reconocimiento: El modelo del clasificador entrena y asigna nuevos datos de entrada a las clases según la similitud de sus características [58].

FUNCIONAMIENTO

El aprendizaje consiste en la construcción de un hiperplano de separación optima (HSO) que está definido por el margen máximo de separación entre las clases.

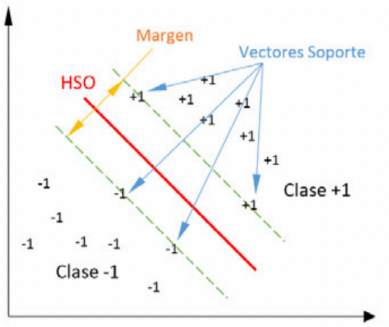


Figura 6. Hiperplano de Separación Óptimo para SVMs.
Fuente: González et al [58].

Siendo:

$\vec{x}_i \in \Re^n$	Las clases de entrenamiento: Clase -1 y Clase +1 son
$\vec{w}\cdot\vec{x}+b=+1$ $\vec{w}\cdot\vec{x}+b=-1$	Muestras fronterizas para las clases (color verde)
$\vec{w}\cdot\vec{x}+b=0$.	Hiperplano de Separación Optima – HSO (color rojo)
$2/\|\|\vec{w}\|\|$	Margen entre los planos (color naranja). Para que HSO clasifique de manera correcta se debe imponer restricciones para determinar las clases "-1" y "+1".

Tabla 3. Definiciones de Maquina de Soporte Vectorial

En la Figura 6, HSO representa el hiperplano de separación optima entre las clases -1 y +1, además las líneas entrecortadas en color verde definen las muestras fronterizas para las clases. El margen se define entre los hiperplanos paralelos y el HSO [58].

El caso descrito es el modelo para caso de datos linealmente separables, pero las aplicaciones reales casi nunca son lineales, además se debe considerar que se puede introducir datos muy atípicos. En consecuencia, encontrar HSO no es tarea trivial.

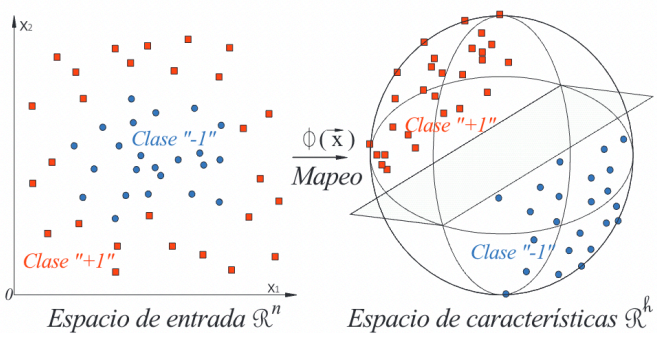


Figura 7. Mapeo de datos a espacio de mayor dimensión SMVs.
Fuente: González et al [58].

En el caso de datos no lineales los datos de entrada son mapeados a un espacio de dimensión mayor utilizando función llamada kernel o también denominado el truco de kernel [55]. En la Figura 7 se muestra el espacio de entrada donde no se puede obtener HSO debido a la distribución no línea, entonces mediante la aplicación de kernel se amplía el plano que permita encontrar HSO para la separación de clases.

Las funciones kernel que utiliza para SVM son: lineal, polinomial y gaussiana; siendo esta ultima la que se utiliza en los modelamientos de clasificación que permite separar las clases de entrenamiento.

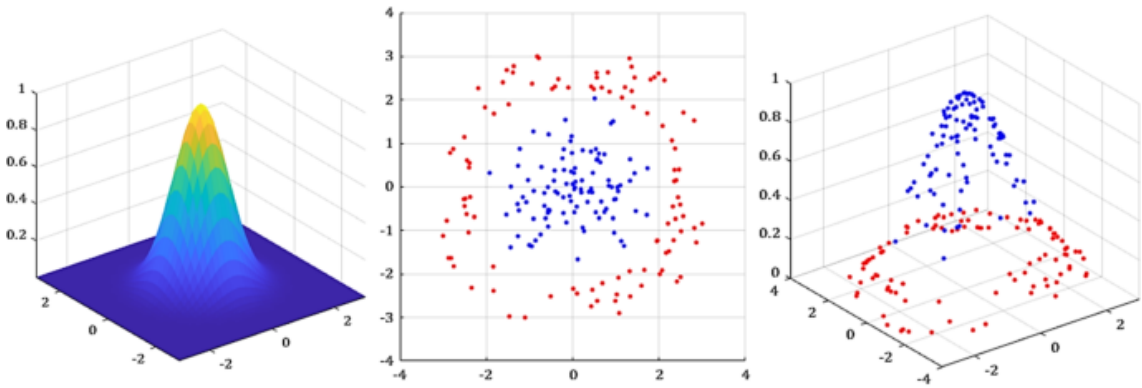


Figura 8. Kernel Gaussiano SVMs.
Fuente: A. Elen, S. Baş, y C. Közkurt (2022) [59].

1.2.10 Arboles de clasificación y regresión (CART).

Propuesta desarrollada por Leo Breiman en 1984, Classification And Regression Tress (CART, por sus siglas en inglés), consiste en un conjunto de reglas que calcula fronteras de decisión y mediante un procesos obtiene particiones o grupos homogéneos que permite determinar la categoría a la que pertenece un dato de entrada respecto a una variable discriminante [54] [60]. CART se puede usar para regresión o problemas de clasificación, en el presente estudio se utiliza para clasificación.

El proceso consiste en 3 pasos [61] [60].

Paso 1. Construcción del árbol máximo: En el primero paso se construye un árbol binario partiendo desde el nodo raíz de donde se incrementa los niveles según los características o variables de discriminación.

En este proceso se evalúa el grado de impureza del nodo con el índice Gini, el valor 0 del índice significa nodos puros (pertenece a una sola categoría), mientras el indicé mayor a 0 son nodos con impureza (datos con más de una categoría).

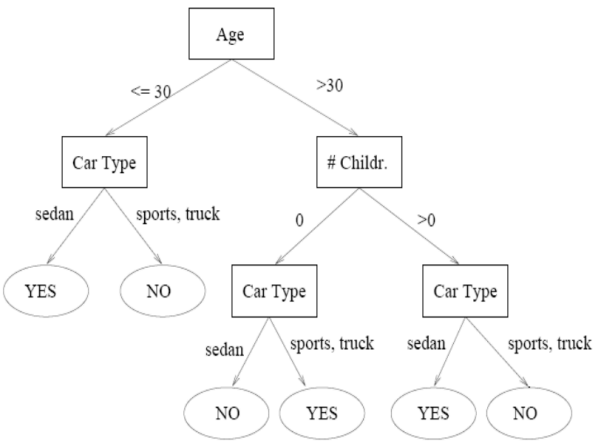
$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Donde: Pi es la probabilidad de un variable que clasifique en una clase particular.

Paso 2. Poda del árbol: El árbol obtenido es sobre ajustado cortando sucesivamente los nodos terminales hasta encontrar su tamaño adecuado.

Paso 3. Selección del árbol optimo: Se realiza mediante un proceso de comparación o penalización de complejidad estimando el error de clasificación, procedimiento se realiza mediante la validación cruzada.

A continuación, en la Figura 9 se presenta un ejemplo de árbol de clasificación si un conductor vive o no en los suburbios.



Car type	Age	Chd	Sub
sedan	23	0	yes
sports	31	1	no
sedan	36	1	no
truck	25	2	no
sports	30	0	no
sedan	36	0	no
sedan	25	0	yes
truck	36	1	no
sedan	30	2	yes
sedan	31	1	yes
sports	25	0	no
sedan	45	1	yes
sports	23	2	no
truck	45	0	yes

Figura 9. Ejemplo CART
Fuente: Pineda (2009) [60]

1.2.11 Arboles aleatorios (RF)

Desarrollado por Leo Breiman en 2001 [17], los bosques aleatorios o Random Forest (RF, por sus siglas en inglés) es una técnica de aprendizaje automático supervisado. Se construye un conjunto de árboles basado en una muestra de datos de entrenamiento y se promedian con una predicción para obtener el modelo único [62] [63].

Cada árbol del conjunto se obtiene en dos etapas [62]:

1. Genera un número considerable de árboles de decisión con el conjunto de datos.
   m < M, donde cada árbol contiene un subconjunto de muestras, donde:
   m: predictores y,
   M: total predictores.
2. Cada árbol crece hasta su máxima extensión.

Un nodo raíz, un conjunto de nodos internos y nodos hoja (nodos finales) conforman el árbol de decisión. Los nodos intermedios representan etapas de decisión y los nodos finales la clasificación final [64].

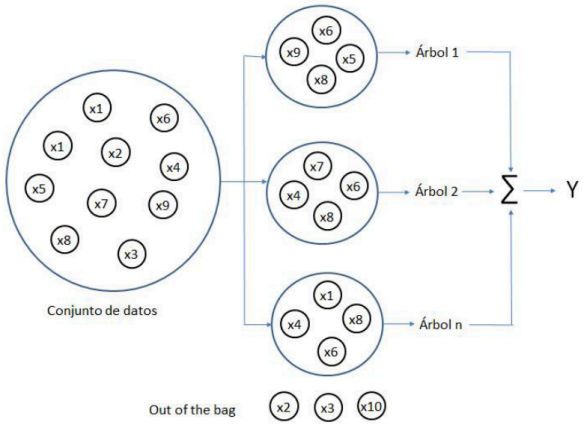


Figura 10. Ejemplo Random Forest
Fuente: Espinosa-Zúñiga (2020) [62].

Cada árbol generado contiene un grupo de observaciones aleatorias elegidas mediante la técnica estadística boostrap. Esto significa que una observación se puede considerar en una o varias muestras [63], como se muestra en la Figura 10 donde x6 incluye en los 3 árboles. Cuando llega al nodo final (hoja) asigna la etiqueta, clase que corresponde al objeto clasificado.

Random Forest es una técnica que se puede usar para clasificación o regresión, para el caso del estudio se utiliza para clasificar, donde cada árbol asigna una clase siendo el resultado la etiqueta o clase con mayor número de asignaciones [62].

RF, es un modelo simple de entrenar con altos rendimientos y desempeños eficientes con grandes cantidades de datos y cientos de predictores sin excluir ninguna [62].

1.2.12 Computación en la nube y GEE

Las últimas décadas se ha reforzado los mecanismos de uso y explotación de recursos computacionales de manera centralizada, esto tiene que ver la utilización de recursos desde grandes infraestructuras de sistemas informáticos y de almacenamiento a través de internet, en lugar de una maquina local [65].

La computación en la nube es el camino por donde empresas como Amazon, Apple, Google, HP, IMB, Microsoft entre otros, han optado de manera comercial o libre ofrecer servicios de recursos hardware y software ofreciendo acceso mediante clientes ligeros como navegadores, apis.

Google Earth Engine, GEE.

Para el presente estudio se ha seleccionado la plataforma Google Earth Engine (GEE, por sus siglas en ingles) que ofrece infraestructura y almacén de datos en la nube para científicos, investigadores y desarrolladores de manera gratuita para uso académico y de investigación [66].

En GEE haciendo uso de las herramientas Code Editor y Google Colab se puede utilizar grandes almacenes de datos (petabytes) de imágenes capturados por satélites de todo el planeta tierra para detectar cambios, mapear tendencias, cuantificar diferencias etc. En el caso de este estudio se utiliza para la clasificación de cobertura del suelo.

1.2.13 Metodología CRISP DM

El procesamiento de grandes volúmenes de datos para soporte en toma de decisiones cada vez toma alta importancia en las estrategias institucionales o empresariales.

La ciencia de datos es una disciplina para obtener información valiosa mediante la aplicación de técnicas y modelos matemáticos, se pueden beneficiarse mediante la buena

gestión de proyectos y procesos. Sin embargo, seguir de manera estricta una metodología es un desafío. En esto el modelo CRISP-DM y la ciencia de datos mejoran de manera significativa incluso en enfoques ágiles [67].

Proceso Estándar Intersectorial Para Minería De Datos (CRISP-DM, por sus siglas en inglés) es un marco de trabajo para proyectos con minería de datos aplicable independientemente de la área o tecnología [68]. Consta de 6 fases iterativas desde la comprensión del problema hasta el despliegue.

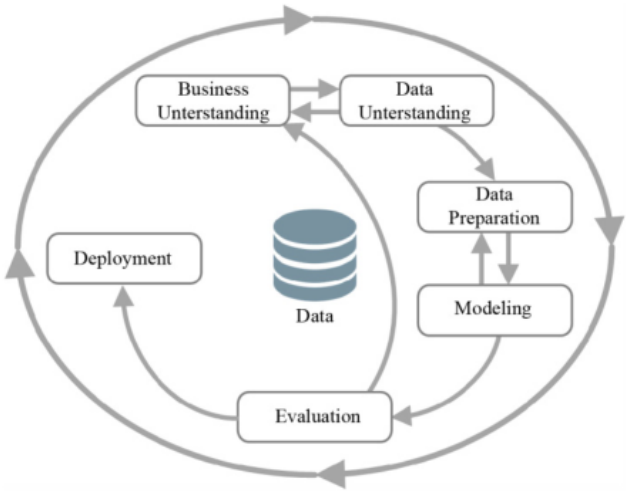


Figura 11. El proceso CRISP-DM
Fuente: Huber S et al (2019) [68]

COMPRENSIÓN DEL NEGOCIO (PROBLEMA)

En la fase 1, se evalúa el estado de la situación actual para obtener la visión general del negocio, recurso disponibles y necesarios; involucra el nivel del conocimiento del problema. A partir de esto se determina el objetivo, beneficios y al final se describe el plan del proyecto que describa la tecnología y herramientas a utilizar [67], [69].

COMPRENSIÓN DE DATOS

La fase 2 inicia con la recopilación de datos, su exploración y descripción y la verificación de la calidad de los datos como tareas esenciales en esta fase. A demás, se describe los datos mediante estadística o la determinación de atributos y sus interrelaciones [67]. La recopilación puede ser desde plataformas informáticas [69].

PREPARACIÓN DE DATOS

La fase 3 comprende la integración y limpieza de los datos mediante los criterios de inclusión y exclusión, eliminación de inconsistencias. En dependencia de la fase anterior se construye el dataset con las características y sus derivados [67]–[69].

MODELADO

La fase 4 es el flujo de trabajo mediante procesamiento de datos que se realiza utilizando técnicas y modelos ML. Probablemente se construye y evalúa varios modelos, es el punto donde se emplea los escenarios de entrenamiento y prueba hasta obtener el modelo adecuado u optimo. Se puede regresar a la fase 3 si fuera necesario [67], [70].

EVALUACIÓN

En la fase 5 se evalúa el modelo entrenado con un conjunto de datos reales. Es la evaluación de los resultados frente a los objetivos empresariales definidos. Es la tarea para verificar el cumplimiento de criterios de éxito empresarial, de manera que el modelo sea aceptado. A demás de revisar que todos los pasos previos hayan cumplido de manera correcta [67].

DESPLIEGUE

La fase 6 y final es la explotación de resultados mediante la visualización en un software para soporte de decisiones, o un mecanismo para acceso a resultados. Esto requiere la implementación de infraestructura puesta en marcha del modelo, supervisión y mantenimiento [67].

1.3 Antecedentes Contextuales

En Ecuador el Consejo Técnico de Uso y Gestión del Suelo en su Resolución Nro. 003-CTUGS-2019, expide la "NORMA TÉCNICA PARA EL PROCESO DE FORMULACIÓN O ACTUALIZACIÓN DE LOS PLANES DE DESARROLLO Y ORDENAMIENTO TERRITORIAL DE LOS GOBIERNOS AUTÓNOMOS DESCENTRALIZADOS". En el artículo 1, dentro de su ámbito de competencia, faculta la utilización de los instrumentos de desarrollo y ordenamiento territorial a cualquier régimen.

En la misma resolución, el Capítulo II, Sección I, Artículo 7, literal a, párrafo primero, avala el desarrollo de los contenidos:

a) *"Modelo de territorial actual*.- debe evidenciar las potencialidades y los problemas; y su relación sobre la red de asentamientos humanos caracterizados y la clasificación del suelo, zonas de importancia para la conservación, cuencas y microcuencas, zonas de riesgo y aquellas con amenazas climáticas, complementado por la relaciones multinivel y vinculaciones con los GAD circunvecinos, proyectos nacionales de carácter estratégico (en caso de existir), así como las redes de infraestructura logística, transporte, movilidad, accesibilidad, energía, telecomunicaciones, áreas de explotación de recursos naturales, entre otros elementos importantes que fueron identificados en el diagnóstico estratégico y que varían según el nivel de gobierno y sus competencias" [71].

En el mismo contexto MUDUVI [15] en el Artículo 27, Plan de uso y gestión del suelo (PUG), resalta la importancia de para los GAD contar con la información "del límite urbano existente, los datos catastrales actualizados, mapas de valoración de suelo, planes maestros de redes de infraestructura: de agua potable, alcantarillado, energía eléctrica, componente vial, red de espacios públicos, equipamientos de salud, educación, sociales y áreas verdes; el esquema y jerarquización vial, sistema de movilidad y transporte público, datos de densidad poblacional en el territorio, entre otros".

Considerando avales legales y constitucionales mencionados, en el presente estudio se considera la provincia de Chimborazo del Ecuador como área de estudio para la clasificación de la cobertura del suelo, esto contribuirá al análisis, determinación y conocimiento de áreas de interés mencionadas en el Artículo 27 de PUG y colaborar el cumplimiento de análisis y valoración de planes estratégicos y de sostenibilidad mencionados emitidas por el Consejo Técnico de Uso y Gestión del Suelo [71].

En el ámbito de clasificación de la cobertura del suelo, el estudio realizado a nivel de país en 2014 por MAE y MAGAP determina los tipos de cobertura en su mayoría cubierta por pastizal y páramo. Además, se puede resaltar la existencia de bosques y en menor medida los cuerpos de agua y zona antrópica [15], [72].

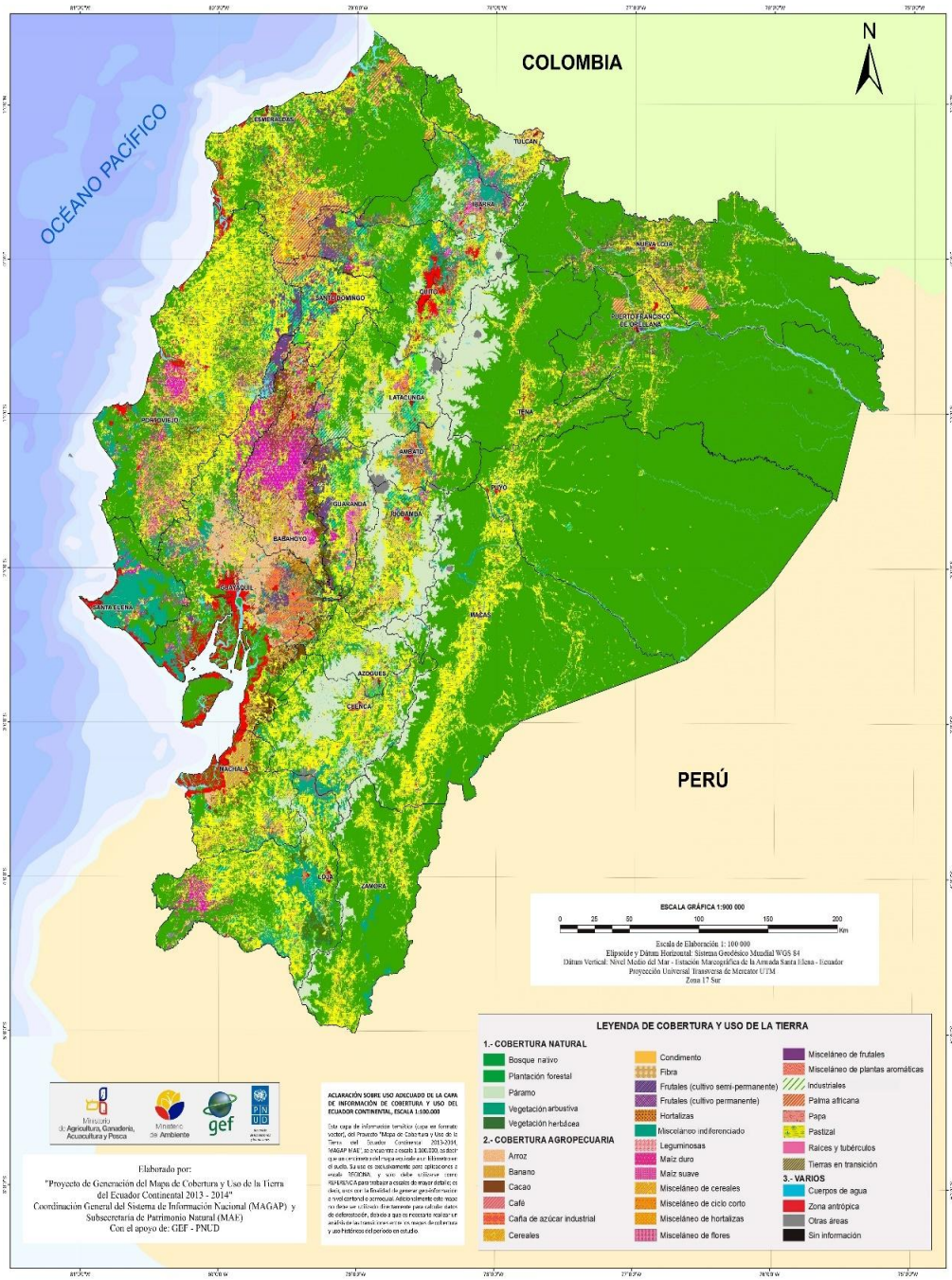


Figura 12. Mapa de cobertura y uso del suelo del Ecuador MAE, MAGAP
Fuente: MAE, MAGAP, 2014 [72]

También se ha podido clasificar cinco tipos de cobertura: agua, no vegetativo, vegetativo, nieve y árboles a partir de imágenes satelitales obtenidos de S2 [7].

La provincia de Chimborazo se encuentra ubicada en la zona centro de Ecuador, cuenta con una superficie territorial de con una extensión jurisdiccional de 6.578,10 Km2, dividido políticamente en 10 cantones y 45 parroquias rurales [14]. A continuación, se presenta información general del área de estudio.

Dato	Descripción
Nombre del GAD	Gobierno Autónomo Descentralizado de la Provincia de Chimborazo
Fecha de creación de la provincia	25 de junio de 1824
Población	524.004 habitantes (proyección INEC 2.020)
Extensión	La provincia de Chimborazo, ubicada en el centro del Ecuador, con una extensión jurisdiccional de 6.578,10 Km2 (GADPCH 2.019), políticamente se subdivide en 10 cantones y 45 parroquias rurales (INEC-2.010).
Rango altitudinal	El rango altitudinal de la provincia de Chimborazo desde los 135 m.s.n.m. a 6310 m.s.n.m.
Nacionalidad	Kichwa; Pueblo Puruwá
Límites	Norte: Provincia de Tungurahua. Sur: Provincia de Cañar Este: Provincia de Morona Santiago Oeste: Provincia de Bolívar y Guavas

Tabla 4. Datos generales de la provincia de Chimborazo
Fuente: Prefectura de Chimborazo, 2020 [14]

La clasificación de la cobertura del suelo de la provincia de Chimborazo será una herramienta de soporte para el análisis y toma de decisiones para el establecimiento y ejecución de planes estratégicos, desarrollo, innovación, movilidad, urbanización, conservación ambiental y otras factores de interés; ya que el desarrollo económico y ecológico implica el conocimiento de los tipos del suelo para ser aprovechados de manera responsable en beneficio de las familia y la sociedad riobambeña y del país.

CAPITULO II

2. MATERIALES Y MÉTODOS

2.1 Tipo de estudio o investigación

Por el tipo de estudio que se realiza en la investigación se ha considerado 3 tipos: exploratoria, descriptiva y correlacional. De acuerdo con Sampieri et al [73], esto es posible debido al análisis y estrategias de la investigación.

EXPLORATORIA

En el Capítulo I, mediante la revisión sistemática de la literatura (SLR, por sus siglas en inglés) se ha expuesto que los estudios relacionados al tratamiento de imágenes satelitales y aprendizaje automático son temas nuevos y aun joven en al área de la TIC.

El estudio exploratorio ha permitido determinar la escaza información digital de la clasificación de cobertura terrestre en la provincia de Chimborazo, también la búsqueda y levantamiento fuentes de datos con imágenes satelitales desde repositorio GEE.

DESCRIPTIVO

En el contexto de la investigación el estudio exploratorio ha permitido definir 6 clases de coberturas del suelo fundamentado en los estudios [15], [30], [28]. Se define el dataset y recolecta los datos de entrenamiento y prueba del área geográfica de la provincia de Chimborazo que se utiliza para el modelamiento mediante ML. La descripción de tipos de cobertura se aborda en la 2.4. Métodos Teóricos, fase 2 de la metodología CRISP-DM.

CORRELACIONAL

El estudio correlacional permite determinar la correlación de bandas de las imágenes satelitales y su asociación para aplicación del algoritmo ML. El presente estudio describe las correlaciones de bandas de imágenes satelitales Sentinel 2 para la composición de dataset que permite la discriminación de clases de coberturas terrestres.

Se ha seleccionado las bandas B2, B3, B4 que corresponde a los colores azul, verde y rojo respectivamente. Esta correlación se traduce al sistema de composición de colores primarios RGB, para representación de cualquier imagen computacional.

La correlación de las bandas B2, B6 (vegetación borde rojo), B8 (NIR), B11 (SWIR, infrarrojo de onda corta) permiten la detección de zonas agrícolas que son representados en tonalidad verde brillante.

Las combinaciones de las bandas (NIR - RED) / (NIR + RED) representa el índice de vegetación diferencial normalizada (NDVI, por sus siglas en inglés), permite la detección de áreas de vegetación, forestal y zonas agrícolas. Las combinaciones de bandas espectrales (SWIR + RED) - (NIR - BLUE) / (SWIR + RED) + (NIR + BLUE) representa el índice de suelo desnudo (BSI, por sus siglas en ingles), permite determinar la composición de mineral de suelo y detección de áreas desérticas o superficies desnudos.

La utilización de las bandas y sus composiciones es necesario para la clasificación de coberturas, esto permite discriminar los tipos basados en pixeles de colores según las etiquetas que se asigna. La construcción del dataset completo se aborda 2.4. Métodos Teóricos, fase 2 y 3 de la metodología CRISP-DM.

2.2 Paradigma de la investigación

Se ha empleado el enfoque del paradigma cuantitativo para probar la hipótesis en base a mediciones de variables cuantitativas y el análisis estadístico. Este enfoque representa cumplir un conjunto secuencial de pasos sin saltar u omitir ninguno de ellos [73].

Se utiliza las variables cuantitativas que se obtiene del modelo de clasificación del suelo que son precisión, f1-score, recall y sensibility para lo cual se desarrolla la matriz de confusión, un modelo matemático para clasificaciones mediante ML.

2.3 Población y muestra

Se define como población el total de imágenes satelitales del planeta tierra capturados por el satélite Sentinel 2. Por lo tanto, se desconoce el total de unidades de la población.

Por lo característica de la investigación se ha empleado un muestreo no probabilístico, en base a objetivos de la investigación [73], por lo tanto es posible construir conjunto de datos de entrenamiento y prueba que se utiliza aplicando los algoritmos ML. Debido que toma relevancia los criterios y decisiones del investigador se ha apoyado en investigaciones previas sobre uso y coberturas del suelo que ha permitido definir 6 clases de coberturas del suelo.

Muestra	Descripción
Entrenamiento	1581 datos construidos mediante el uso de la plataforma GEE.
Prueba	598 datos construidos mediante el uso de la plataforma GEE.

Tabla 5. Población y muestra

En total se tiene 2179 datos, de manera aproximada corresponden a los porcentajes 70%-30% para entrenamiento y prueba respectivamente, según las distribuciones de datos para aplicación mediante ML.

2.4 Métodos teóricos

MÉTODO INDUCTIVO

El método permite ir desde lo particular a lo general, se basa en observaciones particulares para luego poder describir conclusiones generales y probables.

El caso del presente estudio se ajusta a este método debido que en escenarios similares es posible la aplicación de la línea de investigación que sigue para la clasificación de coberturas del suelo, que se obtendrá resultados similares y probables.

2.5 Metodología CRISP-DM

CRISP-DM es una metodología que ofrece un marco de trabajo aplicable en la ciencia de datos. El presente estudio tiene un el enfoque del paradigma cuantitativo. Por lo tanto, CRISP-DM se ajusta a tales exigencias mediante un proceso secuencial sin omitir ninguna fase, partiendo desde la compresión del problema hasta la puesta en producción del modelo implementado.

Las fases de la metodología se han descrito en el capítulo anterior, en este capítulo es la aplicación de cada fase para desarrollo y utilización de imágenes satelitales y ML para clasificación de coberturas del suelo. No se aborda la fase 6 ya que no se despliega en un software para soporte de decisiones.

2.5.1 Comprensión del negocio (problema)

El ordenamiento territorial, definiciones limítrofes, uso y explotación de la tierra son factores que inciden en el desarrollo social en una población. Por lo que, la información geográfica digital y actualizada contribuye como herramienta para toma de decisiones.

Se ha definido la provincia de Chimborazo (Figura 14) como área geográfica del presente estudio. Se ha identificado como problema principal, la escaza y desactualizada información digital sobre la clasificación de los tipos de suelo, corroborado mediante la revisión sistemática de literatura (Capítulo I).

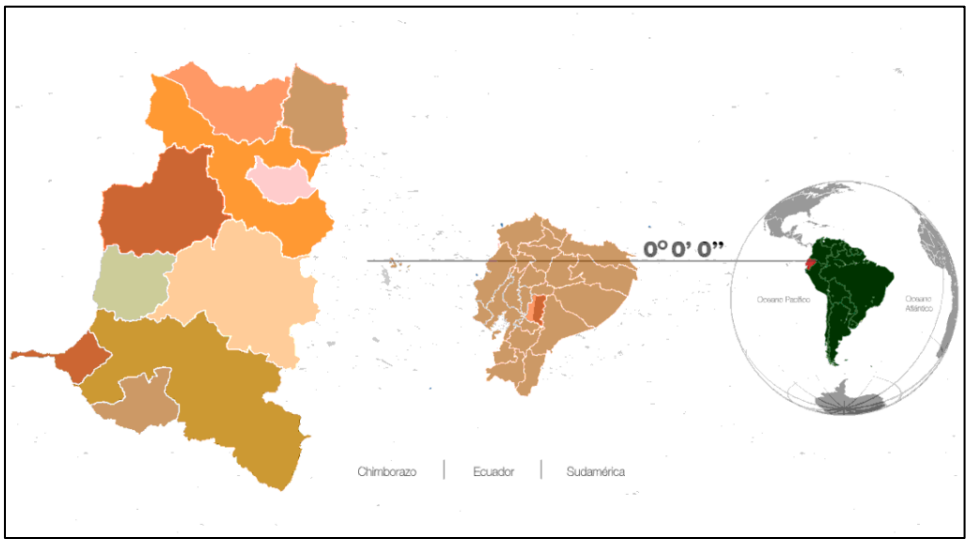


Figura 13. Ubicación de la provincia de Chimborazo en el mapa
Fuente: Chimborazo Travel (2022) [74].

Se propone desarrollar modelos para clasificación de la cobertura del suelo mediante técnicas de ML e imágenes satelitales que permita disminuir la carencia de información digital sobre los tipos de suelos de la provincia de Chimborazo.

Almacén origen de datos	Google Earth Engine
Rango de fechas	2020-01-01 - 2021-12-31
Área geográfica	Ecuador, provincia de Chimborazo.
Satélite	Sentinel 2 – S2
Cantidad de bandas por imagen	13 bandas
Resolución imagen (metro)	10m, 20m, 60

Tabla 6. Información general de la región geográfica de estudio

Según las especificaciones expuestas en la tabla 6 se desarrolla utilizando la plataforma GEE mediante el uso de las imágenes satelitales de Sentinel 2 de los dos últimos años 2020 y 2021 según en el rango establecido.

La region de interes, datos de entrenamiento y prueba se construye utilizando el entorno de desarrollo integrado Code Editor de GEE, asi tambien el desarrollo de los modelos ML utilizando algoritmos CART, Random Forest y SVM; de donde se obtiene las metricas: precision, recall, f1-score, specificity e indice kappa a partir de la matriz de confusión que permite la comparación de los tres modelos ML.

2.5.2 Comprensión de datos

Para la construcción del dataset, datos de entrenamiento y prueba se utiliza las imágenes satelitales de Sentinel 2, que ofrece imágenes multiespectrales de alta resolución de 10m, 20m, 60m por píxel y cuyo objetivo es proporcionar imágenes para detección de cambios terrestres, mapas de cobertura o variaciones geofísicas.

Bandas de Sentinel 2

S2 posee 13 bandas (Capítulo I) para la representación de imágenes, para el presente estudio se ha utilizado 7 banda que se detalla en la tabla 2, según las discriminaciones aplicables definidas en [75].

Banda	Representa	Descripción
B2	Blue	Discrimina de suelo y vegetación.
B3	Green	Discrimina de agua clara y oscura.
B4	Red	Discrimina de frondosidad, tipos de vegetación, suelos y áreas urbanas.
B6	Vegetation Red Edge	Discrimina de agricultura y vegetación.
B8	NIR	Discriminación de costas y vegetación (infrarrojo cercano).
B11	SWIR	Discrimina de vegetación, diferenciación de nubes, nieve (Infrarrojo lejano).
BSI	Bare Soil Index	Discrimina composición de mineral de suelo para detección de áreas desérticas o superficies desnudo (Índice de suelo desnudo)

Tabla 7. Bandas utilizadas de S2

Tambien se puede combinar bandas para representación estandar RGB, indice de vegetacion estandarizada NDVI o indice de suelo desnudo BSI, las cobinaciones se presentan en las formulas que sigue a continuacion.

$$RGB = (B4, B3, B2)$$

$$NDVI = \frac{(B8 - B4)}{(B8 + B4)}$$

$$BSI = \frac{(B11 + B4) - (B8 - B2)}{(B11 + B4) + (B8 + B2)}$$

Donde:

B2: Color primario azul (BLUE).

B3: Color primario verde (GREEN).

B4: Color primario rojo (RED).

B8: Infrarrojo cercano (NIR).

B11: Infrarrojo lejano (SWIR).

Las bandas y combinaciones descritas se utilizan para la construcción del dataset de la región de interés que servirá como almacén general imágenes para entrenamiento y prueba de datos.

Clases de cobertura del suelo

Se ha definido las clases de suelo según el conocimiento del área por el investigador y apoyando en los estudios y especificaciones desarrolladas en el capítulo II por [15], [28], [30], criterio de homogeneidad de áreas terrestres con o sin vegetación, a excepción de Peña et al [30] que subdivide dos grupos entre agua y tierra.

Del estudio desarrollado por MAGAP y MAE se utiliza los niveles 1 y 2 para converger en igualdad de criterios para definir tipos o clases de coberturas terrestres.

Tipo de cobertura	Referencia
Nombre: Agua **Numero: 0** HEX #2840FF	Área sin vegetación - Masa de agua artificial o natural [28]. Agua [30]. Nivel I: Cuerpo de agua [15].
Nombre: Urbano **Numero: 1** HEX #A65A3A	Tierra - Sin Recubrimiento Vegetal > Industrias, carreteras, residencial [30]. Nivel I: Zonas antrópicas – Nivel II: Área poblada; Nivel I: Otras tierras – Nivel II: Infraestructura [15].
Nombre: Forestal **Numero: 2** HEX #5AFF28	Tierra - Con recubrimiento vegetal - Áreas forestales [30]. Nivel I: Bosque y sus Nivel II asociado [15].
Nombre: Cultivo **Numero: 3** HEX #EEFF25	Área con vegetación - Áreas terrestres, cultivadas [28]. Tierra - Con recubrimiento vegetal - Áreas agrícolas [30]. Nivel I: Tierra agropecuaria y su Nivel II asociado [15].
Nombre: Suelo desnudo **Numero: 4** HEX #BF04C2	Área sin vegetación - Áreas desnudas [28]. Tierra – Sin recubrimiento vegetal [30].

	Nivel I: Otras tierras, Nivel II: Área sin cobertura vegetal [15].
Nombre: Arbustivo **Numero: 5** HEX #FF871B	Área con vegetación [28]. Tierra – Con recubrimiento vegetal – Herbáceas y arbustivos [30]. Nivel I: Vegetación arbustiva y herbácea y su Nivel II asociados [15].
Nombre: Nieve **Numero: 6** HEX #FFE7F8	Área sin vegetación – Nieve y hielo [28]. Nivel I: Otras Tierras, Nivel II: Glaciar [15].

Tabla 8. Tipos de cobertura del suelo

En la Tabla 8 se expone los tipos de coberturas del suelo que se utiliza para el etiquetado de datos de entrenamiento y prueba, está compuesta por: nombre, número y color.

2.5.3 Preparación de los datos

Región de interés

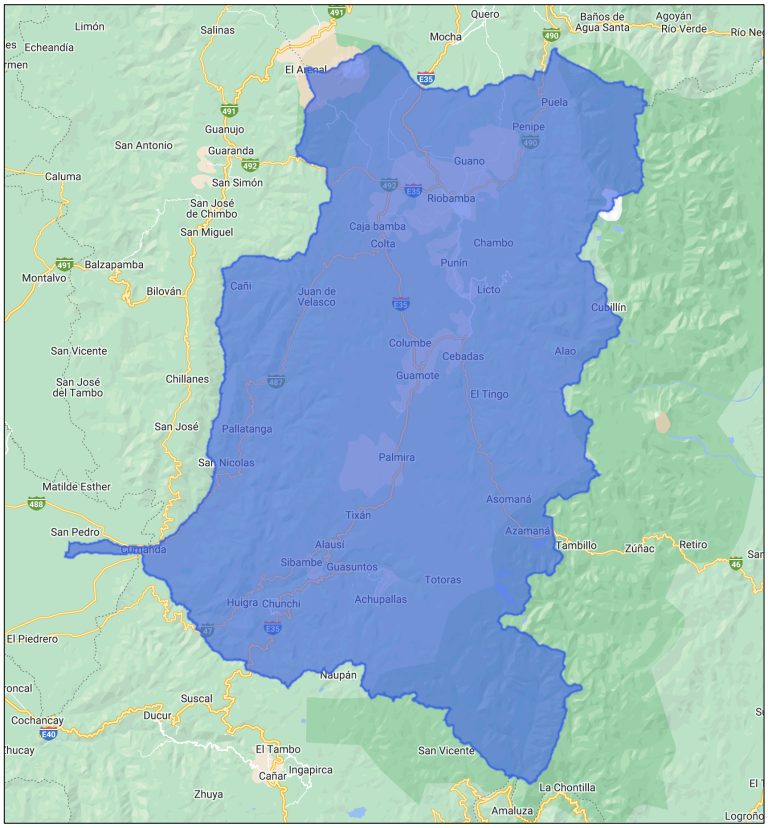


Figura 14. Región de interés provincia de Chimborazo

La región de interés se ha construido y cargado en un archivo comprimido (.shape) que representa la zona geográfica de la provincia de Chimborazo. Se presenta en la Figura 14.

Dataset

El dataset se ha construido introduciendo el periodo de fechas para seleccionar las imágenes satelitales, bandas de entrada y salida.

Periodo de interés	2020-01-01 - 2021-12-31
Bandas de entrada	B2, B3, B4, B6, B8, B11.
Bandas de salida	NDVI, BSI

Figura 15. Datos de entrada para Dataset

Debido a la teledetección algunas áreas componen de capas de nubes. Por lo tanto, se procede a la limpieza de nubosidad haciendo uso de la banda QA60 que contiene conjuntos de bits de las máscaras de nube, para obtener imágenes limpias de nubes.

A continuación, utilizando las bandas de entrada se construye y agrega la composición de las bandas NDVI y BSI que será el resultado final del dataset.

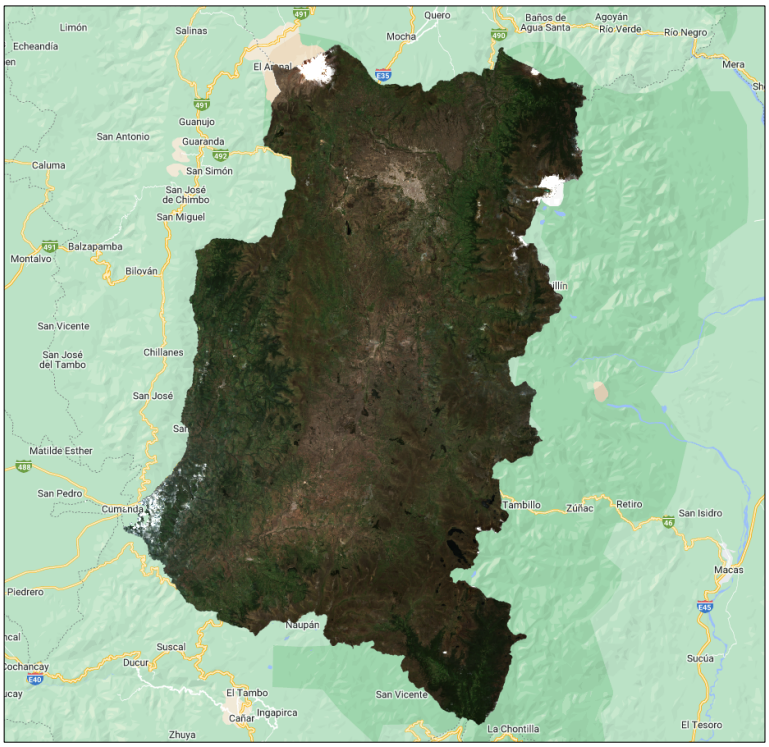


Figura 16. Representación RGB de región de interés.

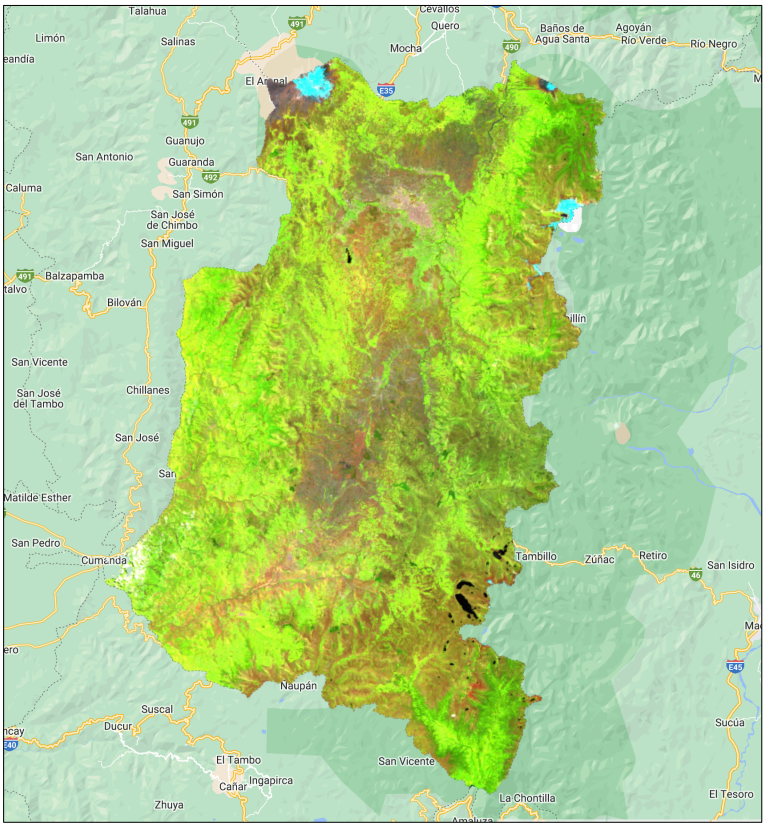


Figura 17. Representación NDVI de la región de interés.

Figura 18. Representación BSI de la región de interés.

Mediante la correlación expuesta en 2.1 y las fórmulas definidas en la fase 2 (comprensión de datos), en las Figuras 16, 17 y 18. Este dataset construido se exporta en un solo conjunto de datos para utilización con los datos de entrenamiento y prueba.

Datos de entrenamiento

El levantamiento de datos de entrenamiento y prueba se ha realizado en la plataforma GEE, de acuerdo con las zonas geográficas que cumplan las características de clases definidas además del conocimiento de la región por parte del investigador.

Para construir los datos de entrenamiento se grafica la región de interés, una vez visualizado el mapa satelital se crea las clases seleccionando FeatureCollection, se asigna: nombre (Agua), propiedad (target), número de clase (0) y color. Esta información se ingresa para cada clase, según la representación de la Figura 17.

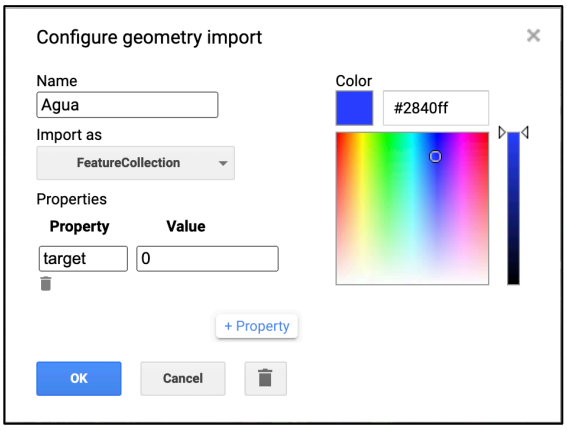
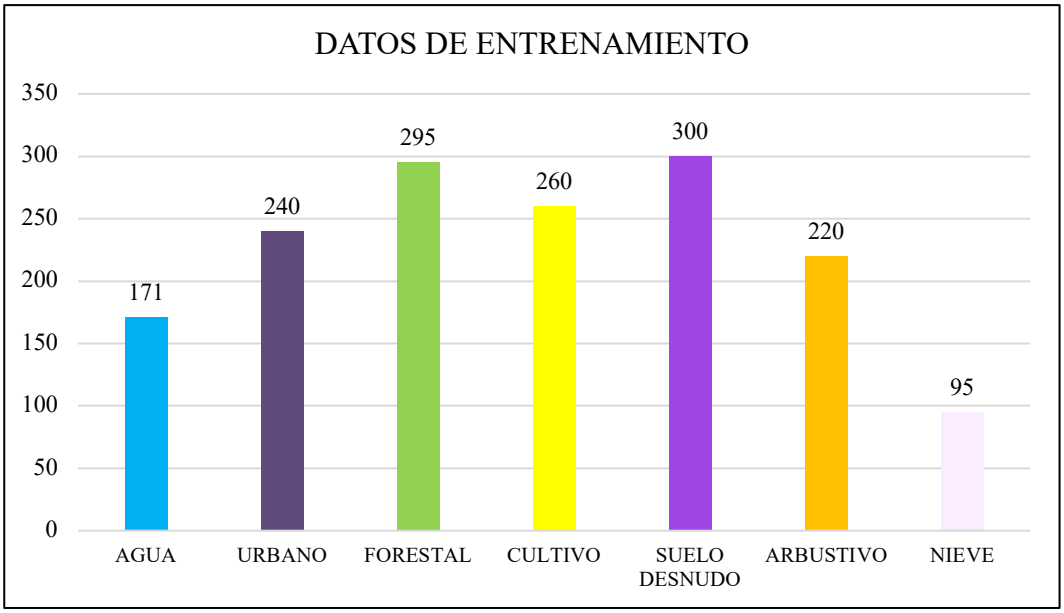


Figura 19. Creación de clases de cobertura del suelo en GEE

El número de clases se asigna en la propiedad "target", mediante esta propiedad se accede al número de clase asignada en el campo "value" definido desde 0 a 6; en total se tiene siete clases para el estudio.

	NAME	TARGET	COLOR	TOTAL
	AGUA	0	2840FF	171
	URBANO	1	A65A3A	240
	FORESTAL	2	5AFF28	295
	CULTIVO	3	EEFF25	260
	SUELO DESNUDO	4	BF04C2	300
	ARBUSTIVO	5	FF871B	220
	NIEVE	6	FFE7F8	95
			TOTAL	**1581**

Tabla 9. Datos de entrenamiento



Gráfica 1. Histograma de datos de entrenamiento.

En total se ha definido 1581 datos etiquetados para los 7 tipos de coberturas, como se representa en la tabla 18 y Gráfica 1.

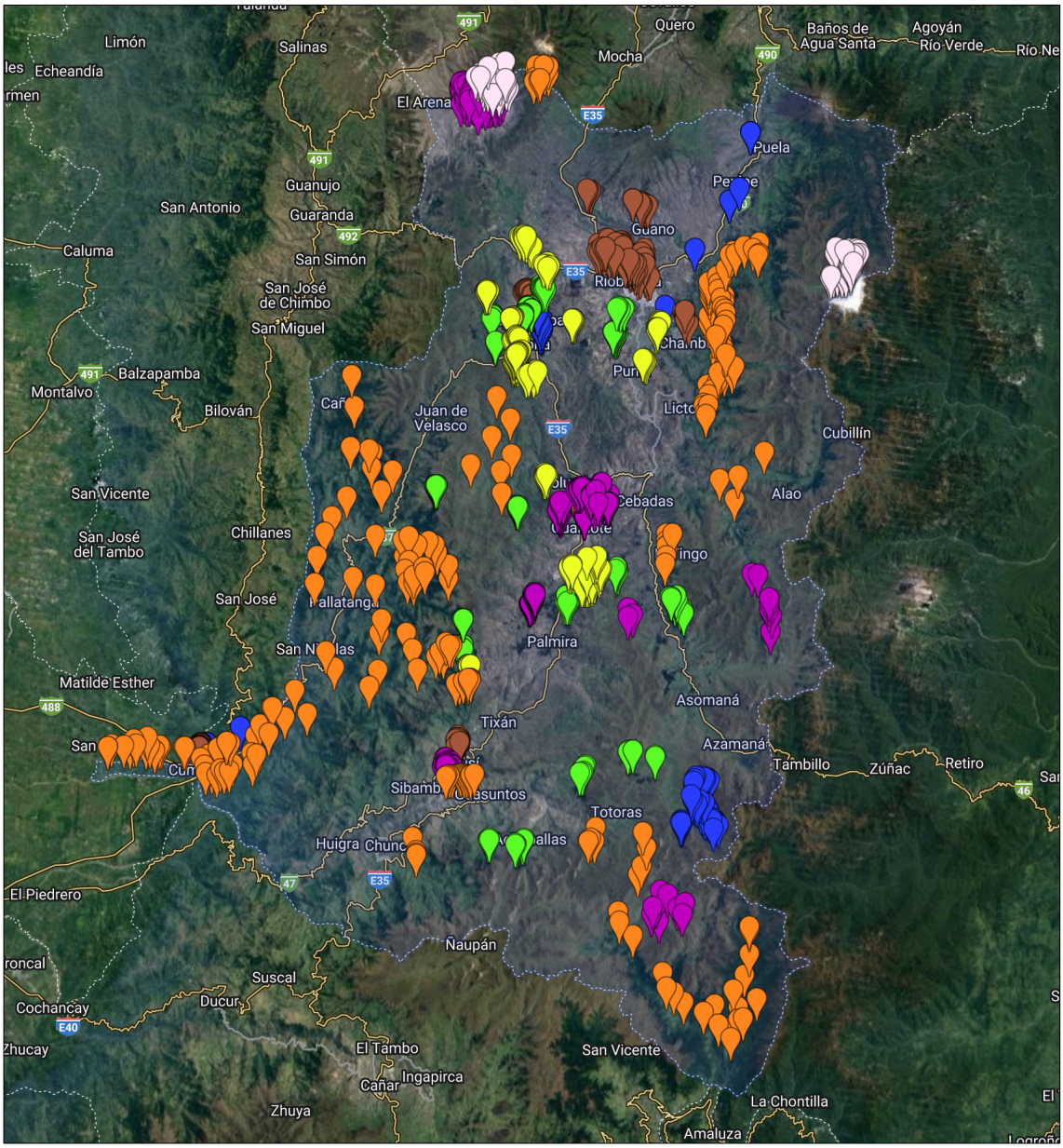


Figura 20. Datos de entrenamiento en GEE

Recopilado los datos de entrenamiento se puede visualizar las etiquetas en la región de interés en la Figura 20. Una vez construido se fusiona las colecciones para transformar en un solo conjunto de datos que contiene cualquier tipo de clase, estas son accesibles mediante el número de clase "value" asignado a la propiedad "target".

```
var puntos_entrenamiento =
Agua.merge(Urbano).merge(Forestal).merge(Cultivo).merge(Suelo_Desnudo).merge(Arbust
ivo).merge(Nieve);
```

Tabla 10. Fusión de datos de entrenamiento

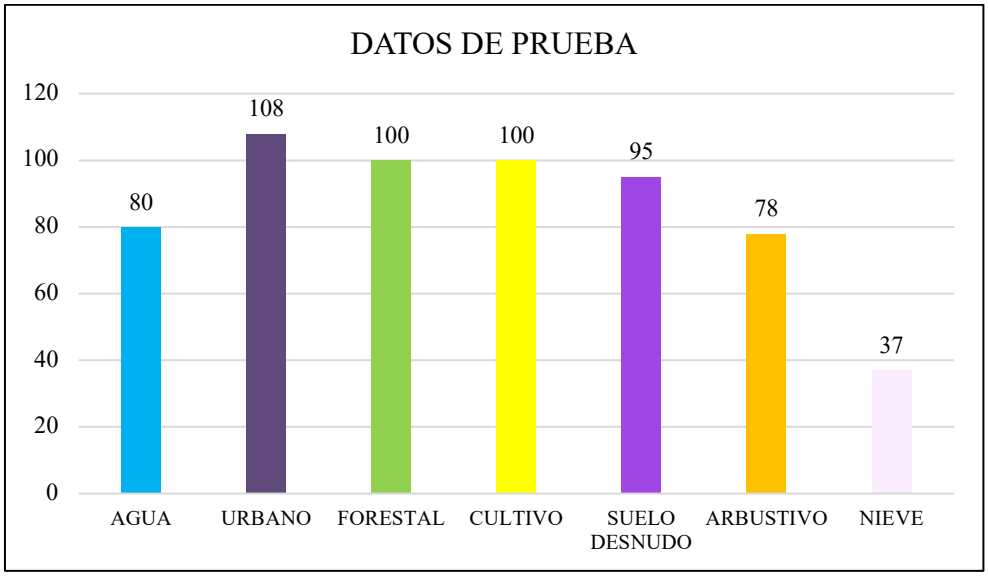
El resultado fusionado se exporta para ser utilizado en el desarrollo de los algoritmos.

Datos de prueba

El mismo proceso o tratamiento se cumple para la construcción de datos de prueba, en la plataforma GEE se ha etiquetado 598 puntos para pruebas, como se presenta en la Tabla 11 y Gráfica 2.

	NAME	TARGET	COLOR	PRUEBA
	AGUA	0	2840FF	80
	URBANO	1	A65A3A	108
	FORESTAL	2	5AFF28	100
	CULTIVO	3	EEFF25	100
	SUELO DESNUDO	4	BF04C2	95
	ARBUSTIVO	5	FF871B	78
	NIEVE	6	2840FF	37
			TOTAL	**598**

Tabla 11. Datos de prueba



Gráfica 2. Histograma de datos de prueba

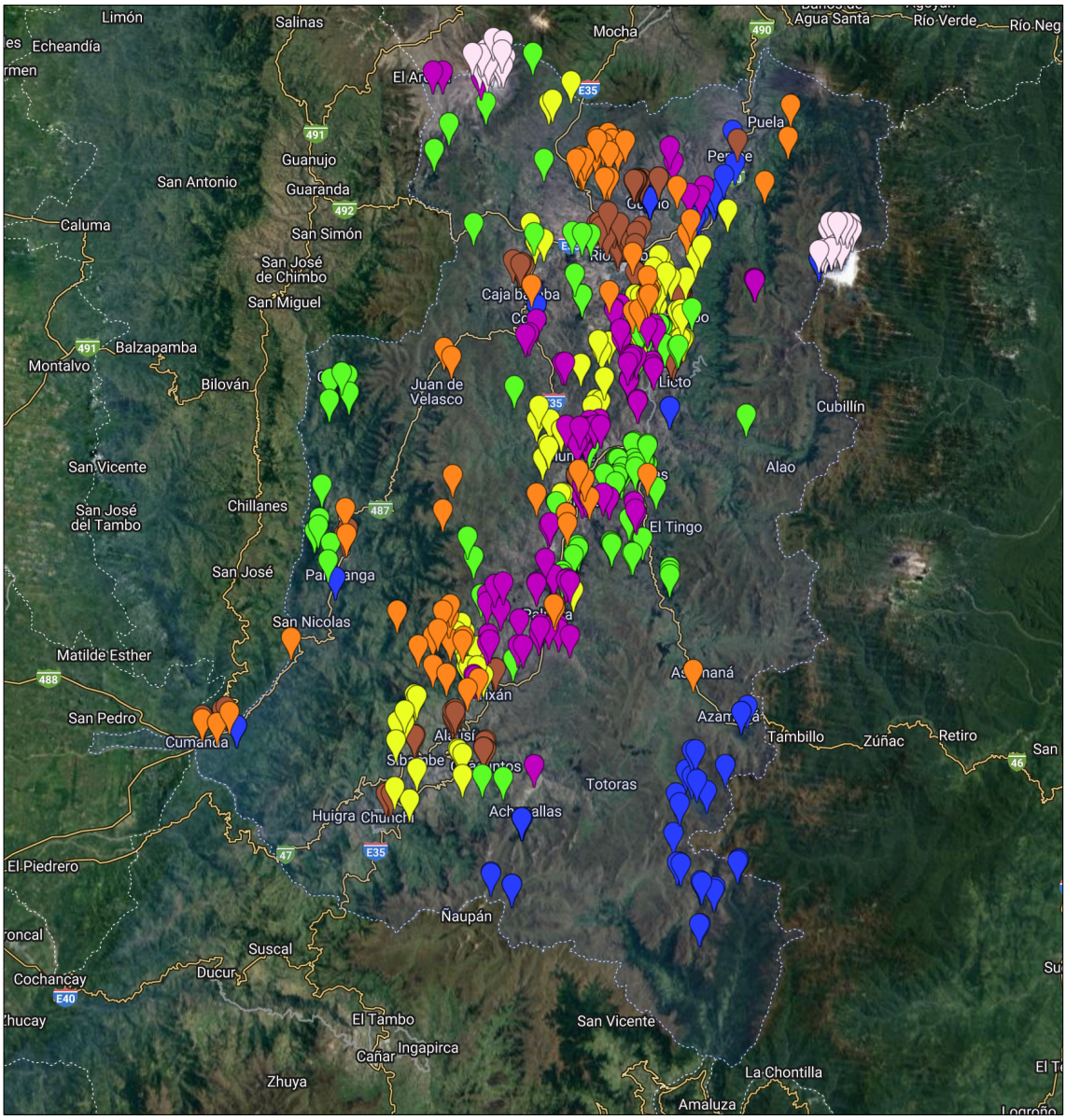
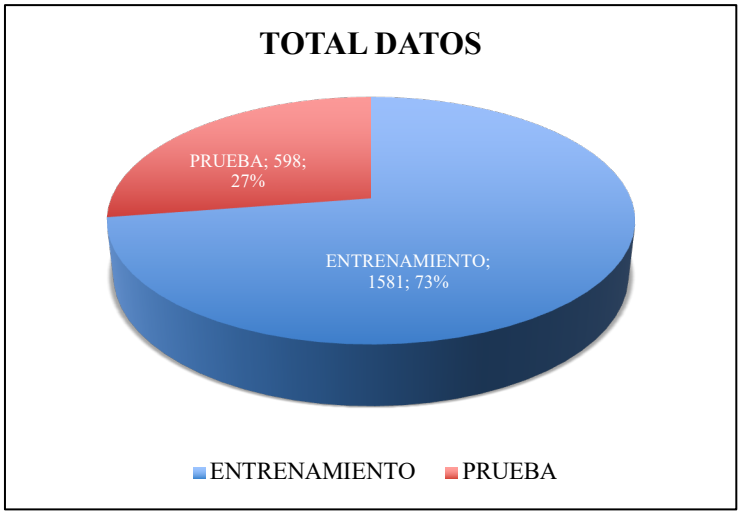


Figura 21. Datos de prueba en GEE.

Los puntos de prueba etiquetados dentro de la región de interés se visualiza en la Figura 21, igual a los datos de entrenamiento se realiza la fusión de clases (Tabla 11) para transformar en un conjunto de datos único que es exportado para utilizar como datos de entrada en los algoritmos de ML.

```
var puntos_prueba =
Agua.merge(Urbano).merge(Forestal).merge(Cultivo).merge(Suelo_Desnudo).merge(Arbust
ivo).merge(Nieve);
```

Tabla 12. Fusión de datos de prueba



Gráfica 3. Datos totales

Conforme a la utilización de datos en ML se aplica la teoría de mayor cantidad de datos para entrenamiento y el restante para prueba; por lo tanto, de manera aproximada se distribuye en 70%-30% para datos de entrenamiento y prueba.

2.5.4 Modelado

En la comprensión del problema (fase 1) se había definido la aplicación de 3 algoritmos de ML, en esta fase se desarrolla haciendo uso del dataset, datos de entrenamiento y prueba construidos en la fase 3.

Los algoritmos de clasificación que se desarrolla son:

- CART: Arboles de clasificación y regresión.
- RF: Bosques aleatorios.
- SVM: Soporte de Maquina Vectorial.

En primera instancia ser hace las importaciones datos de entrada que se había construido previamente (Figura 22): región de interés (region), dataset compuesto (composition), datos de entrenamiento (training) y prueba (test) que utilizará los algoritmos.

```
Imports (4 entries)
  var region: Table projects/ee-lefrainqg/assets/chimborazo_shape
  var composition: Image projects/ee-lefrainqg/assets/chim_data_composicion (7 bands)
  var training: Table projects/ee-lefrainqg/assets/chim_data_entrenamiento
  var test: Table projects/ee-lefrainqg/assets/chim_data_prueba
```

Figura 22. Importación de datos de entrada para algoritmos ML

Uno de los datos de entrada son las bandas, esta información se obtiene desde el dataset compuesto, esto sirve para el filtrado de los datos de entrenamiento y prueba (Figura 22) y como parámetro de los algoritmos.

```
var bands = composition.bandNames();
```

Tabla 13. Obtención de bandas para algoritmos ML

Con el dato de las bandas y utilizando el parámetro "target" se realiza el filtrado de datos de entrenamiento y prueba desde el dataset compuesto, con scala igual a 10 que significa la escala en metros para la proyección de las muestras (Tabla 13).

```
var training_data = composition.select(bands).sampleRegions({
  collection: training,
  properties: ['target'],
  scale: 10
});

var test_data = composition.select(bands).sampleRegions({
  collection: test,
  properties: ['target'],
  scale: 10
});
```

Tabla 14. Filtrado de datos de entrenamiento y prueba algoritmo CART.

Los datos "training_data" y "test_data" filtrados son ahora los datos de entradas para entrenamiento y prueba respectivamente, que utiliza los algoritmos que se desarolla a continuación:

CART (Classification and Regression Tress)

CART es un algoritmo que para su entrenamiento utiliza arboles de decisión por lo que requiere definir los parámetros "maxNodes" número de nodos o profundidad y "minLeafPopulation" tamaño de muestras de hojas para cada nodo.

A demás también requiere los datos de entrenamiento, el nombre del parámetro por el cual accede al valor de cada clase de cobertura "target" y por ultimo las bandas utilizadas.

Entonces, el entrenamiento del algoritmo CART con 100 nodos, y 10 hojas por nodos se define como muestra en la Tabla 15.

```
var cart_classifier = ee.Classifier.smileCart(100, 10).train({
    features: training_data,
    classProperty: 'target',
    inputProperties: bands
```

```
});
```

Tabla 15. Entrenamiento de algoritmo CART.

Se obtiene el resultado desde el dataset general aplicando las bandas utilizadas en el entrenamiento y se visualiza el resultado de mapa de clasificación (Tabla 16).

```
// Imagen clasificada
var classified = composition.select(bands).classify(cart_classifier).focal_mode();

// Visualizar el resultado
Map.addLayer(classified.clip(region), {min: 0, max: 6, palette: palette}, 'CART',
false);
```

Tabla 16. Resultado de clasificación entrenamiento del algoritmo CART

RF (Random Forest)

El algoritmo de clasificación Random Forest o bosques aleatorios construye un conjunto de árboles con muestras aleatorias a partir de los datos de entrenamiento; recibe como parámetro "numberOfTrees" que corresponde al número de árboles que se va a construir.

Así mimo recibe como datos de entrada: datos de entrenamiento el nombre del atributo que permite acceso al tipo de clase determinada y las bandas utilizadas.

El algoritmo se define como se muestra en la Tabla 17, donde: 300 significa el número de árboles que formara el bosque.

```
var rf_classifier =  ee.Classifier.smileRandomForest(300)
  .train({
    features: training_data,
    classProperty: 'target',
    inputProperties: bands
});
```

Tabla 17. Entrenamiento del algoritmo Random Forest

El atributo classProperty con valor "target" es el que permite identificar y clasificar según el valor asociado a cada tipo de cobertura de 0 al 6, de manera que el clasificador partiendo desde un el nodo raíz forma un conjunto de 300 árboles que conforman el bosque.

```
// Imagen clasificada
var img_classified = composition.select(bands).classify(rf_classifier).focal_mode();

// Visualizar el resultado
Map.addLayer(img_classified.clip(region), {min: 0, max: 6, palette: palette}, "RF",
false);
```

Tabla 18. Visualización de resultado clasificación RF.

Realizado el entrenamiento del clasificador se obtiene la imagen clasificada (Tabla 18), que tambien muestra el mapa de clasificación.

SVM (Support Vector Machine)

Realizado el proceso de normalización de datos se obtiene las bandas desde el dataset normalizado para asegurar que no se haya perdido ninguna en este proceso.

El algoritmo máquina de soporte vectorial en su proceso de entrenamiento hace el trazado de los hiperplanos de separación optima (HSO) utilizando un tipo de núcleo, en este estudio se utiliza el kernel: función de base radial (RBF, por sus siglas en inglés).

El clasificador SVM recibe 3 parámetros que deciden el rendimiento del algoritmo:

Parámetro	Descripción
kernel	Puede ser: LINEAR, POLY, RBF o SIGMOID.
gamma	Define que tan cerca o lejos está la influencia de los datos de entrenamiento. De manera técnica establece la distancia de distribución de datos al espacio trazado en HSO. • Mas bajo y más alto: menor precisión. • Valores intermedios: menor distancia de separación.
cost	Parámetro de regularización que controla la penalización de clasificación errónea; también se conoce como *C*. • Mas alta: menor o no tolerable y recae en sobreajuste; sanción más alta por clasificación errónea. • Mas bajo: mayor tolerancia y márgenes suaves; sanción menor por clasificación errónea.

Tabla 19. Parámetros del algoritmo SVM.

En este estudio al ser clasificación multiclass con datos linealmente no separables se utiliza el kernel "RBF", con gamma 0.1 y cost 10, el clasificador SVM se define como se presenta en la Tabla 19.

```
var svm_params = { kernelType: 'RBF', gamma: 0.1, cost: 10 };
var svm_classifier = ee.Classifier.libsvm(svm_params)
  .train({
    features: training_data,
    classProperty: 'target',
    inputProperties: bands
});
```

Tabla 20. Definición del algoritmo SVM.

```
// Imagen clasificada
```

```
var img_classified = composition.select(bands).classify(rf_classifier).focal_mode();

// Visualizar el resultado
Map.addLayer(img_classified.clip(region), {min: 0, max: 6, palette: palette}, "RF",
false);
```

Tabla 21. Mostrar resultado de clasificador SVM

Al igual que los anteriores algoritmos se obtiene el resultado de clasificación utilizando el algoritmo entrenado (Tabla 21).

2.5.5 Evaluación

La fase 5 de CRISP-DM se desarrolla en el capítulo III "Resultados" mediante la obtención de resultados planteando los criterios de validaciones, se obtiene: el mapa de clasificación, matriz de confusión, precisión, recall, f1-score, specificity; para los 3 clasificadores desarrollados.

2.6 Métodos empíricos

Se define como estudio de caso debido que su aplicación se basa en una región especifica, en este caso la zona geográfica que comprende la provincia de Chimborazo ubicado en la zona centro de Ecuador, para definir como condición la utilización de imágenes satelitales y clasificar las coberturas terrestres que compone la región.

2.7 Técnicas estadísticas

En cuanto las técnicas estadísticas se aplica las medidas de tendencia central, gráficos e inferencia estadísticos a partir de los modelos matemático que arrojan resultados de las métricas para los algoritmos ML.

CAPITULO III

3. RESULTADOS

En el capítulo II se realizado el procesamiento de datos y construcción de modelos. En esta sección se desarrolla el análisis, evaluación, comparación de los 3 modelos de clasificación de cobertura terrestre según los resultados y métricas obtenidas con los datos de prueba. Además, en este capítulo se cumple con la fase 5 "Evaluación" de la metodología CRISP-DM.

Para cada algoritmo de clasificación se obtiene los resultados de métricas:

3.1 Matriz de Confusión.

Una de las formas adecuadas de evaluar el desempeño de un clasificador, es obteniendo la matriz de confusión. La idea de esta matriz es contabilizar el total de veces se ha clasificado de manera correcta un tipo de clase [51].

También se denomina matriz de error o de contingencia donde los valores que aparecen en la diagonal indica número de clasificaciones correctas y las que quedan fuera son migraciones o fugas. Entonces según Sánchez Muñoz [76] se define 2 tipos de errores:

- Error de omisión: Son datos que perteneciendo a esa clase no aparecen por estar incluidos en otra de manera incorrecta (datos por debajo de la diagonal principal de la matriz).
- Error de comisión: Son datos que no perteneciendo a una clase aparecen en ella (datos por encima de la diagonal principal de la matriz de confusión).

Al ser una matriz está conformada por filas y columnas, las filas corresponden a los datos reales y las columnas a las predicciones.

Entonces la matriz de confusión se define de la siguiente manera:

		PRED	
		1	0
REAL	1	**TP**	**FN**
	0	**FP**	**TN**

Tabla 22. Definición de matriz de confusión

Donde:

- TP = True Positive (Verdadero Positivo): Cantidad de positivos que fueron clasificados correctamente como positivos.
- FN = False Negative (Verdadero Negativo): Cantidad de negativos que fueron clasificados correctamente como negativos.
- FP = False Positive (Falso Positivo): Cantidad de positivos que fueron clasificados incorrectamente como negativos.
- TN = True Negative (Verdadero Negativo). Cantidad de negativos que fueron clasificados incorrectamente como positivos.

A partir de estos datos se obtiene las métricas como resultado del desempeño del clasificador.

3.2 Matriz de Observación.

La matriz de observación es el resultado de todos los valores totales con las métricas para cada clase una vez procesada mediante las definiciones y fórmulas que corresponde a cada uno.

3.3 Métricas

Sea:

i, j = índices de filas y columnas de la matriz respectivamente.
C = clase.
N = total datos de prueba.

Entonces se define las siguientes fórmulas:

- Precisión General (overall acuracy).

$$\text{overall acuracy} = \frac{\sum_{ij=0}^{C} TP_{i=j}}{N}$$

Se define la precisión general a los valores ubicados en la diagonal principal, por lo tanto es la sumatoria de los valores clasificados correctamente (*TP*) para la clase *C* desde *i,j=0* cuyo índice de fila y columna sea igual valor *i=j*.

- Precisión.

$$\text{precision} = \frac{TP}{TP + FP}$$

La precisión representa la proporción de verdadero positivos, es el porcentaje de casos positivos, es decir las predicciones hechas por el modelo que coincide con la clase positiva real.

- Sensibilidad (recall, sentitivity).

$$\text{recall} = \frac{TP}{TP + FN}$$

La sensibilidad o recall es la tasa de verdaderos positivos. Es el porcentaje de casos positivos que fueron correctamente identificadas por el algoritmo.

- Especificidad (specificity)

$$\text{specificity} = \frac{TN}{TN + FP}$$

La especificidad es la tasa de verdaderos negativos. Es el porcentaje de casos negativos que el algoritmo ha clasificado correctamente.

- F1-Score

$$F1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} * \text{recall}}$$

Esta métrica resume la precisión y sensibilidad, analiza el balanceo entre las dos métricas.

- Índice o coeficiente de Kappa.

Desarrollado por Jacob Cohen (1960) [77], generalmente se conoce como medida de precisión que considera la concordancia de ausencia de azar [78]. El índice o coeficiente kappa mide el grado de concordancia de N elementos con C categorías, entonces Cohen define como la proporción de desacuerdos esperados al azar que no ocurren [79].

El valor kappa está dentro del rango [-1, 1], sea *k* el índice de kappa entonces según [80], se define:

- $k < 0$: Negativo, significa no existe concordancia, puede alcanzar hasta -1.
- $k = 0$: Los valores observados son independientes.
- $k > 0$: Positiva, existe concordancia; siendo +1 concordancia perfecta.

Coeficiente Kappa	Fuerza de concordancia
0,00	Pobre (Poor)
0,01 - 0,20	Leve (Slight)
0,21 - 0,40	Aceptable (Fair)
0,41 - 0,60	Moderada (Moderate)
0,61 - 0,80	Considerable (Substantial)
0,81 - 1,00	Casi perfecta (Almost perfect)

Tabla 23. Valoración de concordancia del coeficiente Kappa.
Fuente: [80], [81].

Así mismo Cerda et al [80] recoge del estudio de Landis y Koch [81] cuyas valoraciones del coeficiente kappa que se presenta en la Tabla 23.

3.4 Criterios de pruebas de los algoritmos

Se considera las siguientes pruebas por cada clasificador:

Clasificador	Criterio de prueba
CART	Variación de árboles de decisión y sus hojas. 1. 50 árboles con 5 hojas 2. 100 árboles con 25 hojas 3. 300 árboles con 50 hojas
RF	Variación de cantidad de árboles para el bosque. 1. 50 árboles para el bosque. 2. 150 árboles para el bosque. 3. 300 árboles para el bosque.
SVM	Núcleo RBF. 1. Variación de gamma: gamma={0.008, 0.01, 0.05, 0.1, 0.5, 1.0, 3.0, 7.0, 11.0} , cost=0.1 2. Variación de cost: gamma=0.1, cost={0.02, 0.03, 0.08, 1.0, 10.0, 100.0} 3. Svm con gamma y cost optimo.

Tabla 24. Criterios de clasificación de los algoritmos de ML utilizados

3.5 Obtención de métricas en GEE

Para obtener los resultados se introduce los datos de prueba al clasificador; de este resultado se puede obtener la matriz de confusión y luego las métricas.

```
print('VALIDACION DATOS DE ENTRENAMIENTO')
var validated = test_data.classify(cart_classifier);

// Matriz de confusion
var mc = validated.errorMatrix("target", "classification");
print("Matriz de Confusión: ", mc);

// Metricas
print('Accuracy (precision general): ', mc.accuracy());
print('Precision (specificity): ', mc.consumersAccuracy());
print('Sensitivity (recall): ', mc.producersAccuracy());
print('f1-score: ', mc.fscore());
print("Indice Kappa: ", mc.kappa());
```

Tabla 25. Obtención de resultados de un algoritmo de clasificación.

En la Tabla 25 se muestra el procesamiento de datos de prueba, obtención de la matriz de confusión, donde:

- test_data: Corresponde a los datos de prueba procesados en capitulo II, Metodología CRISP-DM, Fase 3, Figura 20, Tabla 11.
- cart_classifier: Corresponde al tipo de clasificador, puede ser: cart_classifier, rf_classifier o svm_classifier; desarrollados en el capítulo II, Metodología CRISP-DM, Fase 4.

3.6 Desarrollo de pruebas del algoritmo CART

Según los criterios de prueba para el algoritmo CART se obtiene los resultados para cada variación de la cantidad de árboles.

1. 50 árboles con 5 hojas.

RESULTADOS.

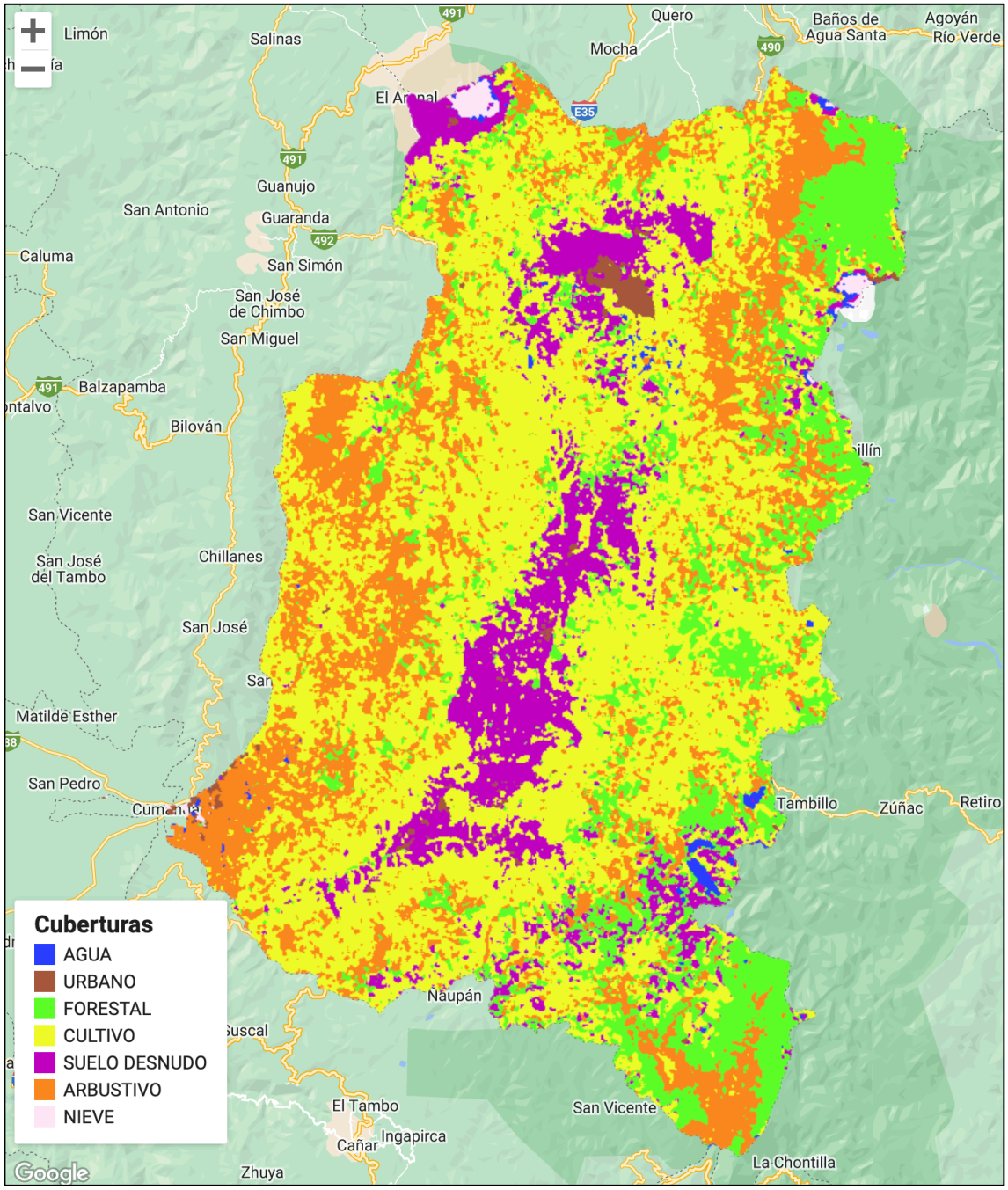


Figura 23. Mapa de clasificación algoritmo CART, criterio de prueba 1.

MATRIZ DE CONFUSIÓN									
			PREDICCIÓN						
			0	**1**	**2**	**3**	**4**	**5**	**6**	
			AGUA	URBANO	FORESTAL	CULTIVO	SUELO DESNUDO	ARBUSTIVO	NIEVE	**TOTAL**
REAL	**0**	AGUA	**71**	0	1	4	4	0	0	**80**
	1	URBANO	0	**92**	0	1	10	0	5	**108**
	2	FORESTAL	0	0	**84**	7	3	6	0	**100**
	3	CULTIVO	1	2	0	**81**	2	14	0	**100**
	4	SUELO DESNUDO	0	8	4	6	**76**	1	0	**95**
	5	ARBUSTIVO	1	1	13	21	1	**40**	1	**78**
	6	NIEVE	1	0	0	0	2	0	**34**	**37**

	SUMA TOTAL	**74**	**103**	**102**	**120**	**98**	**61**	**40**	**598**
	Precisión	96%	89%	82%	68%	78%	66%	85%	
	Error de Omisión	4%	11%	18%	33%	22%	34%	15%	
	Error de Comisión	11%	15%	16%	19%	20%	49%	8%	
	Precisión General	**79,93**							

Tabla 26. Matriz de confusión algoritmo CART, criterio de prueba 1.

		MATRIZ DE OBSERVACIÓN							
		TP	**TN**	**FP**	**FN**	**precision**	**recall**	**f1-score**	**specificity**
0	AGUA	71	515	3	9	0,959459459	0,8875	0,922077922	0,994208494
1	URBANO	92	479	11	16	0,893203883	0,851851852	0,872037915	0,97755102
2	FORESTAL	84	480	18	16	0,823529412	0,84	0,831683168	0,963855422
3	CULTIVO	81	459	39	19	0,675	0,81	0,736363636	0,921686747
4	SUELO DESNUDO	76	481	22	19	0,775510204	0,8	0,787564767	0,956262425
5	ARBUSTIVO	40	499	21	38	0,655737705	0,512820513	0,575539568	0,959615385
6	NIEVE	34	555	6	3	0,85	0,918918919	0,883116883	0,989304813
TOTAL						**0,804634381**	**0,803013041**	**0,801197694**	**0,966069187**

Tabla 27. Matriz de observación algoritmo CART, criterio de prueba 1.

índice kappa = 0.7630212805304942

ANÁLISIS.

Obtenido los resultados del mapa de cobertura (Figura 23), matriz de confusión (Tabla 26) y la matriz de observación (Tabla 27), se deduce:

- Se ha alcanzado el 79,93 de precisión general.
- Las clases 3 (cultivo), 4 (suelo desnudo), 5 (arbustivo) poseen menor precisión con 0.67, 077, 0.65 (Tabla 24) respectivamente. Esto significa que no se pudo predecir de manera correcta las clases que pertenecen a la clase real. En consecuencia, para estas clases se tiene mayores porcentajes en los errores según se presenta en la Tabla 23.
- La clase 5 (arbustivo) con 0.51, implica que posee menor porcentaje de clases correctamente clasificadas.
- Con los parámetros de prueba definidos, con 0.76 de índice kappa, el grado de acuerdo del clasificador es considerable.

2. 100 árboles con 25 hojas.

RESULTADOS.

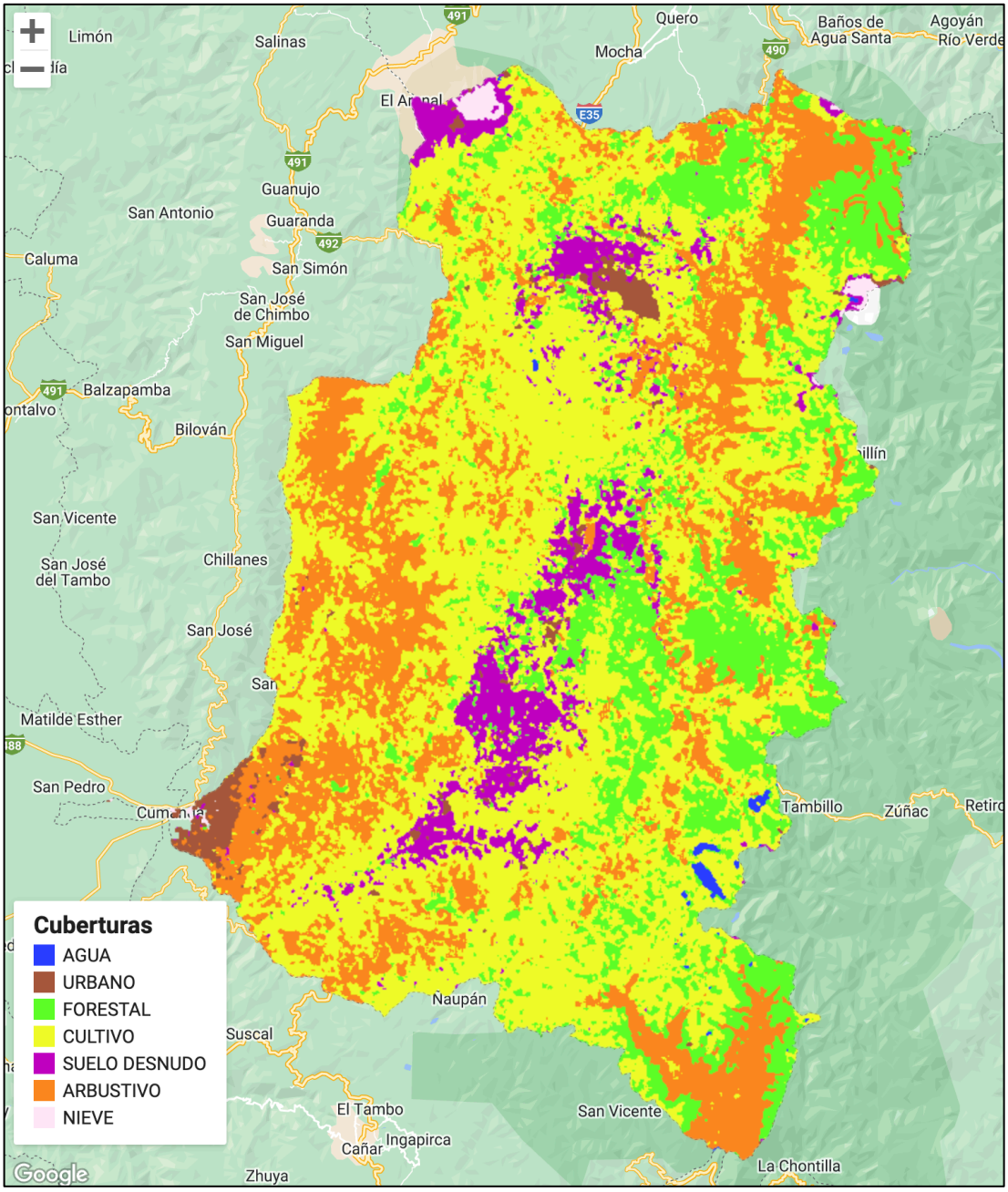


Figura 24. Mapa de clasificación algoritmo CART, criterio de prueba 2.

MATRIZ DE CONFUSIÓN									
		PREDICCIÓN							
		0	1	2	3	4	5	6	
		AGUA	URBANO	FORESTAL	CULTIVO	SUELO DESNUDO	ARBUSTIVO	NIEVE	**TOTAL**
REAL	0 AGUA	**70**	0	2	3	4	1	0	**80**
	1 URBANO	0	**81**	0	2	20	0	5	**108**
	2 FORESTAL	0	0	**84**	8	0	8	0	**100**
	3 CULTIVO	0	1	0	**76**	4	19	0	**100**
	4 SUELO DESNUDO	0	6	8	14	**67**	0	0	**95**
	5 ARBUSTIVO	0	1	7	22	1	**46**	1	**78**
	6 NIEVE	0	0	0	1	1	0	**35**	**37**
	SUMA TOTAL	**70**	**89**	**101**	**126**	**97**	**74**	**41**	**598**
	Presición	100%	91%	83%	60%	69%	62%	85%	
	Error de omisión	0%	9%	17%	40%	31%	38%	15%	
	Error de comisión	13%	25%	16%	24%	29%	41%	5%	

Precisión General	**76,76**

Tabla 28. Matriz de confusión algoritmo CART, criterio de prueba 2.

		MATRIZ DE OBSERVACIÓN							
		TP	**TN**	**FP**	**FN**	**precision**	**recall**	**f1-score**	**specificity**
0	AGUA	70	518	0	10	1	0,875	0,933333333	1
1	URBANO	81	482	8	27	0,91011236	0,75	0,822335025	0,983673469
2	FORESTAL	84	481	17	16	0,831683168	0,84	0,835820896	0,965863454
3	CULTIVO	76	448	50	24	0,603174603	0,76	0,672566372	0,899598394
4	SUELO DESNUDO	67	473	30	28	0,690721649	0,705263158	0,697916667	0,940357853
5	ARBUSTIVO	46	492	28	32	0,621621622	0,58974359	0,605263158	0,946153846
6	NIEVE	35	555	6	2	0,853658537	0,945945946	0,897435897	0,989304813
TOTAL						**0,787281706**	**0,780850385**	**0,780667335**	**0,960707404**

Tabla 29. Matriz de observación algoritmo CART, criterio de prueba 2.

índice kappa = 0.7258400728261013

ANÁLISIS.

Obtenido los resultados del mapa de cobertura (Figura 24), matriz de confusión (Tabla 28) y la matriz de observación (Tabla 29), se deduce:

- Con precisión general de 76,76 se evidencia el impacto en el cambio en número de árboles y hojas al caso de prueba 1. Entonces, el número de hojas debe tener un valor equilibrado ni muy grande ni muy bajo.
- Para la clase 5 que corresponde al tipo de cobertura Arbustivo, con el valor de precisión 0,621 y recall 0,589, se observa que casi la mitad de los datos de prueba no se ha clasificado de manera óptima.
- Por el valor de kappa 0.725 y los parámetros ingresados el clasificador tiene fuerza de concordancia considerable, al igual que el caso de prueba 1.

3. 300 árboles con 50 hojas.

RESULTADOS.

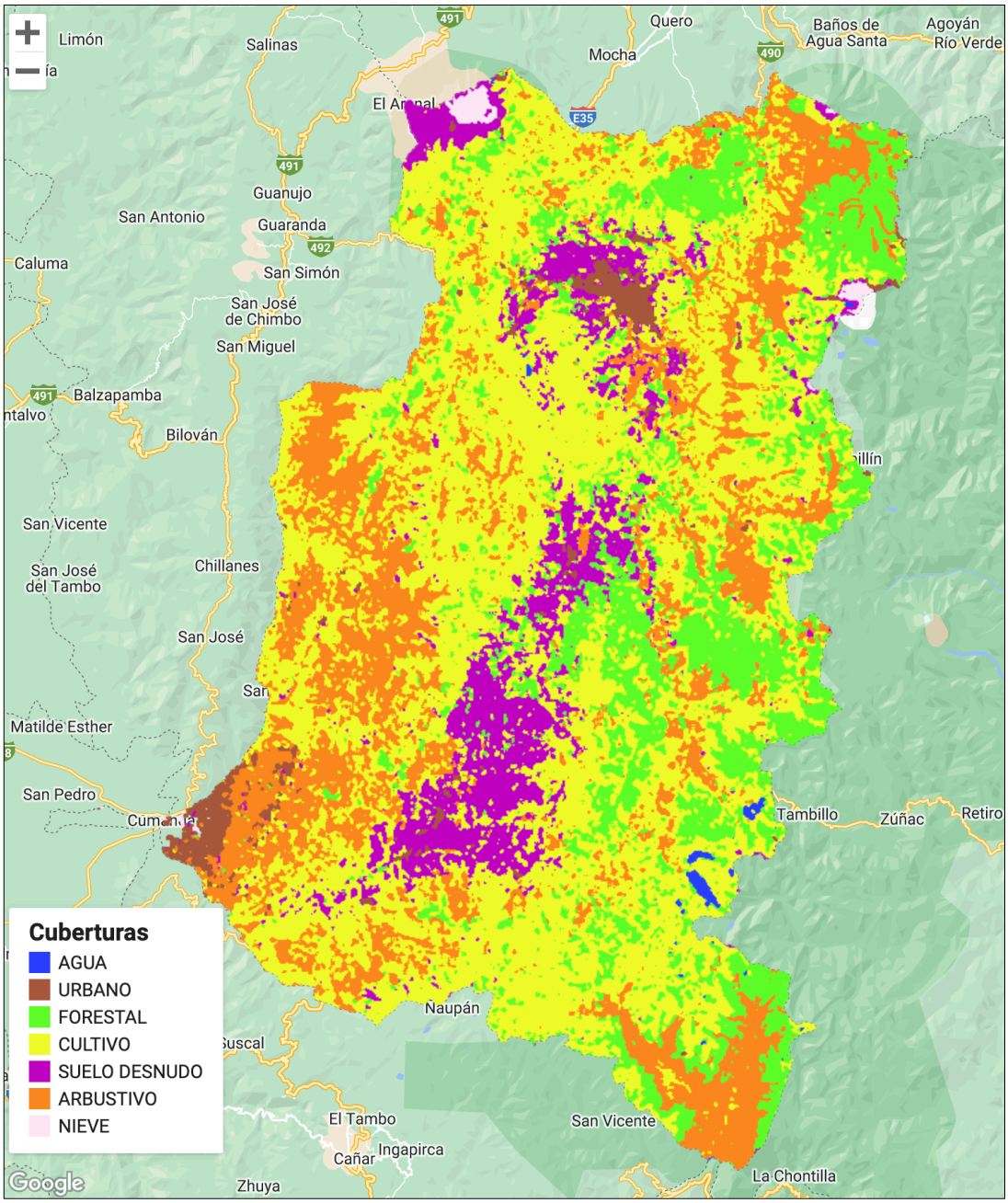


Tabla 30. Mapa de clasificación algoritmo CART, criterio de prueba 3.

MATRIZ DE CONFUSIÓN									
			PREDICCIÓN						
			0	1	2	3	4	5	6	
			AGUA	URBANO	FORESTAL	CULTIVO	SUELO DESNUDO	ARBUSTIVO	NIEVE	TOTAL
REAL	0	AGUA	**70**	0	2	3	4	1	0	**80**
	1	URBANO	0	**92**	0	1	10	0	5	**108**
	2	FORESTAL	0	0	**84**	8	0	8	0	**100**
	3	CULTIVO	0	2	0	**72**	14	12	0	**100**
	4	SUELO DESNUDO	0	7	8	10	**70**	0	0	**95**
	5	ARBUSTIVO	0	1	7	27	1	**41**	1	**78**
	6	NIEVE	0	0	0	1	1	0	**35**	**37**
	SUMA TOTAL		**70**	**102**	**101**	**122**	**100**	**62**	**41**	**598**
	Presición		100%	90%	83%	59%	70%	66%	85%	
	Error de omisión		0%	10%	17%	41%	30%	34%	15%	
	Error de comisión		13%	15%	16%	28%	26%	47%	5%	
	Precisión General		**77,59**							

Tabla 31. Matriz de confusión algoritmo CART, criterio de prueba 3.

		MATRIZ DE OBSERVACIÓN							
		TP	**TN**	**FP**	**FN**	**precision**	**recall**	**f1-score**	**specificity**
0	AGUA	70	518	0	10	1	0,875	0,933333333	1
1	URBANO	92	480	10	16	0,901960784	0,851851852	0,876190476	0,979591837
2	FORESTAL	84	481	17	16	0,831683168	0,84	0,835820896	0,965863454
3	CULTIVO	72	448	50	28	0,590163934	0,72	0,648648649	0,899598394
4	SUELO DESNUDO	70	473	30	25	0,7	0,736842105	0,717948718	0,940357853
5	ARBUSTIVO	41	499	21	37	0,661290323	0,525641026	0,585714286	0,959615385
6	NIEVE	35	555	6	2	0,853658537	0,945945946	0,897435897	0,989304813
TOTAL						**0,791250964**	**0,785040133**	**0,785013179**	**0,962047391**

Tabla 32. Matriz de observación algoritmo CART, criterio de prueba 3.

índice kappa = 0.735393861343636

ANÁLISIS.

Obtenido los resultados del mapa de cobertura (Figura 30), matriz de confusión (Tabla 31) y la matriz de observación (Tabla 32), se deduce:

- Se ha alcanza la precisión de 77,59%. Por lo tanto, en la tercera prueba se infiere que el parámetro cantidad de hojas implica en mayor medida el porcentaje de precisión en el algoritmo.
- En la matriz de observación columna f1-score, el balance de métricas alcanzar valores altos en la prueba 1 y 2; por lo tanto, tiene consistencia en la correlación de las métricas precisión y recall.
- Por el resultado del índice de kappa 0,735, el algoritmo tiene fuerza de concordancia considerable.

3.7 Desarrollo de pruebas del algoritmo RF

1. 50 árboles para el bosque.

RESULTADOS

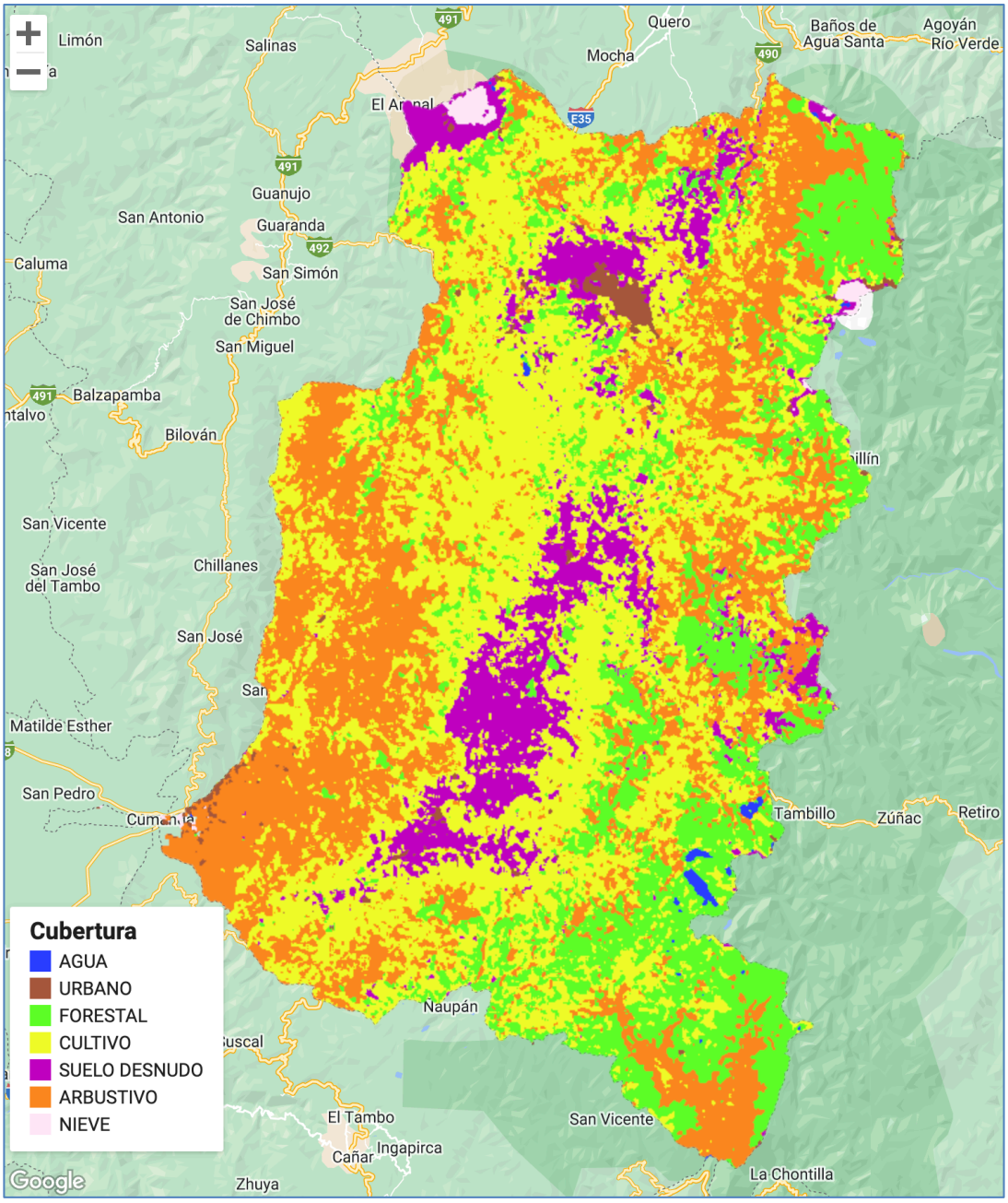


Figura 25. Mapa de clasificación algoritmo RF, criterio de prueba 1.

MATRIZ DE CONFUSIÓN									
			PREDICCIÓN						
			0	1	2	3	4	5	6	
			AGUA	URBANO	FORESTAL	CULTIVO	SUELO DESNUDO	ARBUSTIVO	NIEVE	TOTAL
REAL	0	AGUA	**73**	0	2	1	3	1	0	**80**
	1	URBANO	0	**100**	0	1	6	0	1	**108**
	2	FORESTAL	0	0	**87**	3	3	7	0	**100**
	3	CULTIVO	0	2	0	**83**	3	12	0	**100**

4	SUELO DESNUDO	0	3	2	8	**81**	1	0	**95**
5	ARBUSTIVO	0	0	4	7	1	**65**	1	**78**
6	NIEVE	1	0	0	0	0	0	**36**	**37**
SUMA TOTAL		**74**	**105**	**95**	**103**	**97**	**86**	**38**	**598**
Presición		99%	95%	92%	81%	84%	76%	95%	
Error de omisión		1%	5%	8%	19%	16%	24%	5%	
Error de comisión		9%	7%	13%	17%	15%	17%	3%	
Precisión General		**87,79**							

Tabla 33. Matriz de confusión algoritmo RF, criterio de prueba 1.

		MATRIZ DE OBSERVACIÓN							
		TP	**TN**	**FP**	**FN**	**precision**	**recall**	**f1-score**	**specificity**
0	AGUA	73	517	1	7	0,986486486	0,9125	0,948051948	0,998069498
1	URBANO	100	485	5	8	0,952380952	0,925925926	0,938967136	0,989795918
2	FORESTAL	87	490	8	13	0,915789474	0,87	0,892307692	0,983935743
3	CULTIVO	83	478	20	17	0,805825243	0,83	0,81773399	0,959839357
4	SUELO DESNUDO	81	487	16	14	0,835051546	0,852631579	0,84375	0,968190855
5	ARBUSTIVO	65	499	21	13	0,755813953	0,833333333	0,792682927	0,959615385
6	NIEVE	36	559	2	1	0,947368421	0,972972973	0,96	0,996434938
TOTAL						**0,885530868**	**0,885337687**	**0,884784813**	**0,979411671**

Tabla 34. Matriz de observación algoritmo RF, criterio de prueba 1.

índice kappa = 0.8560295499892815

ANÁLISIS

El caso de prueba para el algoritmo RF con 50 arboles presentado en los resultados Figura 24, Tabla 32 y 33, alcanza la precisión general de aproximadamente ≈87.79%. En las pruebas con numero de árboles menores a 50 se ha notado que no baja desde ≈80%. Por lo tanto, según el valor de índice kappa 0,856 el clasificador está en el rango de clasificación casi perfecta según los criterios definidos para esta métrica.

En la tabla de observación se puede evidenciar que los valores de f1-score alcanzar valores altos, esto implica el balance de correlación de las métricas precisión y recall que determina que efectivamente clasificaron de forma correcta las clases.

Según resultados de la matriz de observación la clase Agua (0) posee valor más alto en las métricas; caso opuesto con la clase Arbustivo (5) con valores bajos. Entonces existe menores errores en la clase 0 y mayores en la clase 5 como se evidencia en la Tabla 31.

2. 150 árboles para el bosque.

RESULTADOS

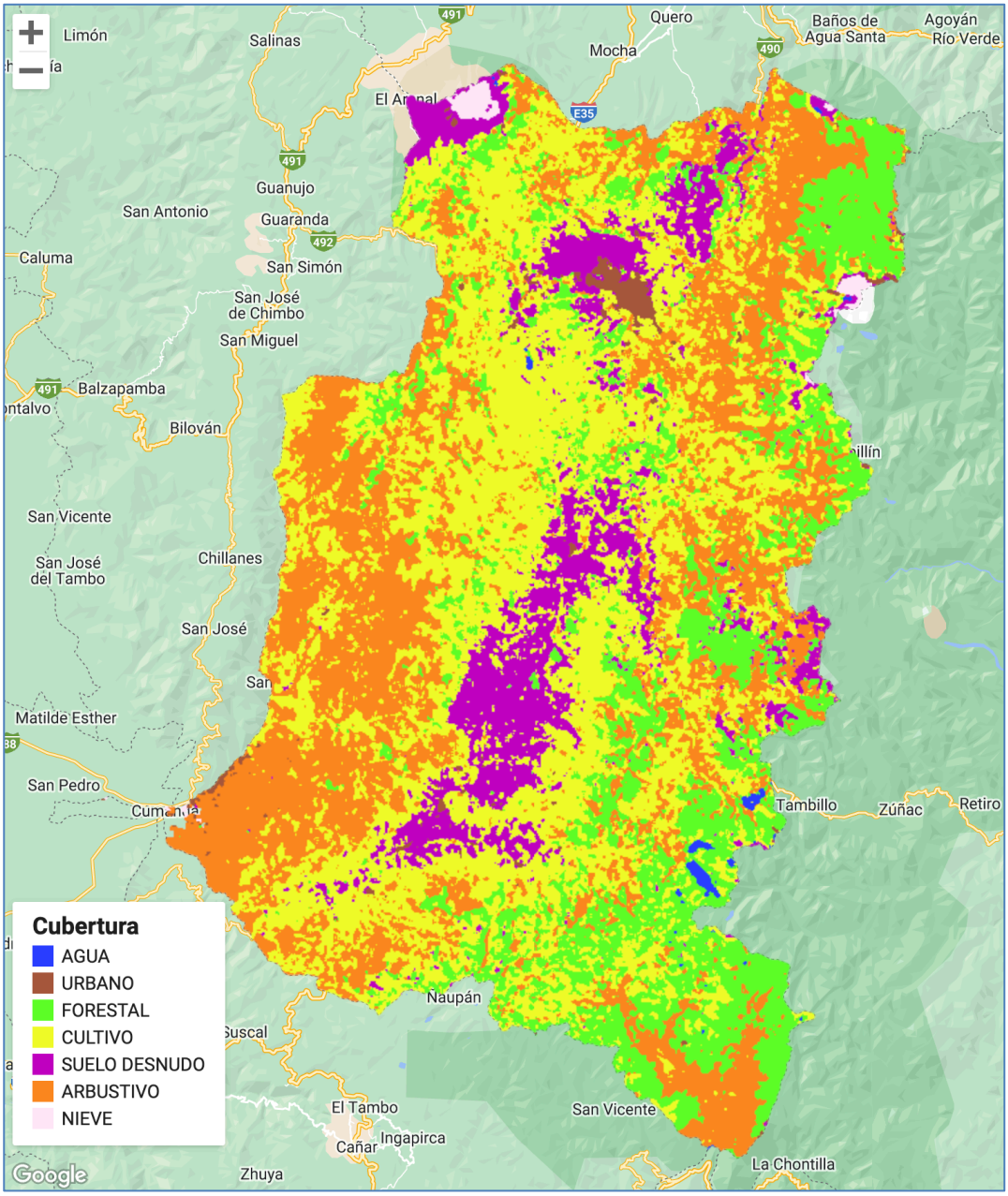


Figura 26. Mapa de clasificación algoritmo RF, criterio de prueba 2.

MATRIZ DE CONFUSIÓN									
		PREDICCIÓN							
		0	**1**	**2**	**3**	**4**	**5**	**6**	
		AGUA	URBANO	FORESTAL	CULTIVO	SUELO DESNUDO	ARBUSTIVO	NIEVE	**TOTAL**
REAL	**0** AGUA	**73**	0	1	1	4	1	0	**80**
	1 URBANO	0	**99**	0	1	6	0	2	**108**
	2 FORESTAL	0	0	**87**	3	3	7	0	**100**
	3 CULTIVO	0	2	0	**84**	2	12	0	**100**
	4 SUELO DESNUDO	0	3	2	8	**81**	1	0	**95**
	5 ARBUSTIVO	0	0	4	5	0	**68**	1	**78**
	6 NIEVE	1	0	0	0	0	0	**36**	**37**
	SUMA TOTAL	**74**	**104**	**94**	**102**	**96**	**89**	**39**	**598**
	Presición	99%	95%	93%	82%	84%	76%	92%	

Error de omisión	1%	5%	7%	18%	16%	24%	8%	
Error de comisión	9%	8%	13%	16%	15%	13%	3%	
Precisión General	**88,29**							

Tabla 35. Matriz de confusión algoritmo RF, criterio de prueba 2.

		MATRIZ DE OBSERVACIÓN							
		TP	**TN**	**FP**	**FN**	**precision**	**recall**	**f1-score**	**specificity**
0	AGUA	73	517	1	7	0,986486486	0,9125	0,948051948	0,998069498
1	URBANO	99	485	5	9	0,951923077	0,916666667	0,933962264	0,989795918
2	FORESTAL	87	491	7	13	0,925531915	0,87	0,896907216	0,985943775
3	CULTIVO	84	480	18	16	0,823529412	0,84	0,831683168	0,963855422
4	SUELO DESNUDO	81	488	15	14	0,84375	0,852631579	0,848167539	0,970178926
5	ARBUSTIVO	68	499	21	10	0,764044944	0,871794872	0,814371257	0,959615385
6	NIEVE	36	558	3	1	0,923076923	0,972972973	0,947368421	0,994652406
TOTAL						**0,88833468**	**0,890938013**	**0,888644545**	**0,980301619**

Tabla 36. Matriz de observación algoritmo RF, criterio de prueba 2.

índice kappa = 0.8620062173022974

ANÁLISIS

El algoritmo RF con 150 arboles de decisión con 88,29% de precisión general presenta un incremento de casi 1% respecto a la prueba 1, como primera apreciación se puede decir que el número de árboles implica el resultado de las métricas.

Con coeficiente o índice de kappa 0.862, RF con 150 arboles también recae en concordancia de clasificación casi perfecta, igual que la prueba 1; entonces tiene desempeño optimo en discriminar las 7 clases de coberturas de suelo para los que fue entrenado.

También observamos en la columna de especificidad (specificity) que tiene porcentajes altos en clasificar clases que efectivamente no corresponde. Un claro ejemplo es la clase Nieve (6) con 0,980 que determina que los datos fugados efectivamente no pertenecen a la categoría Agua.

3. 300 árboles para el bosque.

RESULTADOS

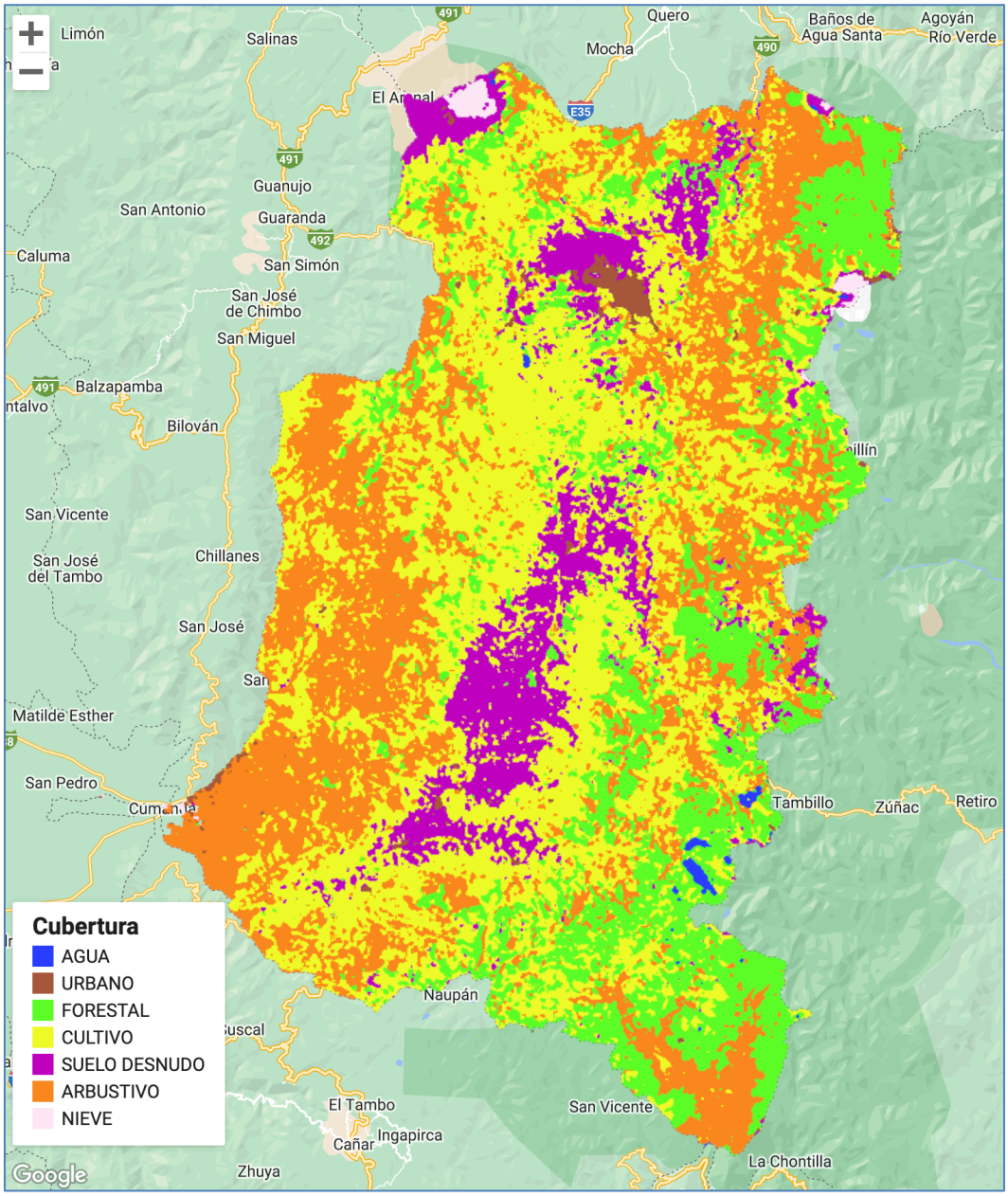


Figura 27. Mapa de clasificación algoritmo RF, criterio de prueba 3.

MATRIZ DE CONFUSIÓN									
			PREDICCIÓN							
			0	1	2	3	4	5	6	
			AGUA	URBANO	FORESTAL	CULTIVO	SUELO DESNUDO	ARBUSTIVO	NIEVE	TOTAL
REAL	0	AGUA	**72**	0	2	1	4	1	0	**80**
	1	URBANO	0	**100**	0	1	6	0	1	**108**
	2	FORESTAL	0	0	**87**	3	3	7	0	**100**
	3	CULTIVO	0	2	0	**85**	2	11	0	**100**
	4	SUELO DESNUDO	0	3	2	8	**82**	0	0	**95**
	5	ARBUSTIVO	0	0	4	4	0	**69**	1	**78**
	6	NIEVE	1	0	0	0	0	0	**36**	**37**

	SUMA TOTAL	**73**	**105**	**95**	**102**	**97**	**88**	**38**	**598**
	Presición	99%	95%	92%	83%	85%	78%	95%	
	Error de omisión	1%	5%	8%	17%	15%	22%	5%	
	Error de comisión	10%	7%	13%	15%	14%	12%	3%	
	Precisión General	**88,80**							

Tabla 37. Matriz de confusión algoritmo RF, criterio de prueba 3.

		MATRIZ DE OBSERVACIÓN							
		TP	**TN**	**FP**	**FN**	**precision**	**recall**	**f1-score**	**specificity**
0	AGUA	72	517	1	8	0,98630137	0,9	0,941176471	0,998069498
1	URBANO	100	485	5	8	0,952380952	0,925925926	0,938967136	0,989795918
2	FORESTAL	87	490	8	13	0,915789474	0,87	0,892307692	0,983935743
3	CULTIVO	85	481	17	15	0,833333333	0,85	0,841584158	0,965863454
4	SUELO DESNUDO	82	488	15	13	0,845360825	0,863157895	0,854166667	0,970178926
5	ARBUSTIVO	69	501	19	9	0,784090909	0,884615385	0,831325301	0,963461538
6	NIEVE	36	559	2	1	0,947368421	0,972972973	0,96	0,996434938
TOTAL						**0,894946469**	**0,895238883**	**0,894218204**	**0,981105717**

Tabla 38. Matriz de observación algoritmo RF, criterio de prueba 3.

índice kappa = 0.8678731957301007

ANÁLISIS

Con 300 arboles alcanzado el 88.80% en precisión general ha mejorado de manera significativa la clasificación de casos positivos; este resultado se refleja en la matriz de confusión (Tabla 36) y observación (Tabla 37) para las clases Agua (0), Urbano (1) y Nieve (6) con menores porcentajes en errores y mayores en las métricas.

3.8 Desarrollo de pruebas del algoritmo SVM

Para el algoritmo de SVM cuyo desempeño depende del núcleo (kernel), establecer los hiperparámetros es un paso crucial a la vez nada trivial. Por lo tanto, al usar un kernel de tipo RBF recibe dos hiperparámetros: gamma y cost, la utilidad de cada uno ya se ha explicado en el capítulo anterior.

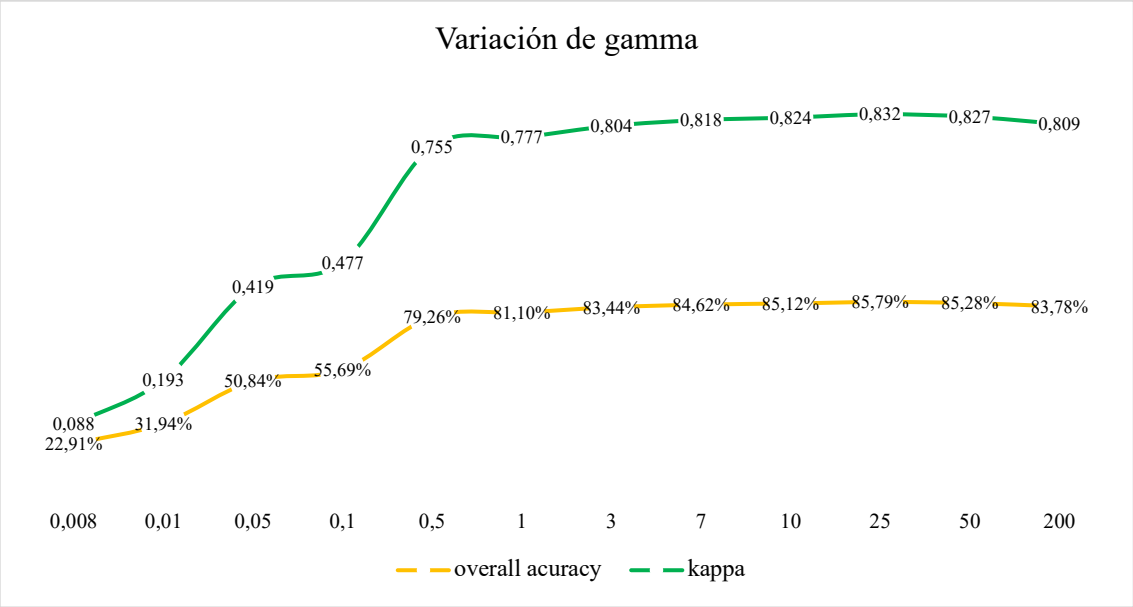
Debido que gamma y cost no se puede establecer al azar, en el presente estudio se ha realizado validaciones variando desde valores muy bajos hasta muy altos para estos dos hiperparámetros.

1. Variación de gamma

VARIACION DE GAMMA			
gamma	**cost**	**overall acuracy**	**kappa**
0,008	1	0,229	0,088

0,01	1	0,319	0,193
0,05	1	0,508	0,419
0,1	1	0,557	0,477
0,5	1	0,793	0,755
1	1	0,811	0,777
3	1	0,834	0,804
7	1	0,846	0,818
10	1	0,851	0,824
25	1	0,858	0,832
50	1	0,853	0,827
200	1	0,838	0,809

Tabla 39. Variación de heperparámetro gamma algoritmo SVM, criterio 1.



Gráfica 4. Variación de heperparámetro gamma algoritmo SVM, criterio 1.

En el primero criterio de prueba se ha variado el valor de gamma desde lo más bajo hasta lo más alto (Tabla 39, Gráfica 4). Con gamma=0.008 valor mínimo se obtiene precisión (0.229) muy baja, a medida que aumenta gamma también aumenta la precisión e índice kappa. Sin embargo, cuando el valor es muy alto gamma=200 nuevamente las métricas tienden a bajar (0.838). Con esto se valida la teoría de mientras más bajo o alto menor precisión.

Por lo tanto, al fijar en los valores intermedios de gamma (1,3,7,10) la precisión fluctúa entre el ≈80% y el ≈85%, esto indica que en este rango de valores gamma se concentra el valor optimo y no se recae en sobreajustes.

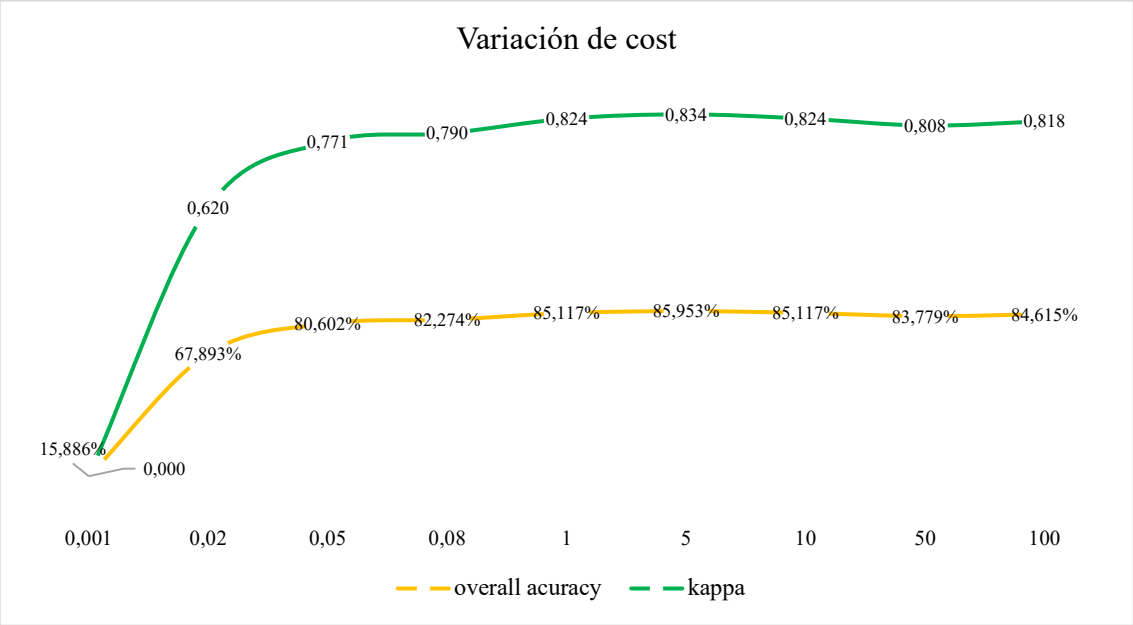
Por las observaciones expuestas y por su teoría, para el presente estudio se ha seleccionado gamma=10 que recae en los intermedios de cotas inferior y superior de

gamma. También se apoya la índice kappa con 0.824 que recae en el rango de clasificación casi perfecta.

2. Variación de cost

VARIACIÓN DE COST			
gamma	**cost**	**overall acuracy**	**kappa**
10	0	0,159	0,000
10	0,02	0,679	0,620
10	0,05	0,806	0,771
10	0,08	0,823	0,790
10	**1**	0,851	0,824
10	5	0,860	0,834
10	10	0,851	0,824
10	50	0,838	0,808
10	100	0,846	0,818

Tabla 40. Variación del hiperparámetro cost algoritmo SVM, criterio 2.



Gráfica 5. Variación de heperparámetro gamma algoritmo SVM, criterio 2.

Con el resultado de gamma=10 obtenido en criterio de prueba 1, en esta prueba la variación afecta a cost, empezando en coste=0,001 obteniendo la precisión general de 15.9% lo cual es muy bajo, asi mismo por el indice kappa=0 implica que los datos son totalmente independientes y no mantienen correlación.

Se ha incrementado cost hasta alcanzar la concordancia considerable y casi perfecta a partir de coste=0.02. Sin embargo al seguir aumentando se puede observar que tanto la precision y la concordancia tienden a bajar (coste=50, 100).

Observando la Tabla 40, y la curva de kappa en grafica 5, se ha seleccionado el valor de cost=1 que alcanza la precisión ≈0,851 e indice kapp ≈0,824 (rango de concordancia casi perfecta), que se encuentra en el intermedio de las cotas inferior y superior.

3. Svm con gamma y cost optimo

Por los criterios de prueba 1 y 2 se ha seleccionado y establecido los hiperparámetros con los valores: gamma = 10 y cost = 1.

Entonces es posible obtener los resultados de clasificación de coberturas utilizando el algoritmo SVM.

RESULTADO

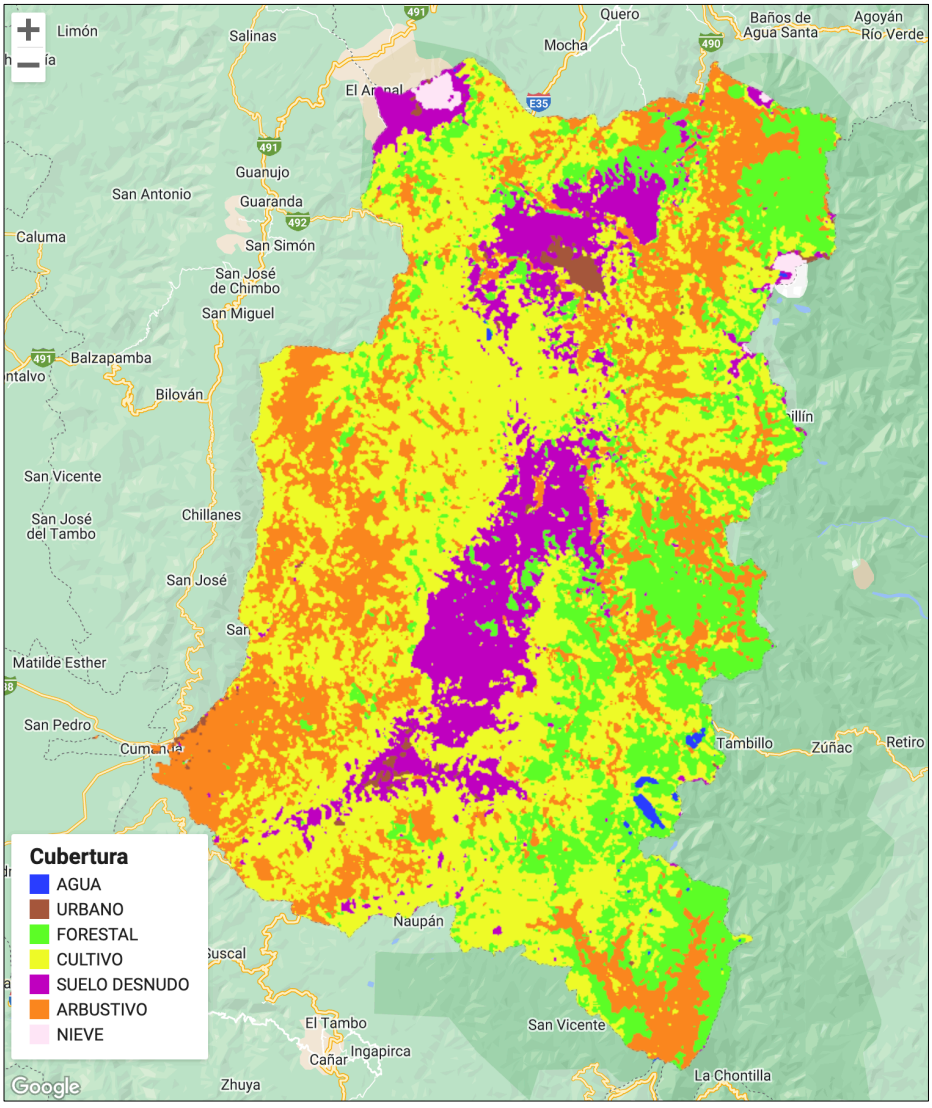


Figura 28. Mapa de clasificación algoritmo SVM, criterio de prueba 3.

MATRIZ DE CONFUSIÓN									
		PREDICCIÓN							
		0	1	2	3	4	5	6	
		AGUA	URBANO	FORESTAL	CULTIVO	SUELO DESNUDO	ARBUSTIVO	NIEVE	TOTAL
REAL	0 AGUA	**71**	0	4	0	3	2	0	**80**
	1 URBANO	0	**97**	0	0	10	1	0	**108**
	2 FORESTAL	0	0	**87**	3	5	5	0	**100**
	3 CULTIVO	0	3	2	**73**	3	19	0	**100**
	4 SUELO DESNUDO	0	3	4	3	**85**	0	0	**95**
	5 ARBUSTIVO	0	0	9	8	1	**60**	0	**78**
	6 NIEVE	1	0	0	0	0	0	**36**	**37**
	SUMA TOTAL	**72**	**103**	**106**	**87**	**107**	**87**	**36**	**598**
	Presición	99%	94%	82%	84%	79%	69%	100%	
	Error de omisión	1%	6%	18%	16%	21%	31%	0%	
	Error de comisión	11%	10%	13%	27%	11%	23%	3%	
	Precisión General	**85,12**							

Tabla 41. Matriz de confusión algoritmo SVM, criterio de prueba 3.

		MATRIZ DE OBSERVACIÓN							
		TP	**TN**	**FP**	**FN**	**precision**	**recall**	**f1-score**	**specificity**
0	AGUA	71	517	1	9	0,986111111	0,8875	0,934210526	0,998069498
1	URBANO	97	484	6	11	0,941747573	0,898148148	0,91943128	0,987755102
2	FORESTAL	87	479	19	13	0,820754717	0,87	0,844660194	0,96184739
3	CULTIVO	73	484	14	27	0,83908046	0,73	0,780748663	0,97188755
4	SUELO DESNUDO	85	481	22	10	0,794392523	0,894736842	0,841584158	0,956262425
5	ARBUSTIVO	60	493	27	18	0,689655172	0,769230769	0,727272727	0,948076923
6	NIEVE	36	561	0	1	1	0,972972973	0,98630137	1
TOTAL						**0,867391651**	**0,860369819**	**0,862029846**	**0,974842698**

Tabla 42. Matriz de observación algoritmo CART, criterio de prueba 3.

índice kappa = 0.8244292184721759

ANÁLISIS

Los hiperparámetros gamma y cost asignados en la Figura 28, Tabla 41 y Tabla 42 presentan los resultados de mapa de cobertura terrestre, matriz de confusión y matriz de observación, respectivamente; de estos resultado se ha determinado que:

- Se ha alcanzado la precisión general de 85.12%, con casos más altos de verdaderos positivos en la clase 0 (Agua), 1 (Urbano) y 6 (Nieve) este último alcanza la precisión de 100% evaluado de forma individual que se presenta en la matriz de observación.
- La clase 5 (Arbustivo) con f1-score 0.727 presenta la más baja de las métricas, esto tiene sentido debido que su precisión = 0.689 y recall=0.769 tiene

concordancia en el balance de resultados debido que no existe distancia muy signiicativa entre sus rangos de resultado.

- Otro de los resultados a abordar recae en la clase 6 (Nieve) con metricas igual a 1 (100%) o muy cercanos, al verificar los verdaderos positivos y falsos negativos se observa que solamente 1 (FN=1) dato se ha clasificado como no correspondiente a la clase 6.
- Por ultimo, con indice de kappa = 0.824 el algoritmo de clasificación SVM con kernel=RBF, gamma=10 y cost=1, alcanza la concordancia de clasificación casi perfecta conforme a las tabla de valoraciones del coeficiente de kappa.

3.9 Comparación de algoritmos CART, RF y SVM.

Los casos de prueba ejecutados han arrojado tres resultados para cada algoritmo de clasificación, en base a los análisis expuestos se ha seleccionado un caso de prueba para los algoritmos con desempeño optimo según los valores de las métricas (Tabla 43).

Algoritmo	Parametros	Datos prueba	overall acuracy	precisio n (avg)	recall (avg)	f1-score (avg)	specificit y (avg)	kappa
CART	300 árboles 50 hojas.	598	0,776	0,791	0,785	0,785	0,962	0,735
RF	300 árboles	598	0,888	0,895	0,895	0,894	0,981	0,868
SVM	kernerl=RBF gamma=10 cost=1	598	0,851	0,867	0,860	0,862	0,975	0,824

Tabla 43. Matriz de comparación de los algoritmos CART, RF y SVM.

Se ha desarrollado la matriz comparativa con las métricas de cada algoritmo de clasificación. Los resultados de precisión, recall, f1-score, specificity corresponde a los valores promedios obtenidos de la matriz de observación del caso de prueba correspondiente para cada clase de cobertura terrestre.

De la matriz comparativa se resalta lo siguiente:

- La clasificación con bajas métricas se ha presentado para las clases 5 (Arbustivo), 2 (Forestal) y 3 (Cultivo) por a la similitud en el uso de bandas NDVI, B3 y B6.
- Según las pruebas realizadas el algoritmo Random Forest alcanza mayores métricas respecto a CART y SVM.
- Los algoritmos RF y SVM alcanzan el rango de clasificación casi perfecta según los intervalos de valoración del coeficiente kappa.

- El algoritmo CART tiene las métricas más bajas, esto está relacionada en dependencia del grado discriminante que define por la cantidad de árboles y hojas por árbol de decisión.

3.10 Prueba de hipótesis

Según Sampieri et al [73], las investigaciones donde cuya finalidad es comparar, se plantea la hipótesis de la diferencia entre grupos, que a su vez algunos investigadores consideran también como tipo correlacional que implica la relación de dos o más variables.

Por lo expuesto en lo anterior y según los objetivos del presente estudio se plante las siguientes hipótesis:

Descripción:

> "Las imágenes satelitales proporcionan suficiente información discriminatoria para la clasificación de la cobertura del suelo, mediante aprendizaje automático."

Hipótesis:

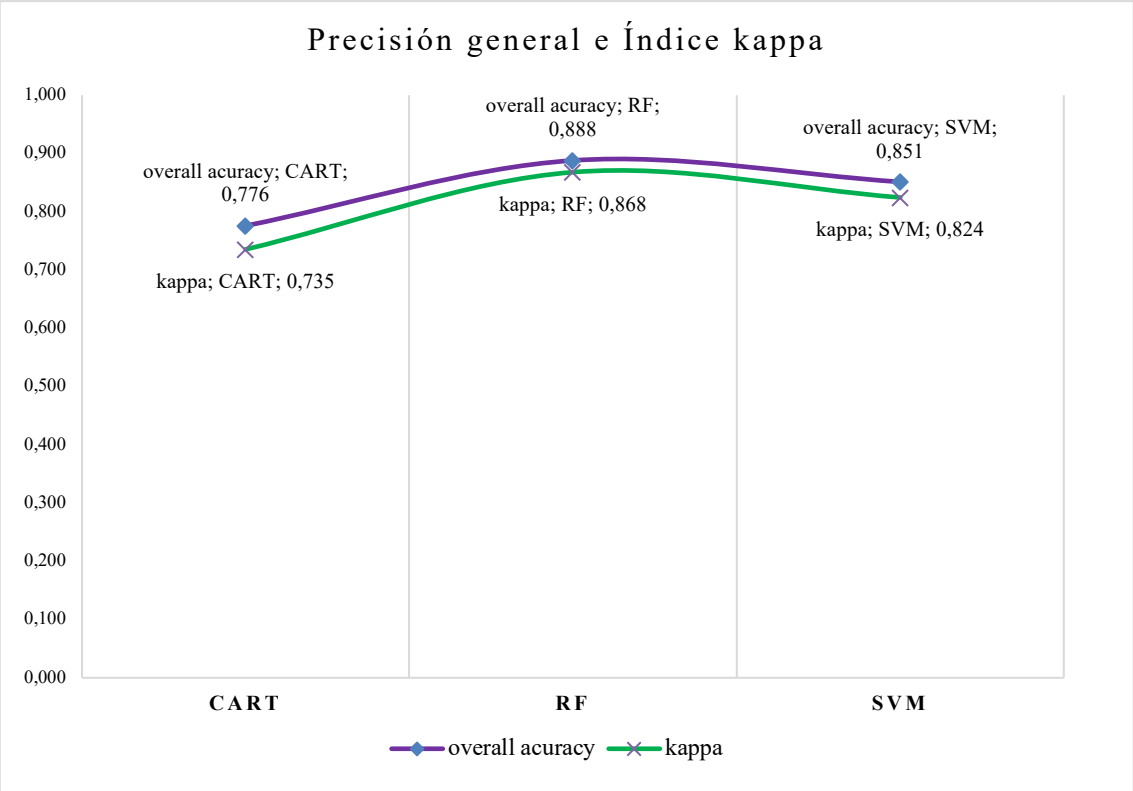
> **H_0**: kappa = 0.
> **H_1**: kappa != 0.

Donde:

> H_0: Hipótesis nula.
> H_1: Hipótesis alternativa.
> kappa: Coeficiente o índice kappa.

Inferencia estadística:

En el presente capitulo 3.3. Métricas, se ha descrito el fundamento teórico del índice o coeficiente kappa. A demás, conforme a al rango para valoraciones de concordancia desarrollada en [80], [81] se ha definido como métrica de validación el índice kappa para determinar las concordancias y correlaciones de las 7 clases de coberturas del suelo clasificadas para la provincia de Chimborazo mediante ML.



Gráfica 6. Comparación de algoritmos, precisión general e índice kappa.

Con los datos de la matriz de comparación de la Tabla 43, se ha graficado los valores de las métricas: precisión general (overall acuracy) e índice kappa de los tres algoritmos. Esto permite inferir el rango de concordancia de kappa definida en el eje de las ordenadas (y) en el intervalo [0-1].

Según los valores de índice kappa:

CART	kappa = 0,735	=>	kappa != 0
RF	kappa = 0,868	=>	kappa != 0
SVM	kappa = 0,824	=>	kappa != 0

Por definición: kappa < 0 no existe concordancia, kappa = 0 son valores independientes y conforme resultados de kappa para los tres algoritmos de clasificación: kappa != 0 y a su vez kappa > 0, para CART concordancia "considerable" para RF y SVM concordancia "casi perfecta"; se rechaza la hipótesis nula H_0 y se acepta la hipótesis alternativa H_1. Entonces, se afirma que: "Las imágenes satelitales proporcionan suficiente información discriminatoria para la clasificación de la cobertura del suelo, mediante aprendizaje automático".

CAPITULO IV

4. DISCUSIÓN DE RESULTADOS

Durante el desarrollo del presente estudio se ha ido determinando hallazgos significativos que aportan fundamentos teóricos y prácticos para la línea de investigación que persigue el análisis y procesamiento imágenes satelitales mediante la aplicación de ML para clasificación de coberturas terrestres. Entonces, la TI ha jugado un rol determinante con prestaciones de almacenamiento y procesamiento en la nueve como la plataforma GEE.

En el presente capitulo se discute los principales resultados desde el punto de vista crítico, además se apoya en las informaciones y teorías descritas en el Capítulo I y Capitulo II.

4.1 Como se ha desarrollado los modelos ML

Mediante la SLR se había determinado la carencia de información geográfica referente a tipos de coberturas del suelo en la región de provincia de Chimborazo, considerando que la aplicación de ML es joven puede ser uno de los factores de la carencia de este tipo de informaciones en regiones o entornos específicos.

Una de las limitantes que es la infraestructura computacional para ML debido que demanda alta en prestaciones para procesamiento de imágenes satelitales de altas resoluciones. Por lo tanto se ha seleccionado la plataforma Google Earth Engine como fuente de datos y procesamiento en la nube, también utilizada en trabajos previos por [1], [7], [9]–[12].

4.2 Construcción de los dataset

De las 13 bandas de Sentinel 2 se ha utilizado 7 bandas ["B2", "B3", "B4", "B6", "B8", "B11", "BSI"] además de las composiciones RGB, NDVI y BSI que permite la discriminación de las 7 clases [0 Agua, 1 Urbano, 2 Forestal, 3 Cultivo, 4 Suelo Desnudo, 5 Arbustivo, 6 Nieve] de coberturas del suelo definidos para la provincia de Chimborazo.

Para replicación en estudios similares cabe señalar que la selección de las clases de coberturas es dependiente de los objetivos de la investigación y el conocimiento de la región geográfica por parte del investigador. Por ejemplo, un caso particular a destacar es la definición de la clase 6 (nieve) que no en todas las regiones está presente.

```
// Se obtiene los predictores basados en la propiedad target de los datos
var all_data = data_composition.select(bands).sampleRegions({
  collection: data,
  properties: ["target"],
  scale: 10
}).randomColumn();

// Porcentaje para entrenamiento y prueba
var split = 0.8;
var training_data = all_data.filter(ee.Filter.lt("random", split));
var validation_data = all_data.filter(ee.Filter.gte("random", split));
```

Tabla 44. Asignación aleatoria de datos para entrenamiento y prueba

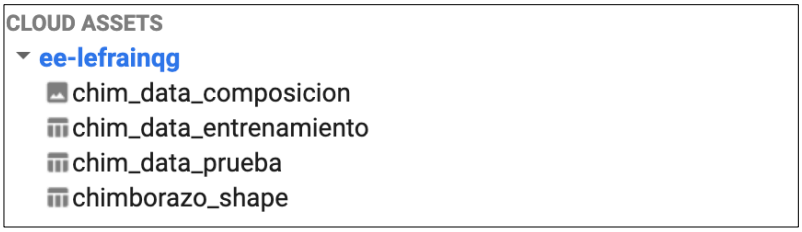


Figura 29. Separación de datos para entrenamiento y prueba.

Se ha construido el dataset general, datos de entrenamiento y prueba de manera independiente para tener la disponibilidad de utilizar para cualquier caso práctico en la implementación de los algoritmos ML. Para tener mayor control y que el ordenador no decida los porcentajes para entrenamiento y prueba de manera aleatoria (Tabla 44), no se unifica los datos de entrenamiento y prueba en un solo dataset.

4.3 Desempeño de algoritmos de clasificación ML

CART

Mediante el procesamiento de los 3 casos de prueba del algoritmo CART se ha podido determinar la afectación en mayor medida del parámetro cantidad de hojas, con numero de hojas menor o igual a 10 alcanza el ≈79,93%. En contraste, mientras se sube el número de hojas la precisión tiende a bajar, como se evidencia en los casos de prueba 2 y 3. En síntesis, mientras menor número de hojas implica mayor precisión y viceversa. Por lo tanto, a mayor número de hojas mejor discriminación.

RF

Con los resultados de las 3 pruebas ejecutadas con 50, 150, 300 arboles de decisión, se evidencia cuanto mayor número sea este parámetro mejor desempeño del clasificador RF, con pruebas extras realizadas con parámetro mayor a 300, por ejemplo 500 arboles el algoritmo tienen a bajar en las métricas.

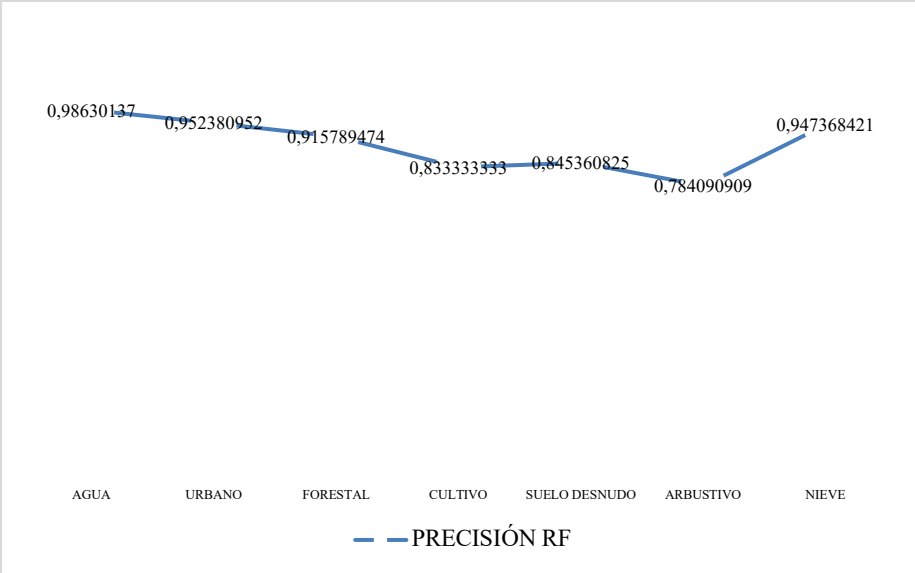
SVM

Al igual que los dos algoritmos anteriores después de los tres casos de prueba para el algoritmo SVM es primordial la búsqueda de los hiperparámetros gamma y cost, estos valores generalmente se concentran entre las cotas inferior y superior, se selecciona el valor intermedio de las cotas que permita el desempeño optimo del clasificador sin caer en el sobreajuste.

Para el caso de clasificación de coberturas de la provincia de Chimborazo se ha definido gamma=10 y cost=1, obtenidos ejecutando los casos de prueba 1, 2 y 3, como resultado rondan entre el ≈80% y ≈85% la métrica de precisión general e índice kappa en rango de concordancia casi perfecta.

4.4 Clasificador optimo

La selección de la clasificación optimo depende de las métricas que arroja el algoritmo después del proceso de entrenamiento y prueba con datos de prueba. También se debe considerar la cantidad y calidad de datos, los parámetros, las bandas utilizadas, los tipos de coberturas definidas.



Gráfica 7. Precisión de cada categoría, algoritmo RF.

Para el caso de clasificación de coberturas del suelo en la provincia de Chimborazo, el algoritmo que mejor desempeño tuvo es Random Forest con 300 arboles de decisión, alcanzando la precisión general de ≈88,80% e índice kappa de 0.867 que se concentra en el rango de valoración de clasificación casi perfecta u optima.

También se debe tener en cuenta los tipos de cobertura que se defina, en el caso del presente estudio se ha podido notar que las clases: forestal, cultivo y arbustivo tienen dificultad en clasificación debido que las bandas B2, B4, B6 y B11discriminan tipos de cobertura asociados a vegetación.

Para una muestra se presenta los datos observados de la precisión general de cada clase de cobertura del algoritmo RF con 300 arboles (Gráfica 7). Se puede observar que su mayor medida la clase Forestal tiene menor precisión con 0.784; mismo comportamiento sucede en los algoritmos CART y SVM que se puede observar en la matriz de observación de los casos de prueba.

4.5 Trabajos futuros

En el área tecnológica de ML al ser una disciplina joven a la vez con mucho progreso el presente trabajo representa un aporte a la línea investigativa en tratamiento de imágenes mediante teledetección y ML. Esto a su vez puede derivar en estudios futuros que completen el abanico de posibilidades y aportes en la explotación de los recursos y uso del suelo apoyados mediante clasificaciones de las coberturas terrestres.

- **Comparación de variabilidad de coberturas**

El presente trabajo aplica y hace comparativa de los algoritmos de clasificación, tomando como referencia desde esta aplicación práctica se puede desarrollar modelos comparativos de cambios de coberturas terrestres a través de tiempo. Esto construye a determinar las tendencias de cambios de la naturaleza para el cuidado del medio ambiente o aprovechamiento responsable de los recursos naturales.

- **Implantación de modelos ML en para soporte de decisiones**

Debido a las prestaciones tecnológicas, computación en la nube y desarrollo de modelos ML, es viable desplegar o implementar en sistemas de soporte de decisiones mediante la construcción de interfaces web o modelos GIS que permita el monitoreo de tipos de suelo que contribuya la toma de decisiones a instituciones seccionales o gubernamental en ámbito público o privado.

CONCLUSIONES

- Conforme al primer objetivo específico planteado, aplicando la revisión sistemática de literatura se ha investigado estudios previos relacionados, esto ha permitido definir el caso de estudio en la provincia de Chimborazo, fuente de datos con imágenes satelitales de S2, herramientas tecnológicas basada en GEE y metodología CRISP-DM, para el desarrollo d los modelos de clasificación ML.

- Según el segundo objetivo específico se ha construido los conjuntos de datos en la plataforma GEE mediante la limpieza de datos, utilizando 7 bandas de S2 y sus composiciones, 1581 datos de entrenamiento, 598 datos de prueba y 7 clase de coberturas del suelo que sirve como datos de entrada en el desarrollo de los algoritmos ML.

- Para cumplir el tercer objetivo se ha desarrollado los modelos ML con los algoritmos de clasificación CART, RF y SVM ejecutando 3 criterios de prueba para cada uno. RF con precisión general ≈88,80% e índice kappa de 0.867 es el algoritmo con mejores desempeños en base a resultado de las métricas. Además, los tipos de cobertura forestal, cultivo y arbustivo tiene métricas bajas debido al uso común de las bandas de discriminación.

- El estudio desarrollado expone mapas de coberturas terrestres de la provincia de Chimborazo de los dos últimos años, lo que permite tener conocimiento y hacer uso de la información para el uso y explotación de la tierra.

RECOMENDACIONES

- El conocimiento de la aplicación de modelos ML es de vital, por lo que se recomienda explorar, analizar y dicernir mediate SLR que permita seleccinar de manera correcta las herramientas, fuente de datos y metodos para el desarrollo de los modelos ML.

- Siendo los recursos computacionales como limitantes en el desarrollo de modelos ML, se sugiere la utilizacion de la plataforma GEE, que provee almacen de datos y entornos integrados de desarrollo en la nube de manera libre en caso de estudios que persigan el enfoque del presente trabajo.

- Se recomienda la aplicación de los algoritmos RF y SVM en clasificación de coberturas terrestres mediante aprendizaje supervisado, debido que tienen mayor fuerza de discriminación sin perder de vista la importancia de los valores de los parámetros que incide en el desempeño.

- Para los estudios futuros se recomienda implementar los modelos ML desarrolados en sistemas de sopote de desciciones para seguimiento de cambio de tipos de suelo y aprovechamiento en uso y explotacion de la tierra.

BIBLIOGRAFÍA

[1] X.-P. Song *et al.*, «An assessment of global forest cover maps using regional higher-resolution reference data sets», en *2011 IEEE International Geoscience and Remote Sensing Symposium*, 2011, pp. 752-755.

[2] M. M. Mishaa, A. D. Andrushia, y T. M. Neebha, «Image based Land Cover Classification for Remote Sensing Applications-A review», en *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, 2021, pp. 152-155.

[3] R. Ravanelli, A. Nascetti, y M. Crespi, «Large Scale Assessment of Free Global DEMs Through the Google Earth Engine Platform», en *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 5242-5245.

[4] R. Yu, G. Wang, T. Shi, W. Zhang, C. Lu, y T. Zhang, «Potential of Land Cover Classification Based on GF-1 and GF-3 Data», en *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2747-2750.

[5] D. Wuyun *et al.*, «The spatiotemporal change of cropland and its impact on vegetation dynamics in the farming-pastoral ecotone of northern China», *Sci. Total Environ.*, vol. 805, p. 150286, sep. 2021.

[6] E. Barrientos-Ávila1 y M. Moya-Calderón2, «El Efecto de la Cobertura del Suelo en la Variación de las Temperaturas Locales; Naranjo, Alajuela, Costa Rica, 2016», *Revista Geográfica de América Central*, vol. 2, n.º 61, pp. 205-219, 2018.

[7] Marco Javier Castelo-Cabay, Gustavo Iván Buñay-Gualoto, Byron Geovanny Pillajo-Landa I, «Uso de Redes Neuronales Artificiales y Computación en la Nube para clasificar la cobertura del suelo en territorio ecuatoriano», *Polo de Conocimiento*, vol. 6, n.º 5, p. 15, 2021.

[8] Scarlet Cartaya Ríos, Shirley Zurita Alfaro, Elvira Rodríguez Ríos, Víctor Montalvo Párraga, «Comparación de técnicas para determinar cobertura vegetal y usos de la tierra en áreas de interés ecológico, Manabí, Ecuador», *REVISTA UD Y LA GEOMÁTICA*, 2014.

[9] W. Wu, X. Zhao, C. Gong, y X. Li, «Obtain the Patterns of Global Forest NPP and its Influence Factors with Google Earth Engine», en *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 2898-2901.

[10] A. S. Suárez L, A. F. Jiménez L, M. Castro-Franco, y A. Cruz-Roa, «Clasificación y mapeo automático de coberturas del suelo en imágenes satelitales utilizando Redes Neuronales Convolucionales», *Orinoquia*, vol. 21, n.º 1, pp. 64-75, 2017.

[11] U. Pimple *et al.*, «Google earth engine based three decadal Landsat imagery analysis for mapping of mangrove forests and its surroundings in the trat province of Thailand», *J. Comput. Commun.*, vol. 06, n.º 01, pp. 247-264, 2018.

[12] T. Mayer *et al.*, «Deep learning approach for Sentinel-1 surface water mapping leveraging Google Earth Engine», *ISPRS Open Journal of Photogrammetry and Remote Sensing*, vol. 2, p. 100005, dic. 2021.

[13] Sistema Nacional de Información, «Mapa de Cobertura y Uso de la Tierra - Sistema Nacional de Información». [En línea]. Disponible en: https://sni.gob.ec/mapa-cobertura-uso. [Accedido: 02-nov-2021].

[14] P. de Chimborazo, «PLAN DE DESARROLLO Y ORDENAMIENTO TERRITORIAL 2020-2030», p. 681, 2020.

[15] Mae-Magap, «PROTOCOLO METODOLÓGICO PARA LA ELABORACIÓN DEL MAPA DE COBERTURA Y USO DE LA TIERRA DEL ECUADOR CONTINENTAL 2013-2014, ESCALA 1:100.00», may 2015.

[16] S. Pineda y S. Carolina, «Comparación de árboles de regresión y clasificación y regresión logística», 2009.

[17] L. Breiman, «Random Forests», *Mach. Learn.*, vol. 45, n.º 1, pp. 5-32, oct. 2001.

[18] C. Cortes y V. Vapnik, «Support-vector networks», *Mach. Learn.*, vol. 20, n.º 3, pp. 273-297, sep. 1995.

[19] K. V. Suresh Babu y V. S. K. Vanama, «Burn area mapping in Google Earth Engine (GEE) cloud platform: 2019 forest fires in eastern Australia», en *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, 2020, pp. 109-112.

[20] G. Mateo-Garcia, J. Muñoz, y L. Gómez-Chova, «Cloud detection on the Google Earth engine platform», en *IGARSS 2017 - 2017 IEEE International Geoscience and Remote Sensing Symposium*, 2017, pp. 1942-1945.

[21] C. Zheng y L. Wang, «Semantic Segmentation of Remote Sensing Imagery Using Object-Based Markov Random Field Model With Regional Penalties», *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, n.º 5, pp. 1924-1935, may 2015.

[22] K. Fukushima, «Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position», *Biol. Cybern.*, vol. 36, n.º 4, pp. 193-202, 1980.

[23] Y. LeCun, L. Bottou, Y. Bengio, y P. Haffner, «Gradient-based learning applied to document recognition», *Proceedings of the IEEE*, 1998.

[24] C. Zhang *et al.*, «Joint Deep Learning for land cover and land use classification», *Remote Sens. Environ.*, vol. 221, pp. 173-187, feb. 2019.

[25] K. Jue, B. Zhong, y A. Yang, «Production of historical classification products based on existing land cover classification products and Google Earth Engine platform», en *2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, 2019, pp. 1-4.

[26] F. L. R. Barbosa, R. F. Guimarães, O. A. de Carvalho, y R. A. T. Gomes, «Land Use/Land Cover (LULC) classification based on SAR/Sentinel 1 image in Distrito Federal, Brazil», *Sociedade & Natureza*, vol. 33, 2021.

[27] A. Ramanath, S. Muthusrinivasan, Y. Xie, S. Shekhar, y B. Ramachandra, «NDVI Versus CNN Features in Deep Learning for Land Cover Clasification of Aerial Images», en *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 6483-6486.

[28] O. Ahlqvist, D. Varanka, S. Fritz, y K. Janowicz, *Land use and land cover semantics principles, best practices, and prospects*. CRC Press Taylor & Francis Group, 2016.

[29] «Análisis geoespacial de la interacción entre el uso de suelo y de agua en el área peri-urbana de Cuauhtémoc, Chihuahua. Un estudio socioambiental en el norte de México», *Investigaciones Geográficas, Boletín del Instituto de Geografía*, vol. 2014, n.º 83, pp. 116-130, abr. 2014.

[30] J. Peña, R. M. Poveda, A. Bonet, J. Bellot, y A. Escarré, «CARTOGRAFÍA DE LAS COBERTURAS Y USOS DEL SUELO DE LA MARINA BAIXA (ALICANTE) PARA 1956, 1978 Y 2000», *Investigaciones Geográficas (Esp)*, n.º 37, pp. 93-107, 2005.

[31] J. B. Campbell y R. H. Wynne, *Introduction to Remote Sensing, Fifth Edition*. Guilford Press, 2011.

[32] H. Cotler, E. Sotelo, J. Dominguez, M. Zorrilla, S. Cortina, y L. Quiñones, «La conservación de suelos: un asunto de interés público», *Gaceta Ecológica*, n.º 83, pp. 5-71, 2007.

[33] R. R. Sokal, «Classification: purposes, principles, progress, prospects», *Science*, vol. 185, n.º 4157, pp. 1115-1123, sep. 1974.

[34] A. Tejedor De León, «Uso de software para el procesamiento de imágenes digitales para la definición de cuencas hidrográficas».

[35] R. C. Gonzalez y R. E. Woods, *Digital Image Processing*. Pearson, 2018.

[36] B. L. Markham, J. C. Storey, D. L. Williams, y J. R. Irons, «Landsat sensor performance: history and current status», *IEEE Trans. Geosci. Remote Sens.*, vol. 42, n.º 12, pp. 2691-2694, dic. 2004.

[37] «Landsat Science - About», *Landsar Science*. [En línea]. Disponible en: https://landsat.gsfc.nasa.gov/about. [Accedido: 07-nov-2021].

[38] U. Nasa, «Landsat Benefiting Society for Fifty Years».

[39] H. Van der Werff y F. Van der Meer, «Sentinel-2A MSI and Landsat 8 OLI Provide Data Continuity for Geological Remote Sensing», *Remote Sensing*, vol. 8, n.º 11, p. 883, oct. 2016.

[40] «Landsat 8», *NASA, LandSat Science*. [En línea]. Disponible en: https://landsat.gsfc.nasa.gov/landsat-8. [Accedido: 07-nov-2021].

[41] Nasa y Usgs, «Landsat 8», *Geological Survey*, 2013.

[42] «Sentinel-2 delivers first images», *THE EUROPEAN SPACE AGENCY*. [En línea]. Disponible en: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2/Sentinel-2_delivers_first_images. [Accedido: 06-nov-2021].

[43] M. Berger, J. Moreno, J. A. Johannessen, P. F. Levelt, y R. F. Hanssen, «ESA's sentinel missions in support of Earth system science», *Remote Sens. Environ.*, vol. 120, pp. 84-90, may 2012.

[44] M. Drusch *et al.*, «Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services», *Remote Sens. Environ.*, vol. 120, pp. 25-36, may 2012.

[45] ESA, «The Sentinel missions», 2022. [En línea]. Disponible en: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/The_Sentinel_missions. [Accedido: 12-oct-2022].

[46] ESA, «Sentinel-2 - Mission Objectives», 2022. [En línea]. Disponible en: https://sentinel.esa.int/web/sentinel/missions/sentinel-2/mission-objectives. [Accedido: 12-oct-2022].

[47] G. Kaplan y U. Avdan, «Object-based water body extraction model using Sentinel-2 satellite imagery», *European Journal of Remote Sensing*, vol. 50, n.º 1, pp. 137-143, mar. 2017.

[48] A. Stumpf, D. Michéa, y J.-P. Malet, «Improved Co-Registration of Sentinel-2 and Landsat-8 Imagery for Earth Surface Motion Measurements», *Remote Sensing*, vol. 10, n.º 2, p. 160, ene. 2018.

[49] C. Soto y C. Jiménez, «APRENDIZAJE SUPERVISADO PARA LA DISCRIMINACIÓN Y CLASIFICACIÓN DIFUSA», *Dyna*, vol. 78, n.º 169, pp. 26-33, 2011.

[50] A. L. Samuel, «Some Studies in Machine Learning Using the Game of Checkers», *IBM J. Res. Dev.*, vol. 3, n.º 3, pp. 210-229, jul. 1959.

[51] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc, 2019.

[52] C. G. Cambronero y I. G. Moreno, «Algoritmos de aprendizaje: knn & kmeans», *Intelgencia en Redes de Comunicación*, 2006.

[53] C. J. Carmona, F. Pulgar, A. M. Garcia, P. Gonzalez, y M. J. del Jesus, «Análisis descriptivo mediante aprendizaje supervisado basado en patrones emergentes». [En línea]. Disponible en: https://simidat.ujaen.es/sites/default/files/biblio/2015%20-%20TAMIDA-a.pdf. [Accedido: 05-nov-2021].

[54] C. Bruno, M. Balzarini, y S. Rosales Heredia, «IDENTIFICACIÓN DE RELACIONES ENTRE RENDIMIENTOS Y VARIABLES AMBIENTALES VÍA ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN (CART)», *Interciencia*, vol. 35, n.º 12, pp. 876-882, 2010.

[55] G. A. Betancourt, «LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs)», *Scientia Et Technica*, vol. XI, n.º 27, pp. 67-72, 2005.

[56] J. Jara Estupiñan, D. Giral, y F. Martínez Santa, «Implementación de algoritmos basados en máquinas de soporte vectorial (SVM) para sistemas eléctricos: revisión de tema», *Tecnura*, vol. 20, n.º 48, pp. 149-170, 2016.

[57] D. Amaya Hurtado, O. F. Aviléres, y L. F. Niño Sierra, «Pruebas de estanqueidad en envases de tereftalato de polietileno basado en máquina de soporte vectorial», *Ingeniare. Revista Chilena de Ingeniería*, vol. 23, n.º 4, pp. 630-637, 2015.

[58] R. G. A. B. M. T. J. del Cerro, «Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial», *Revista Iberoamericana de Automática e Informática Industrial RIAI*, vol. 14, n.º 4, pp. 394-405, oct. 2017.

[59] A. Elen, S. Baş, y C. Közkurt, «An Adaptive Gaussian Kernel for Support Vector Machine», *Arab. J. Sci. Eng.*, vol. 47, n.º 8, pp. 10579-10588, ago. 2022.

[60] S. C. S. Pineda, «Comparacion de Arboles de Regresion y Clasificacion y regresion logıstica». [En línea]. Disponible en: https://repositorio.unal.edu.co/bitstream/handle/unal/2421/42694070_2009.pdf?sequence=1&isAllowed=y. [Accedido: 14-oct-2022].

[61] S. M. Hamze-Ziabari y T. Bakhshpoori, «Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5′ and CART algorithms», *Appl. Soft Comput.*, vol. 68, pp. 147-161, jul. 2018.

[62] J. J. Espinosa-Zúñiga, «Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito», *Ingeniería. Investigación y Tecnología*, vol. XXI, n.º 3, 2020.

[63] B. Takoutsing y G. B. M. Heuvelink, «Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors», *Geoderma*, p. 116192, oct. 2022.

[64] J. A. Valero Medina y B. E. Alzate Atehortúa, «Comparison of maximum likelihood, support vector machines, and random forest techniques in satellite images classification», *Available in: https:///articulo.oa?id=*, vol. 25705, p. 954, 2002.

[65] D. C. Marinescu, «Chapter 1 - Introduction», en *Cloud Computing*, D. C. Marinescu, Ed. Boston: Morgan Kaufmann, 2013, pp. 1-19.

[66] Google, «Google Earth Engine». [En línea]. Disponible en: https://earthengine.google.com/. [Accedido: 07-nov-2022].

[67] C. Schröer, F. Kruse, y J. M. Gómez, «A Systematic Literature Review on Applying CRISP-DM Process Model», *Procedia Comput. Sci.*, vol. 181, pp. 526-534, ene. 2021.

[68] S. Huber, H. Wiemer, D. Schneider, y S. Ihlenfeldt, «DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model», *Procedia CIRP*, vol. 79, pp. 403-408, ene. 2019.

[69] M. Hinojosa, I. Derpich, M. Alfaro, D. Ruete, A. Caroca, y G. Gatica, «Procedimiento de agrupación de estudiantes según riesgo de abandono para mejorar la gestión estudiantil en educación superior», *Texto Livre: Linguagem e Tecnologia*, vol. 15, pp. 1-22, 2022.

[70] M. Gallego Gallego y J. Hernández Cáceres, «Identificación de factores que permitan potencializar el éxito de proyectos de desarrollo de software», *Scientia Et Technica*, vol. 20, n.º 1, pp. 70-80, 2015.

[71] C. T. D. U. Y. G. del Suelo, «RESOLUCIÓN Nro. 003-CTUGS-2019», p. 19, 2019.

[72] M. M. G. Pnud, «MAPA DE COBERTURA Y USO DE LA TIERRA DEL ECUADOR CONTINENTAL ESCALA 1:100 000, AÑO 2013-2014». 2014.

[73] R. H. Sampieri, C. F. Collado, y P. B. Lucio, *Metodología de la investigación*. McGraw-Hill Education, 2014.

[74] Chimborazo Travel, «Chimborazo Travel – Conoce la Provincia de las Cumbres Andinas», *Chimborazo Travel*. [En línea]. Disponible en: https://chimborazo.travel/. [Accedido: 08-nov-2022].

[75] Sinergise, «A repository of custom scripts that can be used with Sentinel-Hub services», *custom-scripts*. [En línea]. Disponible en: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/bands/. [Accedido: 08-nov-2022].

[76] J. M. S. Muñoz, «Análisis de Calidad Cartográfica mediante el estudio de la Matriz de Confusión», *Pensamiento matemático*, vol. 6, n.º 2, pp. 9-26, 2016.

[77] J. Cohen, «A Coefficient of Agreement for Nominal Scales», *Educ. Psychol. Meas.*, vol. 20, n.º 1, pp. 37-46, abr. 1960.

[78] W. Al-Fares, *Historical Land Use/Land Cover Classification Using Remote Sensing*. Springer Cham Heidelberg New York Dordrecht London, 2013.

[79] J. Pérez, J. Díaz, J. Garcia-Martin, y B. Tabuenca, «Systematic literature reviews in software engineering—enhancement of the study selection process using Cohen's Kappa statistic», *J. Syst. Softw.*, vol. 168, p. 110657, oct. 2020.

[80] J. Cerda L, L. Villarroel del P, y Others, «Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa», *Rev. Chil. Pediatr.*, pp. 54-58, 2008.

[81] J. R. Landis y G. G. Koch, «The measurement of observer agreement for categorical data», *Biometrics*, vol. 33, n.º 1, pp. 159-174, mar. 1977.