ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE INGENIERIA EN ESTADÍSTICA INFORMÁTICA

"ANÁLISIS DE ARBOLES DE DECISIÓN PARA LA VALORACIÓN DE CARBONO EDAFICO DE LA PROVINCIA DE CHIMBORAZO MEDIANTE EL USO DE VARIABLES DE EVALUACIÓN NACIONAL FORESTAL MAE - FAO"

Trabajo de titulación

Tipo: Proyecto de investigación

Presentado para optar el grado académico de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

AUTOR: OSCAR ROBERTO PADILLA SEFLA

DIRECTORA: Dra. SILVIA MARIANA HARO RIVERA

Riobamba – Ecuador

2020

Yo, OSCAR ROBERTO PADILLA SEFLA, declaro que el presente trabajo de titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autor asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación. El patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 21 de agosto de 2020

Oscar Roberto Padilla Sefla
060471015-2

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE ESTADÍSTICA INFORMÁTICA

El Tribunal del trabajo de titulación certifica que: El trabajo de titulación: Tipo Proyecto de Investigación, **ANÁLISIS DE ARBOLES DE DECISIÓN PARA LA VALORACIÓN DE CARBONO EDAFICO DE LA PROVINCIA DE CHIMBORAZO MEDIANTE EL USO DE VARIABLES DE EVALUACIÓN NACIONAL FORESTAL MAE - FAO,** realizado por el señor: **OSCAR ROBERTO PADILLA SEFLA,** ha sido minuciosamente revisado por los Miembros del Tribunal del trabajo de titulación, El mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal Autoriza su presentación.

	FIRMA	**FECHA**
Dra. Magdy Mileni Echeverría Guadalupe **PRESIDENTE DEL TRIBUNAL**	Firmado electrónicamente por: MAGDY MILENI ECHEVERRIA GUADALUPE	2020-08-21
Dra. Silvia Mariana Haro Rivera **DIRECTORA DEL TRABO DE TITULACION**	SILVIA MARIANA HARO RIVERA Firmado digitalmente porSILVIA MARIANA HARO RIVERA DN: cn=SILVIA MARIANA HARO RIVERA c=EC o=SECURITY DATA S.A. 1 ou=ENTIDAD DE CERTIFICACION DE INFORMACION Motivo:Soy el autor de este documento Ubicación: Fecha:2020-09-17 21:25-05:00	2020-08-21
Ing. Pablo Javier Flores Muñoz **MIEMBRO DE TRIBUNAL**	PABLO JAVIER FLORES MUNOZ Firmado digitalmente por PABLO JAVIER FLORES MUNOZ Fecha: 2020.09.17 18:17:16 -05'00'	2020-08-21

DEDICATORIA

A Dios por bendecirme con salud, por permitirme tomar mejores decisiones ante las adversidades y por finalmente cumplir mi meta.

Con amor a mis Padres; Juan y Olga, quienes a lo largo de mi vida han velado por mi bienestar y educación siendo mi apoyo en todo momento. Depositando su entera confianza en cada reto que se me presentaba sin dudar ni un solo momento de mi inteligencia y capacidad. Y por el orgullo que sienten por mí, fue lo que hizo ir hasta el final.

Con especial amor a mis hermanos, Edison, Elena, Estefanía y Joselyn que han sido mi principal motivo de inspiración y de superación, por contagiarme de su alegría de vivir y recordarme el valor de una sonrisa y el de un llanto, además de estar apoyándome y compartiendo los momentos más hermosos de mi vida.

Con amor a Marisol por ser pilar central en mi vida, por acompañarme siempre y por compartir conmigo la etapa más hermosa de mi vida.

A mi familia por estar en los momentos que más necesite de su apoyo, gracias por la confianza y apoyo incondicional para seguir adelante y cumplir otra etapa en mí vida, siempre ocuparan una parte muy especial en mi corazón.

A todos aquellos que confiaron en mí y me brindaron una palabra de aliento en los momentos difíciles.

Oscar

AGRADECIMIENTO

El agradecimiento infinito a Dios, por bendecirme cada día, por iluminar mi mente de conocimiento y perseverancia para cumplir cada meta y por darme la fortaleza que se necesita para no rendirme en el camino.

A mis padres por ser los principales gestores de mi formación, por inculcarme el ejemplo de ser una persona de bien ya que con sus consejos, enseñanzas y sustento han logrado guiarme por el duro camino de la vida.

A la Escuela Superior Politécnica de Chimborazo y particularmente a la Facultad de Ciencias, Escuela de Física y Matemática, por brindarme la oportunidad de formarme profesional e íntegramente.

A los Ingenieros, Carlos Bonilla, Diego Damián, y Franklin Cargua del Grupo de Investigación para el Desarrollo Ambiente y Cambio Climático (GIDAC) de la ESPOCH por su apoyo, guía, asesoría y predisposición, donde además brindaron su confianza y dirección en el desarrollo de esta investigación.

Al Ing. Miguel Chinchero por otorgarme un espacio para acceder a los datos del Ministerio del Ambiente del Ecuador (MAE).

Al Ministerio del Ambiente del Ecuador (MAE) como entidad auspiciante de la base de datos de Evaluación Nacional Forestal MAE – FAO a nivel del callejón Interandino, bajo la cual fue posible realizar el presente trabajo.

De manera especial a la Dra. Silvia Haro y al Ing. Pablo Flores, al ser grandes apoyos y guías, que por medio de sus saberes, consejos y experiencias en el campo de la investigación lograron generar un punto de apoyo y fuente de motivación que impulsa, a su vez orienta a continuar el camino de la ciencia mediante una preparación profesional; además agradecer de sobremanera su paciencia, voluntad y disposición que a pesar de tener múltiples ocupaciones contribuyeron para la culminación de este proyecto de investigación.

A cada uno de mis profesores, por transmitirme su conocimiento que permitió mi formación profesional.

Oscar

ABREVIATURAS

C	Carbono
CO	Carbono Orgánico
MO	Materia Orgánica
COS	Carbono Orgánico del Suelo
MOS	Materia Orgánica del Suelo
Mg	Mega gramo
Pg	Peta gramo
Ha	Hectárea
IA	Inteligencia Artificial
SIG/GIS	Sistema de Información Geográfica
CART	Classification and Regression Tree
DA	Densidad Aparente
DEM	Modelos de Elevación Digital
NDVI	Índice diferencial de vegetación normalizado
SAVI	Índice de vegetación ajustado al suelo
VARI	Índice de resistencia atmosféricamente visible
NDWI	Índice diferencial de agua normalizada
BI	Índice de área calcinada
NBR2	Índice de relación de calcinación normalizado 2
EVI2	Índice de vegetación mejorada

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ANEXOS

RESUMEN

El presente trabajo de investigación tuvo como objetivo evaluar la técnica de árboles de decisión mediante el mejor algoritmo de clasificación, para la valoración de carbono edáfico en la provincia de Chimborazo; considerando la base de datos de evaluación Nacional Forestal MAE – FAO. Para el estudio se realizó la limpieza del conjunto de datos, luego se determinaron las variables útiles para la categorización del carbono orgánico del suelo (COS), obteniendo 4 clases: Muy Alto, Alto, Medio y Bajo, posterior a ello se generaron variables espectrales derivadas de imágenes satelitales Landsat 8 (sensor OLI y TIRS), utilizando Sistema de Información Geográfica (SIG). Se encontraron doce variables que controlan la dinámica de distribución de COS, éstas fueron: Ecosistema, Taxonomía, Textura, Pendiente, Modelos de Elevación Digital (DEM), Índice de vegetación de diferencia normalizado (NDVI), Índice de vegetación ajustado al suelo (SAVI), Índice de resistencia atmosféricamente visible (VARI), Índice diferencial de agua normalizada (NDWI), Índice de área calcinada (BI), Índice normalizado de áreas quemadas 2 (NBR2), Índice de vegetación mejorado de dos bandas (EVI2). El algoritmo que proporcionó un mejor porcentaje de eficiencia y resultados relevantes fue el algoritmo de clasificación y regresión (CART) utilizando el método de validación cruzada, el modelo generó una precisión del 65.72% y un error de predicción de 34.28%; estos resultados se presentan como una nueva alternativa de cuantificación de COS. El modelo calibrado puede ser extendido sin necesidad de muestrear *in situ*, muy útil en zonas complejas como el ecosistema de bosque. El mapeo digital de COS permitió revelar los niveles de COS existentes en suelos de Chimborazo y del callejón interandino. Se recomienda continuar con investigaciones bajo esta línea, mismas que permitirán identificar el potencial de la técnica de árboles de decisión, para que puedan ser aplicados es situaciones de interés nacional.

Palabras Clave: <ESTADÍSTICA>, <ÁRBOLES DE DECISIÓN>, <ALGORITMOS DE CLASIFICACIÓN SUPERVISADA>, <ALGORITMO DE CLASIFICACIÓN Y REGRESIÓN>, <CARBONO EDÁFICO>, <SISTEMA DE INFORMACIÓN GEOGRÁFICA>

SUMMARY

The present research work aimed to evaluate the decision tree technique by means of the best classification algorithm, for the evaluation of edaphic carbon in the province of Chimborazo; considering the MAE - FAO National Forest Assessment database. For the study the data set was cleaned, then the useful variables for the categorization of soil organic carbon (SOC) were determined, obtaining 4 classes: Very High, High, Medium and Low, after which variables were generated spectrals derived from Landsat 8 satellite images (OLI and TIRS sensor), using Geographic Information System (GIS). Twelve variables were found that control the SOC distribution dynamics, these were: Ecosystem, Taxonomy, Texture, Slope, Digital Elevation Models (DEM), Normalized Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), Atmospheric Visible Resistance Index (VARI), Normalized Water Differential Index (NDWI), Calcined Area Index (BI), Normalized Burned Area Index 2 (NBR2), Two-band Enhanced Vegetation Index (EVI2). The algorithm that provided a better percentage of efficiency and relevant results was the classification and regression algorithm (CART) using the cross-validation method, the model generated a precision of 65.72% and a prediction error of 34.28%; These results are presented as a new alternative for the quantification of SOC. The calibrated model can be extended without the need for in situ sampling, very useful in complex areas such as the forest ecosystem. The digital mapping of SOC allowed to reveal the existing SOC levels in soils of Chimborazo and the inter-Andean alley. It is recommended to continue with research along this line, which will allow to identify the potential of the decision tree technique, so that they can be applied in situations of national interest.

Keywords: <STATISTICS>, <DECISION TREES>, <SUPERVISED CLASSIFICATION ALGORITHMS>, <CLASSIFICATION AND REGRESSION ALGORITHM>, <SOIL CARBON>, <GEOGRAPHICAL INFORMATION SYSTEM>

INTRODUCCIÓN

En muchas áreas como la ingeniería, medicina, biología, entre otras; aparecen problemas de valoración y de clasificación, siendo una de las técnicas más comunes los árboles de decisión mismos que son efectivos y de fácil interpretación (Roche, 2009, p. 7).

El presente trabajo de titulación consta de varios capítulos, entre ellos el capítulo marco teórico referencial que abarca la temática de los algoritmos de clasificación, su uso en la minería de datos, aprendizaje automático, inteligencia artificial y las herramientas empleadas para su aplicación. El capítulo marco metodológico abarca los materiales, métodos, técnicas y metodologías empleadas para la realización del trabajo. Otro capítulo es el marco de resultados y su discusión, que tiene la siguiente estructura: Depuración y Preprocesamiento de los datos, aplicación y evaluación de los algoritmos de clasificación e implementación de los resultados.

La Depuración del conjunto de datos consistió en limpiar la información irrelevante del conjunto de datos. El preprocesamiento de los datos se realizó con el fin de comprenderlos y preparar para usarlos como entrada en los modelos. Para el modelamiento, se buscaron, seleccionaron y aplicaron los algoritmos de clasificación, más utilizados y óptimos para ser empleados en el conjunto de datos, con los que se estableció el modelo a partir de una data de entrenamiento, posterior a ello se realizó la evaluación de los modelos efectuando los datos de prueba en los modelos generados de cada algoritmo, mediante el cual se seleccionó el algoritmo óptimo para emplear en el conjunto de datos, considerando su mayor eficiencia y menor error de predicción. Y por último los resultados son exportados en mapas geo referenciales generando una mejor interpretación.

También están las secciones de Conclusiones y Recomendaciones, en la sección de Conclusiones se expone las deducciones de las experiencias obtenidas durante el cumplimiento de los objetivos y en la sección de Recomendaciones se detallan sugerencias para un mejor desarrollo de temas similares al presente trabajo.

Antecedentes

Después de una amplia revisión bibliográfica, se ha podido evidenciar que no existe estudios relacionados al tema; lo cual hace que la investigación sea de interés pues permitirá describir variables predominantes para identificar los contenidos de COS en la provincia de Chimborazo y en la región Interandina; sin embargo, existen múltiples aplicaciones de los árboles de decisión de distintas áreas como: Ingeniería, Medicina, Bilogía, Jurídica, Administración de operaciones, Minería de Datos, Para los proyectos de inversión, Tecnología e Inteligencia Artificial.

En Apizaco, Tlaxcala, México se realizó un análisis de árboles de decisión basado en arquitectura ID3 para la detección de apendicitis aguda en niños entre 4 y 15 años, en el que se recolectó un total de 41 casos en el Hospital General Regional. Los resultados mostraron un buen desempeño en el diagnóstico de apendicitis, sin embargo es necesario recolectar una mayor cantidad de datos para maximizar el grado de precisión y factibilidad (Sánchez et al., p. 38).

En Colombia 2017, se construyó un modelo de clasificación basado en árboles de decisión que permitió descubrir patrones de muertes por causa externa. Detectándose que la mayoría de homicidios se presentaron en la Comuna 5 de Pasto, los fines de semana, en la madrugada, el segundo semestre del año, en la vía pública y las víctimas fueron hombres adultos, de oficios varios, siendo la principal causa las riñas y se produjeron con arma de fuego (Timarán et al., 2017, p. 388).

En 2018, en el mismo país, se construyeron árboles de decisión como herramienta para el análisis de riesgos de proyectos, lo cual mostró esa herramienta como una buena opción para la evaluación de situaciones en los proyectos de riesgos, además que es utilidad para resolver problemas que se presentan en el día a día como evaluar una decisión o probabilidad de ocurrencia de una situación a mediano y largo plazo (Maya, 2018, p. 9).

En Brasil 2016, se analizó la exactitud de la clasificación del exceso de peso de escolares mediante la aplicación de un árbol de decisión difusa, se utilizó una base de datos de Itaupú, Paraná (Brasil) conformado por 5962 estudiantes (3024 del sexo femenino y 2938 del sexo masculino), con un rango de edad entre los 6 a 17 años, en el cual determinó una exactitud del 84% en sexo masculino y 89% en sexo femenino (Sulla, et al., 2018, p. 128).

En Santiago de Chile, julio 2014, analizó la estimación de la tasa libre de riesgo para el mercado chileno utilizando herramientas de Minería de datos siendo uno de ellos árboles de decisión;

cuales tienen la ventaja de manejar una gran cantidad de variables junto a sus relaciones no lineales. El rendimiento de los árboles resultó ser superior al modelo de redes neuronales y modelo económico básico (Dupouy, 2014, p. 8).

En Huancayo Perú 2014, analizó la técnica de árboles de decisión como modelo de predicción basado en el rendimiento académico de estudiantes, durante los primeros ciclos en la carrera de ingeniería civil de la Universidad Continental, siendo las variables académicas las más influyentes en el estudiante, que permite una predicción de mayor exactitud y definió el rendimiento académico según intervalos de notas (Camborda, 2014, p. 5).

En la estación Alao, Chimborazo, Ecuador 2016, se aplicó la técnica supervisada árbol de clasificación en data mining mediante el algoritmo CART utilizando ocho variables meteorológicas: temperatura del aire, humedad relativa, presión barométrica, radiación solar difusa, radiación solar global, temperatura del suelo a -20cm y velocidad de viento, con la información comprendida de 24 horas durante todo el año, en el cual el modelo tuvo un rendimiento de 61% donde además la principal variable que mostro el árbol fue la radiación solar global en las horas de 06h00 a 08h00 (Haro Rivera, 2020, pp. 40-42).

Planteamiento del problema

El carbono edáfico es un componente del suelo de vital importancia, su perdida afecta negativamente en la salud del suelo; así como en la producción de alimentos y agrava el cambio climático; esto puede deberse a la falta de conocimiento que se tiene sobre la afectación de las variables: Taxonomía, Textura, Pendiente, Ecosistema, DEM, y los Índices Espectrales como NDVI, SAVI, VARI, NDWI, BI, NBR2, EVI2.

A nivel de la provincia de Chimborazo no existen estudios que realicen una clasificación para los contenidos de carbono edáfico, es por eso que surge la necesidad de identificar los niveles de carbono edáfico en los distintos tipos de suelos, por tal razón el estudio presentado se titula: "Análisis de árboles de decisión para la valoración de carbono edáfico de la provincia de Chimborazo mediante el uso de variables de Evaluación Nacional Forestal MAE-FAO", ya que la identificación y clasificación de los diferentes niveles de carbono evitaran agrupaciones erróneas de las muestras obtenidas en los diferentes tipos de suelo. Este estudio pretende caracterizar el carbono edáfico mediante árboles de decisión, seleccionando el mejor algoritmo de acuerdo a su precisión; para posteriormente mostrar mediante mapas las zonas con mayores contenidos de COS.

Formulación del problema

¿La construcción y análisis de árboles de decisión, permitirá clasificar adecuadamente el contenido de carbono edáfico en la provincia de Chimborazo, en los distintos tipos de suelo?

Problema General

¿Cuál es el mejor algoritmo que permite clasificar el contenido de carbono edáfico en los diferentes tipos de suelo?

Problemas Específicos

- ¿Cuál es la exactitud de los algoritmos empleados en los árboles de decisión que permita clasificar el contenido de carbono edáfico en los diferentes tipos de suelo?
- ¿Cuál es el error mediante la técnica de árboles de decisión para predecir el contenido de carbono edáfico de acuerdo con el tipo de suelo?

Justificación

Justificación teórica

El presente trabajo es de vital importancia para el cumplimiento de los objetivos además ésta investigación tiene como propósito realizar una clasificación de carbono edáfico en la provincia de Chimborazo, para lo cual se pretende realizar una identificación de los diferentes niveles; mediante la utilización de árboles de decisión, el cual se basa en métodos de aprendizaje inductivo, permitiendo comprender el fenómeno del estudio desde el punto de vista de la causalidad, utilizando grandes bases de datos y establecido relaciones y jerarquías entre las variables involucradas. La metodología integra técnicas de procesamiento de imágenes multiespectrales mediante SIG y el uso de algoritmos de autoaprendizaje, para obtener un mapeo digital del COS a nivel de Bosque en la provincia de Chimborazo. Los resultados obtenidos servirán para extender la predicción de COS mediante el algoritmo calibrado en el ecosistema de estudio a nivel nacional. Como también será referente para nuevas investigaciones a nivel nacional e internacional. Esta investigación se la realiza dentro del proyecto "Soil Organic Carbon Evaluation and Sequestration in Ecuadorian Páramo Ecosystems" perteneciente al Grupo de Investigación y Desarrollo para el Ambiente y el Cambio Climático (GIDAC - ESPOCH)

Justificación práctica

Mediante la aplicación de técnicas estadísticas se podrá dar a conocer los tipos de suelos más representativos en el área de estudio, y a la vez cuantificar y clasificar los contenidos de COS obtenidos en diferentes tipos de suelo, para de esta manera contribuir a la generación de nuevas políticas, incrementando y mejorando las distintas acciones enfocadas al manejo adecuado, planificado y sostenible del contenido del COS; además esta investigación pretende proyectar mediante mapas geo referenciales la predicción de los contenidos del COS, con el fin de obtener una visualización legible, misma que permita interpretar los resultados y ayuden a mejorar la toma de decisiones con respecto la valoración del contenido del COS y conservación efectiva a mediano y largo plazo.

Para la investigación se empleará R Studio, software estadístico libre que permite autonomía, tecnológica, estandarización e integración, seguridad, democratización de la información y ahorro de recurso; logrando así la innovación nacional y la optimización del gasto estatal para fortalecer el desarrollo local y la inclusión digital. Adicionalmente la base de datos a ser manipulada proviene de una fuente secundaria gratuita proporcionada por el Ministerio del Ambiente del Ecuador (MAE) y por el Ministerio de Agricultura y Ganadería (MAG), las mismas que son instituciones públicas que se encuentra al servicio de la población.

Objetivos

Objetivo General

Evaluar la técnica de árboles de decisión mediante el mejor algoritmo que se ajuste de forma significativa, a las variables de la base de datos de Evaluación Nacional Forestal MAE - FAO, para la valoración del contenido de carbono edáfico en la provincia de Chimborazo.

Objetivos Específicos

- Analizar las variables de la base de datos de Evaluación Nacional Forestal MAE - FAO.
- Identificar la variable clasificadora considerando los niveles de carbono presente en los diferentes tipos de suelo.
- Seleccionar el óptimo algoritmo de clasificación, para la estimación del contenido de carbono edáfico mediante árboles de decisión.
- Visualizar en mapas geo referenciales el contenido de carbono edáfico obtenido con los árboles de decisión.

Planteamiento de la Hipóstasis

Hipótesis General

La clasificación con árboles de decisión mediante el mejor algoritmo permitirá catalogar los niveles de carbono edáfico en las distintas zonas de la provincia de Chimborazo con mayor exactitud.

Hipótesis Específicos

- La técnica de árboles de decisión mediante el mejor algoritmo permite una exactitud mayor al 70%.
- La técnica de árboles de decisión permite un error de predicción inferior al 30%.

CAPITULO I

1. MARCO TEÓRICO REFERENCIAL

1.1 Inteligencia Artificial

La inteligencia artificial surge en la década de 1940, tiene como principal objetivo imitar el comportamiento humano con el fin de obtener el mejor resultado esperado, es un campo de estudio muy amplio y en constante cambio que busca una comprensión profunda de la inteligencia y de la capacidad de esta a través de la comprensión de sus límites y alcance (Leyva Vázquez et al., 2018, p. 28). También permite dar solución a problemas referentes al análisis de la información con el fin de optimizar el aprendizaje y la toma de decisiones (Marlon Gómez, 2014, p. 35).

1.2 Aprendizaje Automático

El aprendizaje automático conocido también como Machine Learning son técnicas de la minería de datos los cuales son una rama de la Inteligencia Artificial (IA) que desarrollan técnicas capaces de extraer de forma automática conocimientos subyacentes en la informática, donde se implementan algoritmos que procesan conjuntos de datos, identifican patrones y extraen conocimientos suficientes genéricos como para aplicarlos a nuevos conjuntos de datos (Coello Blanco et al., 2018, p. 1421).

En la investigación el aprendizaje automático fue utilizado para comprender la clasificación de los algoritmos de clasificación y se dividen en dos grandes bloques: Aprendizaje no Supervisado o Análisis Clúster, y Aprendizaje Supervisado o Reconocimiento de patrones.

1.2.1 Aprendizaje Supervisado

Estos algoritmos tienen como objetivo determinar cuál es la clase a la que pertenece una nueva muestra sin clase, en base a las clases de las que ya se tiene conocimiento, así como patrones de entrada y salida (Villanueva Morales et al., 2015, p. 264). Estos algoritmos trabajan con un conjunto de datos o de entrenamientos que permite construir modelos confiables (Núñez Reyes et al., 2016, p. 135), para clasificar nuevas observaciones o interpretar la información para transformarla a conocimiento (Rodríguez Tapia, 2018, p. 140). Este tipo de aprendizaje es el más preciso debido a que los clasificadores trabajan con datos ya entrenados es decir que se parten de un conjunto de datos

en el que ya se conoce la variable clase o etiqueta asignada de forma correcta (Corrales Gasca et al., 2015, p. 46).

Los algoritmos que pertenecen a este bloque intentan encontrar relaciones entre las variables independientes y las clases del problema. Con las relaciones o patrones que se logren encontrar en el conjunto de entrenamiento, se construye un modelo de clasificación lo suficientemente genérico como para ser capaz de clasificar correctamente nuevos casos en los que ya no se conozca el valor de la variable clase.

Dentro del bloque del aprendizaje supervisado existen dos modelos de clasificación

- **Modelos No Explicativos:** Entre estos modelos se encuentran paradigmas de clasificación como el análisis discriminante, redes neuronales, técnicas basadas en medidas de distancias (K-NN).

- **Modelo Explicativos:** Estos modelos además de realizar la clasificación de un problema aportan una justificación del resultado obtenido, entre estos modelos están las reglas de inducción y los árboles de decisión.

1.2.2 Aprendizaje No Supervisado

Estos algoritmos no requieren un conjunto de datos entrenados, ni de la intervención de humanos para elaborar un conjunto de datos categóricos (Godoy Viera, 2017, p. 105). En este tipo de aprendizaje, no existe un conocimiento previo a partir del cual se pueda "entrenar" a la máquina para resolver nuevos problemas, es decir, se desconoce el valor de la variable dependiente en el conjunto de datos (Rivera Camacho et al., 2015, p. 11). Los algoritmos que pertenecen a este bloque se encargan de buscar agrupaciones de casos en el conjunto de entrenamiento en base al valor de sus variables independientes, en que se indica que los casos del mismo grupo se parezcan y los casos de distinto grupo sean lo más diferente posible (Gómez Victoria, 2014, p. 36).

1.3 Minería de datos

La minería de datos es una solución para el análisis de fenómenos no explícitos en bases de datos y la búsqueda de patrones ocultos entre los datos, para posteriormente ser usado en la predicción de comportamientos, es decir, mediante la aplicación de técnicas de Inteligencia Artificial y de Aprendizaje Automático identifica información relevante que estaría oculta (Romero Romero, 2018, p. 18; Acosta et al., 2018, p. 1080). Esta herramienta también es conocida como Descubrimiento de conocimientos de Bases de datos, debido a que permite analizar grandes bases de datos y obtener

una descripción de las tendencias y correlaciones entre los datos, facilitando así la toma de decisiones (Escobar Terán et al., 2016, p. 506).

Además, haciendo uso de diferentes algoritmos a partir de datos pre procesados resuelve problemas de agrupamiento automático, clasificación, asociación y detención de patrones secuenciales, ya que proporcionan nuevos conocimientos (Romero y Paredes, 2013, p. 15). Las técnicas de la minería de datos pueden ser descriptivas o predictivas, y generalmente se dividen en cinco categorías: métodos estadísticos, análisis de clúster, árboles de decisión y reglas de decisión, reglas de asociación y detección de fraudes (Romero y Paredes, 2013, p. 16).

El proceso de la minería de datos consta de cinco partes: selección de la información, transformación de los datos, aplicación de las técnicas, interpretación de los resultados y la incorporación del nuevo conocimiento (Cortés Martínez et al., 2018, p. 17).

1.3.1 Metodología para el análisis de minería de datos

Para el análisis de minería de datos existen tres metodologías principales:

1.3.1.1 KDD

La metodología KDD (*KnowledgeDiscovery in Databases*) es un proceso enfocado a descubrir patrones útiles y comprensibles a partir de un conjunto de datos, está compuesto por cinco fases (Azevedo y Santos, 2008, p. 182):

- **Selección:** Consiste en crear un conjunto de datos objetivo o centrarse en un subconjunto de variables o muestras de datos, en las cuales se realizará el descubrimiento.
- **Preprocesamiento:** Radica en la limpieza de datos de destino y el preprocesamiento para obtener datos consistentes.
- **Transformación:** Reside en la transformación de los datos utilizando la dimensionalidad o métodos de reducción.
- **Minería de datos:** Se basa en la búsqueda de patrones de interés en una determinada forma de representación, según el objetivo de minería de datos.
- **Evaluación e Interpretación:** Está en la interpretación y evaluación de los datos

1.3.1.2 SEMMA

El proceso (*Sample Explore Modify Model and Assess*) fue desarrollado por el Instituo SAS. SEMMA permite un desarrollo organizado y adecuado, y considera un ciclo de cinco etapas para el proceso (Azevedo y Santos, 2008, p. 183):

- **Muestra:** Consiste en muestrear los datos extrayendo una porción de un conjunto de datos grande suficiente para contener la información importante, pero lo suficiente pequeño como para manipularlo rápidamente.

- **Explorar:** Radica en la exploración de los datos mediante la búsqueda de tendencias imprevistas y anomalías para obtener comprensión e ideas.

- **Modificar:** Se refiere a la modificación de los datos mediante la creación, selección y transformación de las variables para enfocar el proceso de selección del modelo.

- **Modelar:** Se basa en modelar los datos permitiendo que el software busque automáticamente para una combinación de datos que predice de manera confiable el resultado deseado.

- **Evaluar:** Está en avaluar los datos mediante la utilidad y confiabilidad de los resultados y estimar que tan bien funciona.

1.3.1.3 CRISP-DM

El proceso CRISP-DM (*Cross Industry Standard Process for Data Mining*) se desarrolló mediante el esfuerzo de un consorcio inicialmente compuesto con DaimlerChryler, SPSS y NCR. Es una de las metodologías más usadas para analizar grandes conjuntos de datos y descubrir información valiosa, consiste en un ciclo que comprende seis etapas (Azevedo y Santos, 2008, p. 184).

- **Definir y analizar el problema:** Es esta etapa se identifica los objetivos del estudio entendiendo el problema y definiendo las variables relevantes.

- **Exploración de la data**: Obtener, describir y unificar los datos en un mismo formato puede ser una de las tareas más costosas del proceso.

- **Preparación de la Data:** En esta parte se estudia la data disponible, se analizan las variables desde un enfoque teórico, estadístico y descriptivo con el objetivo de comprender los datos, finalmente se le transforma y limpia para usarlos como variables de entrada en los modelos. Es necesario verificar que la muestra tenga una cantidad equilibrada de clases, es decir, que no exista una proporción muy pequeña de datos, esto es importante porque las particiones

son sensibles a la cantidad de observaciones que hay en cada una de las clases de la muestra (Giudici & Figini, 2009, p. 17).

- **Modelamiento:** Esta etapa se definen, diseñan, seleccionan y aplican distintas herramientas o modelos sobre la data, además se seleccionan los datos de entrenamiento y prueba, también elige el modelo más adecuado en función del problema y los parámetros que mejor se ajusten a los datos sin caer en el sobreajuste y finalmente se optimizan los modelos tratando de identificar los elementos relevantes que más impacten en su rendimiento.

- **Evaluación y Prueba del Modelo:** Aquí se somete a evaluarlos los modelos estudiados con los datos de prueba para valorar como ajusta a nuevos datos y así, medir la precisión de la predicción.

- **Implementación**: En esta etapa los resultados son utilizados y exportados a reportes o a otras bases de datos, además mejoramos el modelo y establecemos un seguimiento sobre posibles variaciones.

Partiendo de lo explicado en las tres metodologías se especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo que es deseable obtener tras cada fase, sin embargo la metodología más completa es CRISP-DM debido a que tiene en cuenta la aplicación al entorno de negocio de los resultados y es muy útil para comprender los métodos de trabajo ya que cuenta con seis etapas muy importantes (Azevedo y Santos, 2008, p. 185).

1.4 Árbol de decisión

1.4.1 Historia

El uso de los árboles de decisión se remonta en 1944 por una idea de *Von Neumann* (1903 - 1958) utilizada en la teoría de juegos, tal como nos cuentan en su artículo Carlota Gastaldi, Marcel Urrea y Pedro Fernández de Córdova, Neumann recomienda trazar un diagrama en forma de árbol donde se pueden ver todas las diferentes maneras de jugar y descubrir un mejor modo de hacerlo, en donde se obtenga un resultado satisfactorio (Gastaldi et al., 1998, pp. 37-42).

Después, en 1950 como nos cuenta Matsudo en su artículo: "Árboles de decisión, una técnica de data meaning", gracias a un trabajo realizado por Hoveland y Hunt se plantea que esta herramienta sea usada para un conjunto de entrenamiento T en el ámbito de la informática, allí se dio a conocer que los árboles de decisión son de gran y múltiple utilidad (Matsudo, 1991, p. 5). Aunque generalmente en esa época el enfoque de éstos era en temas estadísticos, desde 1964 se empezó a

usar en la solución de problemas de inversión y poco después en el análisis de riesgos y proyectos (Vélez Pareja, 2003, p. 5).

Magee fue el primero en utilizar la técnica de árboles de decisiones para tratar el problema de las disposiciones de inversión de capital y posteriormente Hespos y Strassmann propusieron, con algún detalle, combinar el análisis del riesgo, propuesto por Hertz y Hillier, con la técnica de los árboles de decisión (debe aclararse que Magee había previsto la combinación de estos enfoques cuando planteo la utilización de los árboles de decisión). Por su parte, en 1968, Raiffa desarrolló en forma detallada y muy clara la teoría de la decisión, donde se incluye la técnica propuesta por Magee y en general todo lo relacionado con las decisiones bajo riesgo (Vélez Pareja, 2003, p. 323).

1.4.2 Concepto

Los árboles de decisión son una técnica de aprendizaje inductivo supervisado no paramétrico, se utiliza para la predicción y se emplea en el campo de inteligencia artificial, donde a partir de una base de datos se construyen diagramas en forma de árboles, que sirven para representar y categorizar una serie de condiciones que ocurren en forma repetitiva para la solución de un problema (Quintero y Amézquita Collazos, 2003, p. 9).

Los árboles de decisión son guías jerárquicas multi-vía, porque pueden existir más de dos opciones y es una guía porque al responder una pregunta llega a una decisión, son una de las técnicas más eficientes de la clasificación supervisada.

Un árbol, es un modelo esquemático de las alternativas disponibles y de las posibles consecuencias de cada una, su nombre proviene de la forma que adopta el modelo parecido a la de un árbol. El modelo ésta conformado por múltiples nodos cuadrados que representa puntos de decisión y de los cuales surgen ramas (que debe leerse de izquierda a derecha), que representan las distintas alternativas, las ramas que salen de los nodos circulares representan los eventos. La probabilidad de cada evento, P(E), se indica encima de cada rama, las posibilidades de todas las ramas deben sumar 1 (Krajewski y Ritzman, 2000, p. 76).

1.4.3 Objetivo

Los árboles de decisión tienen como objetivo crear un modelo que predice el valor de una variable destino en función de diversas variables de entrada.

1.4.4 Tipos de árboles de decisión

Según (Hernández Y., 2015) existen diferentes tipos de árboles de decisión, que son catalogados según la situación y el resultado deseado:

- **Árbol de clasificación o binario:** Se usa cuando hay varias alternativas que se han calculado anteriormente para obtener resultados más predecibles, para hacer uso de éstos, hay que trazar esquemas binarios y proyectar las diferentes variables o ramas del árbol, gracias a las probabilidades de éstas se puede predecir un poco el resultado. Es usado en probabilidad, estadística y minería de datos.

- **Árbol de regresión:** Este tipo de árbol ayuda a determinar un único resultado ya que se tiene la información necesaria para identificar una "ruta" óptima. Cuando se construye este árbol se divide la información en secciones o subgrupos. Este tipo de árbol es muy usado en bienes raíces.

- **Árbol de mejora:** Es usado cuando se desea tener un resultado más preciso; el árbol se construye, luego se toma una variable, esta se calcula y se la estructura para reducir la incertidumbre y los errores a la hora de tomar decisiones. Este árbol es usado en contabilidad y matemática.

- **Árboles mixtos entre regresión y clasificación:** Se utiliza para prever un resultado con variables impredecibles, generalmente se usan indicadores que muestren lo que ya ha sucedido en un pasado (Sesgando un poco el resultado), y es aplicado generalmente en la ciencia.

- **Árboles de binomiales y trinomiales:** Son árboles donde cada rama es un modelo binomial su función es recrear el modelo binomial varias veces en el tiempo, suponiendo que el precio o costo de un elemento sube o baja con el tiempo. Estos árboles son muy usados en la aplicación de operaciones reales.

1.4.5 Aplicaciones

Los árboles de decisión son muy usados en diferentes ámbitos de la vida en los que requiere tomar decisiones de cualquier tipo, como dice el *Intaver Institute* en uno de sus artículos, cualquier persona puede usar rutinariamente los árboles de decisión, el análisis y lo útil de la herramienta depende de la consecuencia positiva o negativa de cada una de las alternativas:

- **Jurídica:** Esta herramienta puede ser usada en la toma de decisiones de un abogado sobre demandar o no demandar, en ir o no a juicio, teniendo en cuenta la probabilidad de ganar el juicio y lo que implica monetariamente (Intaver Institute Inc.).

- **Administración de operaciones (OM)[2]:** En esta parte se utilizan en la planificación de productos y administración de procesos o capacidad, en el ámbito de OM este método es valioso, ya que permite evaluar diferentes alternativas de expansión cuando hay problemas de capacidad o cuando la demanda es incierta y también cuando hay involucrada más de una decisión (Krajewski y Ritzman, 2000, p. 4).

- **Minería de Datos (DM)[3]:** En este caso son usados abordando problemas como predicción, clasificación y segmentación de datos con el fin de convertirlos información valiosa para el análisis y toma de decisiones. Lograr una buena minería de datos depende de algoritmos, algunos más potentes y otros menos sofisticados que se aplican a los árboles de decisión (Written et al., 2016, p. 34).

- **Medicina:** En el ámbito de la medicina ha sido usada por más de dos décadas y es de vital importancia en casos de análisis genéticos o a la hora de tomar una decisión de si operar o no un paciente que llega a urgencias o incluso la posibilidad de que una persona desarrolle un efecto adverso a una medicina según su predisposición genética, estos árboles de decisión son de tipo clasificación (Bouza Herrena et al., 2014, pp. 20-25).

- **Para los proyectos de inversión:** Esta herramienta también es muy usada en proyectos de inversión como lo cuenta en su libro Raúl Coss Bu "El enfoque de los árboles de decisión, una técnica muy similar a programación dinámica es un método conveniente para representar y analizar una serie de inversiones hechas a través del tiempo" (Coss Bu, 2005, p. 253). También nos aclara que esta herramienta por sí sola no nos indica la mejor opción, para esto el usuario debe conocerlas alternativas, variables y probabilidades, es decir, no se le puede escapar ningún detalle, lo que en algunos casos requiere de mucha información y tiempo de construcción.

- **Tecnología e Inteligencia Artificial (IA):** Es usada para la parametrización y desarrollo de aplicaciones con el fin de indicarle a un programa cómo "Comportarse" frente a una situación que se presente. El manejo de datos por medio de IA se hace además a través de algoritmos, basados en reglas que sirven para representar una serie de condiciones que ocurren de forma sucesiva, como se esperaría que lo hiciera un humano.

- **Valoración de opciones reales:** Este tipo de árboles es usado en el análisis financiero de proyectos con una perspectiva estratégica, evaluando alternativas futuras como: abandono

del proyecto, expansión en caso en el que los resultados superen lo esperado, aplazamiento de la inversión o suspensión del proyecto para evitar flujos de caja negativos, dependiendo del comportamiento de una variable en el tiempo. Este método, tiene como ventaja sobre otros, que permite calcular el comportamiento financiero mínimo hasta la finalización de proyecto y no solo por un período limitado.

Los árboles de decisión "por su estructura son fáciles de comprender y analizar; su función cotidiana se puede dar en diagnósticos médicos, predicciones meteorológicas, controles de calidad y otros problemas que necesiten de análisis de datos y toma de decisiones", los árboles de decisión pueden ser usados en cualquier ámbito sin importar que sea laboral o personal, siempre y cuando implique toma de decisiones con cierto agrado de incertidumbre (Calancha Zuniga et al., 2010, p. 2).

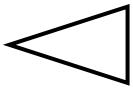
1.4.6 Terminología

- **Nodos de decisión:** Representa los puntos de decisión donde se muestran las distintas alternativas disponibles a elegir. Se escoge la alternativa que presenta mayor valor esperado.
- **Nodos de probabilidad:** Están representado con un círculo, muestran las probabilidades de ciertos resultados.
- **Nodo terminal:** Muestra el resultado definitivo de una ruta de decisión
- **Ramificaciones alternativas:** Cada ramificación representa un resultado probable
- **Alternativa rechazada:** Una vez desarrollado el árbol, las alternativas que no se seleccionan se marcan con dos líneas.

1.4.7 Notación

Los árboles de decisión constan de: nodos de decisión representados por un cuadrado, nodos de probabilidad representados por un círculo y ramas o alternativas representados por líneas o una línea cruzada por otras dos, para notar que es una decisión rechazada y se puede apreciar en la tabla siguiente.

Tabla 1-1: Notación de un árbol decisión.

SIMBOLO	NOMBRE	DESCRIPCION
	Nodo de decisión	Indica una decisión que se tomará
	Nodo de probabilidad	Muestra múltiples resultados inciertos
	Ramificaciones alternativas	Cada ramificación indica un posible resultado o acción
	Alternativa rechazada	Muestra una alternativa que no estaba seleccionada
	Nodo terminal	Indica un resultado definitivo

Fuente: Tópicos de la Inteligencia Artificial.
Realizado por: Padilla S., Oscar R., 2020.

En la Figura 1-1, se muestra las características de un árbol de decisión

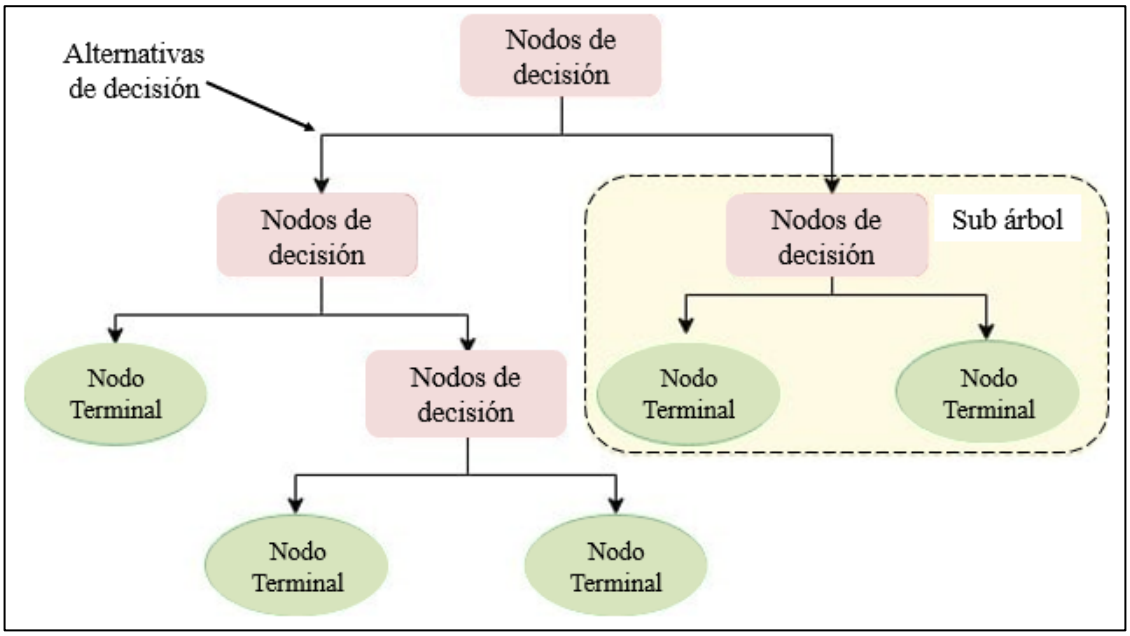


Figura 1-1: Características de un árbol de decisión.
Fuente: Tópicos de la Inteligencia Artificial.

1.4.8 Ventajas y desventajas

Según (Pérez López, 2011, p. 24) se presenta la siguiente tabla:

Tabla 2-1: Ventajas y desventajas de un árbol de decisión.

VENTAJAS	DESVENTAJAS
Son fáciles de construir, interpretar y visualizar.	Tienden al sobreajuste de los datos, por lo que el modelo al predecir nuevos casos no estima con el mismo índice de acierto.
Selecciona las variables más importantes y en creación no siempre se hace uso de todos los predictores.	Se ven influenciados por los outliers, creando árboles con ramas muy profundas que no predicen bien para nuevos casos.
Si faltan datos no podremos recorrer el árbol hasta un nodo terminal, pero si podemos hacer predicciones promediando las hojas del subárbol que alcancemos.	No suele ser muy eficientes con modelos de regresión
No es preciso que se cumplan una serie de supuestos como en la regresión lineal (linealidad, normalidad de los residuos, homogeneidad de la varianza, etc.)	Crear arboles demasiado complejos pueden conllevar a que no se adapten bien a los nuevos datos.
Sirven tanto para variables dependientes cualitativas como cuantitativas, también para variables predictoras o independientes numéricas y categóricas. Además, no necesita variables dummny, aunque a veces mejoran el modelo.	La complejidad resta capacidad de interpretación.
Permiten relaciones no lineales entre las variables explicativas y la variable dependiente	Se pueden crear árboles sesgados si una de las clases es más numerosa que otra.
	Se pierde información cuando se utilizan para categorizar una variable numérica continua.

Fuente: (Pérez López, 2011, p. 24).
Realizado por: Padilla S., Oscar R., 2020.

1.5 Técnicas de validación

El valor de un clasificador reside fundamentalmente en su capacidad de generalización, de modo que sea capaz de identificar en la muestra de entrenamiento patrones que después le sirvan para clasificar correctamente otros casos de clase desconocida A eso le llama *bondad del clasificador*, y es lo que intenta cuantificarse mediante las técnicas de validación.

La técnica más básica podría consistir en ejecutar el clasificador contra la base de datos de entrenamiento y contabilizar el número de errores que comete, aprovechando que es un conjunto de datos para el que conocemos la clase de todos los casos. Sin embargo, esta tasa de error, conocida como *Error de entrenamiento*, no es un criterio de bondad demasiado valioso porque se presupone que será una tasa de error más bien optimista, al estar clasificando los mismos casos con los que se ha entrenado el modelo.

Una tasa de error mucho más interesante de calcular es el *Error poblacional*, que expresa el error cometido al clasificar una población universal de individuos desconocidos para el modelo. A continuación, se expondrán algunas de las técnicas más comunes para calcular este criterio de bondad.

1.5.1 Método H (Holdout)

Este método consiste en dividir la base de datos en dos conjuntos. El primer conjunto, llamado *conjunto de entrenamiento* y que suele comprender 70% de la base de datos, se utilizará para construir el modelo. El segundo conjunto, llamado *conjunto de test* y que comprende del 30% de datos restantes y será aplicado sobre el modelo construido para calcular su bondad.

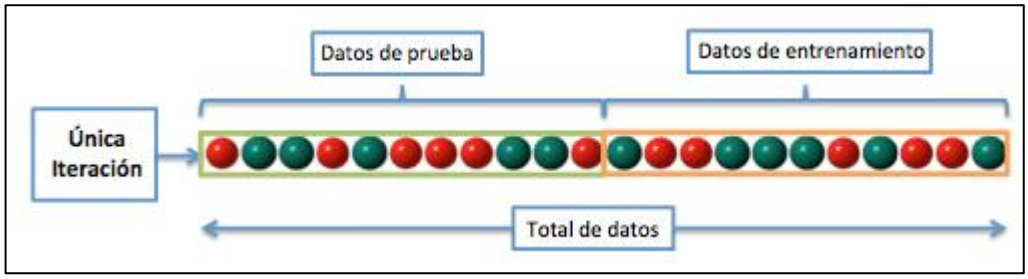


Figura 2-1: Método de Holdout.
Fuente: Validación cruzada (De Joan.domenech91 – Trabajo propio, CC BY-SA 3.0).

Es un método sencillo y no resulta tan robusto como otras técnicas, sin embargo, es mucho más eficiente. Métodos más avanzados como los que explicaran a continuación resultan mucho más caro desde un punto de vista computacional, sobre todo a medida que aumente el número de casos

y/o atributos, hasta el punto de bases de datos realmente grandes el Holdout puede ser el único método de validación viable.

1.5.2 Validaciones cruzadas

Se trata de un tipo de técnicas donde nunca se testea el clasificador con individuos que hayan formado parte de la muestra de entrenamiento. A continuación, se explicarán las dos técnicas más conocidas de este tipo.

1.5.2.1 Leave-one-out

Esta técnica se basa en la idea de las validaciones cruzadas. Partiendo de una base de datos de tamaño "n", ésta se divide en "n" submuestras. Para cada submuestra se separa un único individuo, y con el resto de los individuos se construye un modelo con el que después se intentará clasificar el individuo separado. Al no haber sido parte del conjunto de entrenamiento de la submuestra, si se clasifica mal este individuo se puede considerar un *error poblacional.*

Finalmente, se construirá un clasificador con la base de datos completa y, según esta técnica de validación, su error poblacional será la media de los errores poblacionales cometidos para las "n" submuestras con las que se ha realizado el test.

1.5.2.2 m-fold cross-validation

Esta técnica parte de la misma idea que el *Leave-one-out*, pero en lugar de separar un único individuo de cada submuestra separará un número "n" de individuos. Este grupo de "n" individuos se denominan *fold* y, de la misma manera que en el anterior método, se usarán para medir el error poblacional de la submuestra a la que pertenecen.

De la misma manera, el clasificador final se construye con la base de datos completa y se calcula su error poblacional como la media de los errores poblacionales de las submuestras utilizadas.

En cada una de las k iteraciones de este tipo de validación se realiza un cálculo del error. El resultado final se obtiene realizando la media aritmética de los k valores de errores obtenidos, y se lo representa mediante la siguiente ecuación.

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i \qquad (1.1)$$

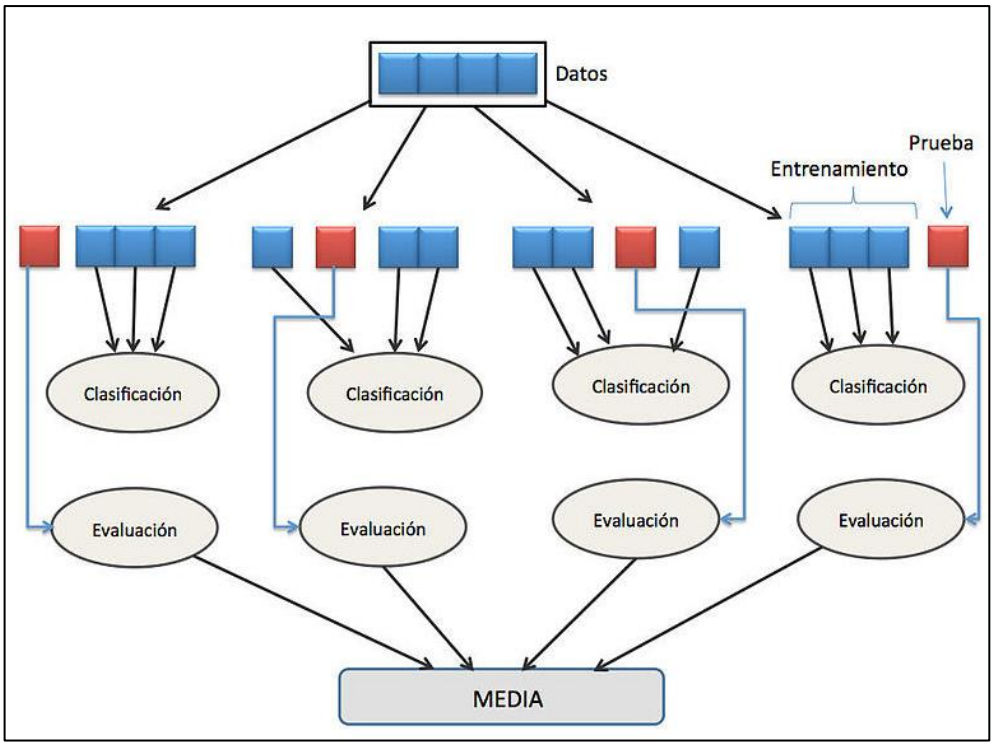


Figura 3-1: Esquema de validación cruzada de k iteraciones.
Fuente: Validación cruzada (De Joan.domenech91 – Trabajo propio, CC BY-SA 3.0).

1.6 Algoritmos de clasificación

Como se mencionó anteriormente los árboles de decisión son un paradigma del Aprendizaje Supervisado que se enmarca dentro de los algoritmos de Machine Learning y que sirve para generar modelos de clasificación explicativos; esto se debe a que los árboles de decisión es una técnica predictiva y los datos disponibles para el desarrollo del modelo corresponden a variables categóricas y discretas, que se ajustan a las características de un árbol de decisión, además por la facilidad de interpretar la información obtenida en forma gráfica.

Los árboles de decisión cuentan con una serie de algoritmos de clasificación que se pueden utilizar para clasificar y predecir, el objetivo de estos algoritmos es que, partiendo de un conjunto de entrenamiento, genera una estructura de árbol de decisión que hará de modelo de clasificación.

1.6.1 Búsqueda de algoritmos de clasificación

El presente trabajo se aplicó diversos algoritmos de clasificación, para su selección se procedió a la búsqueda de información sobre estudios relacionados.

Tabla 3-1: Búsqueda de algoritmos de clasificación.

N	Estudios analizados	Año de publicación
1	Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas (Ortiz Lozano et al., 2017).	2017
2	Aplicación de un árbol de decisión difusa con clasificación de ambigüedad para determinar el exceso de peso en escolares (Sulla Torres et al., 2018).	2018
3	Evaluación de la decisión de obtener el título profesional con la elaboración de la tesis mediante técnicas multivariantes. Caso Universidad Nacional Agraria La Molina (Fernández Jeri y Salinas Flores, 2017).	2017
4	Análisis de la deserción estudiantil en la USB, facultad Ingeniería de Sistemas, con técnicas de Minería de Datos (Romero y Paredes, 2013).	2013
5	Aprendizaje supervisado de funciones de distancia: estado del arte (Nguyen Cong et al., 2015).	2015
6	Aplicación de algoritmos de clasificación para el análisis de tejido mamario y detección de cáncer de mama (Villanueva Morales et al., 2015).	2015
7	Algoritmos de aprendizaje automático: Aplicación en la solución a problemas medio ambientales (Vega Calcines, 2014).	2014
8	Aprendizaje por refuerzo en espacios continuos: algoritmos y aplicación al tratamiento de la anemia renal (Montero, 2014).	2014
9	Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de datos (Eckert y Suénaga, 2015).	2015
10	Aplicación de árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa de interés en Chile (Dupouy Berrios, 2014).	2014
11	Modelo de Sentiment Analysis para la clasificación de noticias en tiempo real en el Mercado de Valores de Buenos Aires (Braña et al., 2016).	2016
12	Aplicación de árboles de decisión para la predicción del rendimiento académico de los estudiantes de los primeros ciclos de la Carrera de Ingeniería Civil de la Universidad Continental (Camborda Zamudio, 2014).	2014

13	Aplicación de Algoritmos de Clasificación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos (Santana Mansilla et al., 2014).	2014
14	Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia (Timarán Pereira et al., 2017).	2017
15	R and Data Mining Examples and Case Studies (Zhao, 2015).	2015
16	Aplicación de Minería de Datos en Marketing (Escobar Terán et al., 2016).	2016
17	Técnicas de aprendizaje de maquina utilizadas para la minería de texto (Godoy Viera, 2017).	2017
18	Algoritmos de clustering y aprendizaje automático aplicados a Twitter (Blanco y Sanz, 2016).	2016

Realizado por: Padilla S., Oscar R., 2020.

Una vez obtenidos los resultados de la Tabla 3-1, se revisaron los algoritmos empleados en cada estudio, se contabilizó el número de veces en las que fue utilizado cada algoritmo, resultado que se presentó en la Tabla 4-1.

Tabla 4-1: Obtención de los algoritmos de clasificación.

Algoritmos	**Estudios analizados**																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**
IDE3			x							x		x							3
AID	x																		1
MARS	x	x					x												3
CHAID	x		x									x						x	4
CART	x			x		x		x		x		x				x			7
QUEST	x														x				2
C4.5						x			x	x		x	x	x					6
C5.0			x				x			x					x	x			5
KNN					x								x				x	x	5
SVM			x								x		x				x	x	5
PAUM													x						1
DHP				x															1
ECLART				x															1

Realizado por: Padilla S., Oscar R., 2020.

En la Tabla 4-1, se observó que los algoritmos de clasificación más utilizados en los estudios son:

- Árboles de Clasificación y Regresión (CART)
- Árboles de Clasificación J48 (C4.5)
- Árboles de Clasificación C5.0
- Máquinas de Soporte Vectorial (SVM)

1.6.2 Selección de los algoritmos de clasificación

Para la selección de los algoritmos de clasificación, se comparó los algoritmos hallados en el punto anterior de acuerdo con características establecidas en base a los resultados y conclusiones de los estudios. Las características planteadas fueron:

- PE = Porcentaje de eficiencia
- GR = Representación gráfica
- EC = Efectivo en conjuntos grandes de datos
- CA = Especificar cantidad de agrupamiento
- DH = Disponibilidad en herramientas elegidas

Los algoritmos que abarcaron más características fueron seleccionados para ser aplicados en este trabajo de titulación, los cuales se detallaron en la Tabla 5-1.

Tabla 5-1: Característica comparativa de los algoritmos de clasificación.

ALGORITMOS	PE	GR	EC	CA	DH
CART	x	x	x	x	x
C4.5	x				
C5.0	x		x	x	x
SVM	x		x		x

Realizado por: Padilla S., Oscar R., 2020.

Como se observó en la Tabla 5-1, los algoritmos que coincidieron con más de dos características fueron: CART, C5.0, SVM; y a continuación se detallan estos algoritmos que fueron aplicados en el conjunto de datos.

1.6.3 Algoritmo C5.0

Se trata de un algoritmo desarrollado por Ross Quinlan en 1993 (Quinlan, 1993, p. 17) como una mejora del C4.5 el cual es una extensión del ID3, desarrollado en 1986 por el mismo autor (Quinlan, 1986, p. 85). Se basan en el concepto de entropía a la hora de construir el árbol, lo cual es la medida de impureza o incertidumbre. Tras un análisis minucioso de las propiedades que debía cumplir la función entropía, Shannon llego a la conclusión de que la mejor función matemática que mide el grado de incertidumbre es la siguiente:

$$E(S) = \sum_{i=1}^{C} -p_i log_2(p_i) \tag{2.1}$$

Donde:

S: es el conjunto de muestras

c: es el número de diferentes clasificaciones

p_I:es la proporción de ejemplos que hay de la clasificación i en la muestra.

En el caso particular de una clasificación binaria la formula anterior queda de la siguiente manera:

$$E(S) = -Plog_2(P) - Nlog_2(N) \tag{3.1}$$

Donde:

P: proporción de ejemplos positivos

N: proporción de ejemplos negativos

La entropía asociada a un atributo X sería:

$$E(T, X) = \sum_{c \in X} p(c)E(D_c) \tag{4.1}$$

Y la ganancia de información que aportaría dividir respecto a los valores de ese atributo:

$$Gain(T, X) = E(T, D) - E(T, X) \tag{5.1}$$

Los árboles de decisión generados por C5.0 pueden ser usados para clasificación, y por esta razón, está casi siempre referido como un clasificador estadístico.

Según (Díaz, 2000, pp. 23-25): El algoritmo C5.0 construye árboles de decisión desde un grupo de datos de entrenamiento de la misma forma en que los hace el C4.5 y ID3, usando el concepto de entropía de información. Los datos de entrenamiento $S = s_1, s_2, ...$ de ejemplos ya clasificados. Cada ejemplo $s_i, = x_1, x_2, ...$ es un vector donde $x_1, x_2, ...$ representan los atributos o

características del ejemplo. Los datos de entrenamiento son aumentados con un vector $C = c_1, c_2, \ldots$ donde $c_1, c_2, \ldots$ representan la clase a la que pertenece cada muestra.

En cada nodo del árbol, C5.0 elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información que resulta en la relación de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión.

Sus principales características son (Weiss et al., 2007, p. 4):

- **Función de división:** Tiene que ver con la eficiencia es el tiempo de construcción de árbol, el uso de memoria y la obtención de árboles considerablemente más pequeños que en el C4.5 con la misma capacidad predictiva.

- **Tipos de variables:** Puede tratar variables discretas y continuas, sin embargo, para poder utilizar gran cantidad de información se recomienda manejar variables continuas para que los árboles obtenidos sean simples y fáciles de entender.

- **Valores *missing:*** El algoritmo tiene la opción de ponderar algunos atributos en manera de enfocar la construcción del árbol

1.6.4 Algoritmo SVM

Máquinas de vectores de soporte (*Support Vector Machines*) son algoritmos de cómputo emergentes desarrolladas por Vapnik y sus colabores, es una de las técnicas más poderosas del aprendizaje automático, que a pesar de su sencillez ha demostrado ser un algoritmo robusto y que generaliza bien en problemas de la vida real (Cristianini et al., 2000, p. 20). La técnica de SVM está basada en la idea de minimización de riesgo estructural (Rest) el cual busca minimizar un límite superior del error de generalización en vez del principio de minimización de riesgo empírico (Remp), puede ser utilizada tanto en problemas de clasificación como de regresión.

Una SVM primero mapea los puntos de entrada a un espacio de características de una dimensión (i.e.: si los puntos de entrada están en R^2). SVM es un clasificador binario que asigna una etiqueta $y \in \{+1, -1\}$ al vector de entrada **x** conforme al signo de la siguiente expresión:

$$f(\boldsymbol{x}) = w^t * \emptyset(\boldsymbol{x}) + b \tag{6.1}$$

Donde: $\emptyset(\boldsymbol{x}) : R^d \rightarrow R^H \ (d < H)$ es una transformación del espacio dimensión, en el que se supone que las clases son linealmente separables. El vector **w** define el hiperplano de separación en dicho espacio y b representa el sesgo respecto al origen de coordenadas (Ureña et al., p. 2).

La solución viene dada por el siguiente problema de minimización cuadrática:

$$min_{w,b,\varepsilon,i} \frac{1}{2} \left|\left|\boldsymbol{w}\right|\right|^2 + C \sum_{i=1}^{n} \xi_i \tag{7.1}$$

$$sujeto\ a: y_i(\boldsymbol{w}^T, \emptyset(\boldsymbol{x_i}) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0; i = 1, \dots, n$$

Donde: $\boldsymbol{x_i} \in R^d (i = 1, \dots, n$ son las muestras de entrenamiento con etiquetas $y_i \in \{+1, -1\}$. Las variables ξ_i miden el error de entrenamiento de cada muestra y C es un factor de ponderación entre el riesgo empírico y el riesgo estructural.

Sus principales características son (Gala García, 2013, p. 25):

- **Función de división:** Tiene que ver hiperplano optimo que mejor separe las clases.
- **Tipos de variables:** Puede tratar variables continuas para que los árboles obtenidos sean simples y fáciles de entender.
- **Valores *missing:*** El algoritmo tiene la opción de ponderar algunos atributos en manera de enfocar la construcción del árbol.

Propiedades de las SVM:

Las principales propiedades y ventajas para el uso de las SVM son las siguientes:

- Tiene una buena generalización con nuevos datos cuando el modelo está bien parametrizado
- El proceso de entrenamiento no depende del número de atributos
- Los modelos dependen de pocos parámetros, C, σ, ε, por lo que la meta modelización es más fácil.
- El modelo final puede ser sencillo, simplemente una combinación de unos pocos vectores soporte.

1.6.5 Algoritmo CART

El algoritmo de clasificación y regresión CART (*Classification and regression trees*), fue desarrollado por Breiman en 1984, consiste en un algoritmo de árbol binario completo que hace particiones de los datos y genera subconjuntos precisos y homogéneos, las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición) formando así grupos homogéneos respecto a la variable que se desea discriminar (Bortolini et al., 2013, p. 741).

CART es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. El nodo inicial es llamado nodo raíz o grupo madre y se divide en grupos hijos o nodos, luego el procedimiento de partición es aplicado a cada grupo hijo por separado. Las particiones se realizan de manera recursiva hasta donde el árbol se construye fragmentando sucesivamente el conjunto de datos (Timofeev, 2004, p. 20).

La construcción del árbol se realiza siguiendo un enfoque de división binaria recursiva, sea N_j el número de casos en la clase j y $\pi(j) = \frac{N_j}{N}$ las probabilidades de que un dato en la clase esté presente en el árbol, donde N es el número de datos. El estimador de probabilidad de que un caso esté en la clase j dado que se ubicó en el nodo t; esta dado por:

$$p(j|t) = \frac{p(j,t)}{p(t)} = \frac{N_j(t)}{N_j} \tag{8.1}$$

Y cumple:

$$\sum_j p(j|t)\, j = 1 \tag{9.1}$$

Así, las $p(j|t)$ son las proporciones relativas de los casos en la clase j en el nodo t (3)

Para elegir la mejor variable debe utilizarse una medida de pureza (*purity*) en la valoración de los 2 nodos posibles. Una de las funciones más utilizadas es la denominada Gini que alcanza un índice de pureza que se considera como máximo (Serna Pineda, 2009, p. 21).

Índice de Gini en el nodo t, i(t), se puede formular de la manera siguiente:

$$i(t) = \sum_{j \neq i} p(j|t) * p(i|t) \tag{10.1}$$

Donde i y j son las características de la variable predictora y p es la proporción.

Sus principales características son:

- **Función de división:** CART divide los datos en segmentos para que sean lo más homogéneos posible respecto a la variable dependiente.

- **Tipos de variables:** Puede tratar con variables discretas -nominales y ordinales- y continuas. Se usará clasificación cuando la variable objetivo es discreta, mientras que se usará regresión cuando la variable objetivo es continua

- **Poda del árbol:** Se basa en la idea de impureza. CART selecciona el corte que conduce al mayor decrecimiento de la impureza.

1.6.6 Pasos para realizar un algoritmo de clasificación

- Importar los datos
  - Ingresar la matriz de datos con la que se presente trabajar.

- Limpiar el conjunto de datos
  - Existen valores NA´S, por lo tanto, deben ser eliminados.
  - Prescindir de variables innecesarias.
  - Crear – convertir variables a tipo factor

- Dividir en conjuntos de entrenamiento y test
  - Antes de entrenar el modelo se debe dividir como conjunto de datos de entrenamiento y de test. La práctica común es 70 – 30. La función que se utiliza para seleccionar aleatoriamente los dos conjuntos de datos se llama *createDataPartition()*.

- Construir el modelo
  - Con los datos de entrenamiento se procede a realizar la construcción del modelo. El comando para generar un modelo de árbol de decisión es usando la función *train()*.

- Hacer la predicción
  - El modelo entrenado se utiliza para predecir nuevas instancias en el conjunto de datos de prueba. Para esto se usa la función *predict().*

- Medir el rendimiento del modelo
  - Para determinar una medida del rendimiento del modelo se utiliza una matriz de confusión (Fawcett, 2006, p. 865). Para esto se utiliza la función *confusionMatrix()*.

Tabla 6-1: Matriz de Confusión.

Clase original	Clase Predicha	
	Positivo (P)	Negativo (N)
Positivo (P)	VP (Verdadero positivo)	FN (Falso Negativo)
Negativo (N)	FP (Falso Positivo)	VN (Verdadero Negativo)

Fuente: (Fawcett, 2006, p. 865).
Realizado por: Padilla S., Oscar R., 2020.

- Medidas de rendimiento a partir de la matriz de confusión o tabla de contingencia

Precisión Global P (Exactitud):

$$P = \frac{VN + VP}{VN + FP + FN + VP} \quad (11.1)$$

Precisión Positiva (Sensibilidad):

$$PP = \frac{VP}{FN + VP} \quad (12.1)$$

Precisión Negativa (Especificidad):

$$PN = \frac{VN}{VN + FP} \quad (13.1)$$

Clasificación errónea (Error):

$$E = \frac{FN + FP}{VN + FP + FN + VP} \quad (14.1)$$

Coeficiente Kappa

Es un coeficiente estadístico que se emplea para cuantificar el grado de acuerdo entre los observadores, corrige el factor azar. Es el estudio de fiabilidad por equivalencia o concordancia entre observadores. Cuando el valor obtenido es menor que -1 se dice que las variables tienen poca relación mientras si el valor es cercano a 1, se dice que existe una fuerte relación entre las variables (López de Ullibarri y Pita Fernández, 2010, p. 171).

La ecuación para K es:

$$k = \frac{P_r(a) - P_r(e)}{1 - P_r(e)} \tag{15.1}$$

Donde:

$P_r(a)$ es el acuerdo observable relativo entre los observadores

$P_r(a)$ es la probabilidad hipotética de acuerdo por azar.

A la hora de interpreta el valor de k es muy útil disponer de una escala como la que se presenta en la Tabla 7-1:

Tabla 7-1: Valoración del índice kappa.

Valor de k	Fuerza de concordancia
< 0.20	Pobre
0.21 – 0.40	Débil
0.41 – 0.60	Moderada
0.61 – 0.80	Buena
0.81 – 1.00	Muy Buena

Fuente: (López de Ullibarri y Pita Fernández, 2010).

Realizado por: Padilla S. Oscar R., 2020.

1.7 Descripción de las variables

1.7.1 Variable predictora

La primera tarea fue analizar la información de la base de datos en la que se extrae las siguientes variables de interés: profundidad de muestreo, densidad aparente (gcm-3), materia orgánica del suelo (%) y concentración o proporción de carbono orgánico (%). La profundidad de muestreo es definida a 30cm por el proyecto a investigar donde el carbono orgánico es más dinámico. Para determinar la concentración de CO se colectan submuestras con barreno a la profundidad definida las submuestras se toman en forma de zigzag o en trayectos, idealmente dentro de la parcela de muestreo, y se mezclan para formar una muestra compuesta que será llevada al laboratorio de suelos. La densidad aparente del suelo se toma a la misma profundidad de donde de donde se extraen las muestras para CO, con un cilindro de volumen conocido (100 cm^3), en al menos 3 sitios por parcela de muestreo.

Posteriormente, la estimación del contenido de carbono orgánico total del suelo se determinó a partir de la siguiente ecuación propuesta por González et al (2008):

$$COS = \%CO * DAP * Ps \quad (16.1)$$

donde:

COS: Carbono orgánico de suelo (Mg/ha)

%CO: Concentración de carbono orgánico en suelos

DAP: Densidad aparente (g/cm^3)

Ps: Profundidad de suelo (cm)

En cuanto a la ausencia de datos de alguna de estas variables, se utilizó modelos de regresión a partir de datos existentes. En el caso de la densidad aparente se procedió aplicar el método de Grigal et. al (1989), a través de la siguiente formula:

$$DAP = 0.669 + 0.941 * e^{-0.06*MO} \quad (17.1)$$

Para las variables restantes se aplicó el factor de conversión de Van Benmelen de 1.724 que resulta de la suposición de que la materia orgánica del suelo contiene un 58% de Carbono (1/0.58 = 1.724).

Para el caso de materia orgánica:

$$MO = 1.724 * \%CO \quad (18.1)$$

Con el fin de determinar la concentración de carbono orgánico:

$$\%CO = \frac{MO}{1.725} \quad (19.1)$$

1.7.2 Variables explicativas

Las variables explicativas que se tomó en consideración para el estudio fueron: ráster de Ecosistema, Taxonomía, Textura, Pendiente, DEM y los índices espectrales, los mismos que están relacionados con el carbono orgánico del suelo (COS).

Búsqueda de los índices espectrales

Para encontrar los índices espectrales se procedió a la búsqueda de estudios relacionados con el suelo.

Tabla 8-1: Búsqueda de los índices espectrales.

N°	Estudios Analizados	Año de publicación
1	El índice normalizado diferencial de la vegetación como indicador de la degradación del bosque (Meneses Tovar, 2012).	2012
2	A soil-adjusted vegetation index (Huete, 1988).	1988
3	Apllication of time series of remotely sensed normalized difference wáter, vegetation and moisture indices in characterizing flood dynamics of large-scale arid zone floodplains (Mohammadi et al., 2017).	2017
4	Assessing the Robusteness of Vegetation Indices to Estimate Wheat N in Mediterranean Environments (Cammarano et al., 2014).	2014
5	Caracterización ecohidrológica de humedades alto andinos usando imágenes de satélite multitemporales en la cabecera de cuenca del río Santa, Ancash, Perú (García y Otto, 2015).	2015
6	Comparison of remote sensing indices for monitoring desert cienegas (Wilson et al., 2016).	2016
7	Comparación espacial y temporal de índices de la vegetación para verdor y humedad y aplicación para estimar LAI en el Desierto Sonorense (Rodríguez Moreno y Bullock, 2013).	2013
8	Delimitación y análisis del incendio forestal de Sierra de Gata (Cáceres) mediante imágenes de los satélites Landsat 8 y Sentinal 2 (Nieto A. et al., 2017).	2017
9	Development of a two-band enhanced vegetation index without a blue band (Jiang et al., 2018).	2018
10	Múltiples índices espectrales para predecir la variabilidad estructurales y funcionales en zonas áridas (Buzzi et al., 2017).	2017
11	Comparación de índices de vegetación a partir de imágenes MODIS en la región del Libertador O'Higgins, Chile en el período 2001-2005 (Carvacho Bart y Marcela, 2010).	2010
12	Incendios forestales y trasferencia de carbono biomasa-suelo en áreas montañosas de clima atlántico (Fernández Menéndez et al., 2011).	2011

Realizado por: Padilla S., Oscar R., 2020.

Selección de los índices espectrales encontrados

Una vez obtenidos resultados de la Tabla 8-1 se buscó los índices usados en cada estudio, se contabilizó la cantidad de veces en las que fue usado cada índice y se los presentó en la siguiente tabla comparativa.

Tabla 9-1: Obtención de los índices espectrales.

Variables	Estudios Analizados												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
NDVI	x		X	x	x	x	x	x	x	x	x		10
SAVI		x				x	x			x	x		5
TSAVI							x						1
VARI		x							x		x	x	4
NDWI			X	x		x							3
LSWI			X										1
NDII			X			x							1
BI			X		x		x	x			x		5
TWI						x							1
NBR2				x				x				x	3
EVI2									x	x	x		3

Realizado por: Padilla S., Oscar R., 2020.

En la Tabla 9-1, se observa que los índices más utilizados en los estudios son:

- Índice de vegetación de diferencia Normalizada (NDVI).
- Índice de Vegetación Ajustado al Suelo (SAVI).
- Índice de resistencia atmosféricamente visible (VARI).
- Índice diferencial de Agua Normalizada (NDWI).
- Índice de área calcinada (BI).
- Índice normalizado de áreas quemadas 2 (NBR2).
- Índice de vegetación mejorado de dos bandas (EVI2).

Por lo tanto, estos índices son aplicados como variables explicativas

1.7.3 Índices espectrales

Los índices espectrales son calculados de la imagen satelital Landsat 8 de órbita hello-sincrónica a 705 km del ecuador terrestre, el tamaño aproximado es de 170km escena de norte a sur por 183 km de este a oeste, incorpora dos instrumentos de barrido: *Operational Land Imager* (OLI) con 9 bandas espectrales que capturan el espectro visible, infrarrojos y espectros de radiación de ondas bajas y un sensor *Thermal infrared Sensor* (TIRS) con 2 bandas espectrales que detecta infrarrojos térmicos y es usado para medir la temperatura de la superficie de la Tierra. La resolución radiométrica de las imágenes Landsat 8 es de 16 bits, la temporal es de 16 días y son imágenes de acceso libre (Survey, 2019, p. 10).

Tabla 10-1: Descripción de la imagen Landsat 8.

Sensor	# Banda	Resolución espacial	Banda	Longitud de onda (µm)
Imagen Operacional de la Tierra (OLI)	Banda 1	30 m	Ultra azul (costero / aerosol)	0.43 - 0.45
	Banda 2	30 m	Azul	0.45 - 0.51
	Banda 3	30 m	Verde	0.53 - 0.59
	Banda 4	30 m	Rojo	0.64 - 0.67
	Banda 5	30 m	Infrarrojo Cercano (NIR)	0.85 - 0.88
	Banda 6	30 m	Onda corta infrarroja SWIR-1	1.57 - 1.65
	Banda 7	30 m	Onda corta infrarroja SWIR-2	2.11 - 2.29
	Banda 8	15 m	Pancromática	0.50 - 0.68
	Banda 9	30 m	Nubes Cirrus	1.36 - 1.38
Sensor térmico infrarrojo (TIRS)	Banda 10	100 m	Infrarrojo térmico (TIRS) 1	10.60 - 11.19
	Banda 11	100 m	Infrarrojo térmico (TIRS) 2	11.50 - 12.51

Fuente: (Survey, 2019, p. 10).
Realizado por: Padilla S., Oscar R., 2020.

Los índices espectrales asociado al COS, son el resultado de la transformación de dos o más bandas, que contribuyen en la mejora del análisis de las propiedades de la vegetación y permiten obtener la distribución espacial de contenido de carbono y otros componentes, así como la estructura de la vegetación de manera fiable y rápida (Huete et al., 2002, p. 197).

A continuación, se mencionan los índices espectrales con más detalle, mismos que más adelante serán puestos en evaluación por su capacidad indicadora como variable que controlan el almacenamiento de COS.

1.7.3.1 Índice de Vegetación de Diferencia Normalizada (NDVI)

Normalized Difference Vegetation Index (por sus siglas en inglés). Es un parámetro numérico que permite estimar y evaluar el estado de salud de la vegetación, calculados a partir de valores de reflectancia (radiación que las plantas emiten o reflejan) a distintas longitudes de onda, es particularmente sensible a la cubierta vegetal (Ortega Gutiérrez, 2015, p. 9).

El NDVI es un índice no dimensional y que al ser normalizado toma valores entre -1 y +1. Si el resultado es menor que 0.1, se trata de cuerpos de agua o tierra desnudas; mientras que, si toma valores cercanos a 1, su interpretación es la existencia de una mayor actividad fotosintética (Meneses Tovar, 2012, p. 40). En las zonas que contienen agua o nubes, el valor de NDVI es siempre menor que 0 (Rouse et al., 1974, pp. 311-313).

La ecuación para el cálculo de este índice en base a las bandas de un sensor se presenta a continuación:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{20.1}$$

Donde:

NIR: Reflectancia de la banda infrarrojo cercano

Red: Reflectancia de la banda verde

El NDVI resta valores de reflectancia de color rojo desde el infrarrojo cercano y lo divide para la suma de las bandas del infrarrojo cercano y rojo (Yates et al., 1984, p. 221). Es muy utilizado para evaluar el vigor de la vegetación, es decir, para diferenciar las coberturas de vegetación, según la densidad y salud (Wilson et al., 2016, p. 465).

Algunos autores han aplicado técnicas de análisis espectral para caracterizar humedades altoandinas, como el caso del NDVI, que ha sido aplicado para identificar humedales perennes y temporales, donde los valores de los umbrales están asociados a los periodos de precipitación (Otto et al., 2017, p. 1715).

1.7.3.2 Índice de Vegetación Ajustado al Suelo (SAVI)

Soil Adjusted Vegetation Index (por sus siglas en inglés). Fue desarrollado por Huete en 1988. Se trata de un índice muy adecuado para trabajos en zonas semiáridas (Ambiente natural de baja productividad y el agua es limitante), donde la contribución del suelo es muy importante, se utiliza

con el fin de corregir algunos efectos externos al valor del NDVI en lugares donde el aporte de la reflectividad del suelo es elevado (Huete, 1988, p. 297).

Para zonas semiáridas el índice SAVI resulta más consistente que el NDVI gracias a esa mayor distinción entre el suelo y la vegetación (Wilson et al., 2016, p. 466).

$$SAVI = \frac{NIR - Red}{NIR + Red + L}(1 + L) \tag{21.1}$$

Donde:

NIR: Reflectancia de la banda infrarrojo cercano

Red: Reflectancia de la banda rojo

L es el factor "línea del suelo". Los valores varían entre 0 y 1, dependiendo de la densidad de la vegetación, L= 0.5 densidades intermedias, L = 0 Mucha vegetación, L = 1 No hay vegetación. Si L = 0 SAVI es igual al NDVI (Sánchez Rodríguez et al., 2000, p. 164).

1.7.3.3 Índice de resistencia atmosféricamente visible (VARI)

Atmospheric Visible Resistance Index (por sus siglas en inglés). Kaufman y Tanré, citados por Chuvieco (2008) proponen un ajuste del NDVI a las condiciones atmosféricas, teniendo en cuenta la diferente dispersión de los canales azul y rojo del espectro (Chuvieco Salinero, 2008, p. 60). Está diseñado para resaltar la vegetación en la parte visible del espectro, a la vez que mitiga las diferencias en la iluminación y los efectos atmosféricos (Cammarano et al., 2014, p. 2831).

$$VARI = \frac{Green - Red}{Green + Red - Blue} \tag{22.1}$$

Donde:

Green: Reflectancia de la banda verde

Red: Reflectancia de la banda rojo

Blue: Reflectancia de la banda azul

1.7.3.4 Índice diferencial de agua normalizada (NDWI)

Normalized Differential Wáter Index (por sus siglas en inglés). Es un indicador numérico, derivada de imágenes satelitales ópticas, usando las ondas de infrarrojo cercano y corta las bandas espectrales del infrarrojo. Es muy empleado para evaluar cuerpos de agua abierta.

El NDWI es útil en muchas aplicaciones de teledetección como en vigilancia de los cultivos utilizados en la salud, la cartografía de la tierra / agua de -1 embarque, la discriminación de agua hacia el interior de los cuerpos de agua de mar abierto, etc. (McFeeters, 1996, p. 1427).

$$NDWI = \frac{Green - NIR}{Green + NIR} \tag{23.1}$$

Donde:

Green: Reflectancia de la banda verde

NIR: Reflectancia de la banda infrarrojo cercano

1.7.3.5 Índice de área calcinada (BI)

Calcined Área Index (por sus siglas en inglés). Es un indicador numérico que combina azul, rojo y onda corta infrarroja con bandas espectrales del infrarrojo cercano, para capturar las variables del suelo. Es capaz de identificar superficies de suelo desnudo, en ambientes de baja intervención antrópica; este tipo de suelo podría aportar información relevante sobre la cantidad de COS debajo del mismo (Wanhui Chen et al., 2004, pp. 3379-3382).

$$BI = \frac{(SWIR1 + Red) - (NIR + Blue)}{(SWIR1 + Red) + (NIR + Blue)} \tag{24.1}$$

Donde:

SWIR1: Reflectancia de la banda infrarrojo de onda corta 1

Red: Reflectancia de la banda rojo

Blue: Reflectancia de la banda azul

1.7.3.6 Índice normalizado de Áreas Quemadas 2 (NBR2)

Normalized Index of Burned Areas 2 (por sus siglas en inglés). Utilizan la reflectancia de la banda infrarrojo de onda corta 1 y 2 para destacar las áreas de calcinación a la vez que mitiga las diferencias en la iluminación y los efectos atmosféricos (Bravo Morales, 2017, p. 84). Los incendios forestales cometidos por el hombre, o fenómenos naturales que destruyen los recursos naturales, desequilibra al medio ambiente (Nieto A. et al., 2017, p. 3).

$$NBR2 = \frac{SWIR1 - SWIR2}{SWIR1 + SWIR2} \tag{25.1}$$

Donde:

SWIR1: Reflectancia de la banda infrarrojo de onda corta 1

SWIE2: Reflectancia de la banda infrarrojo de onda corta 2

1.7.3.7 Índice de vegetación mejorado de dos bandas (EVI2)

Enhanced Vegetation Index of two bands (por sus siglas en inglés). Está orientado en optimizar la respuesta espectral de la vegetación con alta densidad, desarrollando el proceso de reducción de la influencia de la atmósfera; también fue aplicado para monitoreo multitemporal de coberturas terrestres y en análisis de rendimiento de cultivos (Jiang et al., 2018, p. 3835).

$$EVI2 = 2.5 * \frac{NIR - Red}{NIR + 2.4 * Red + 1} \quad (26.1)$$

Donde:

NIR: Reflectancia de la banda infrarrojo cercano

Red: Reflectancia de la banda rojo

1.8 Herramientas de software

Actualmente existen varias herramientas tecnológicas que permiten identificar comportamientos o patrones entre los datos, una de las herramientas más utilizadas tenemos:

1.8.1 R Studio

R es un software estadístico que fue diseñado para hacer análisis estadísticos y gráficas, es un software libre. Es un entorno de desarrollo para R e incluye una consola y un editor de resaltado de sintaxis que admite la ejecución de código, así como herramientas para el historial, depuración y administración de datos (RStudio). R proporciona una amplia variedad de técnicas estadísticas tales como clasificación y técnicas gráficas altamente extensibles para la manipulación de datos. Una de las ventajas de R es la excelente calidad de los gráficos que produce (R-project).Es por ello que para la aplicación de los algoritmos de clasificación supervisados en el presente estudio se utilizó R Studio (Díaz, 2000).

CAPITULO II

2. MARCO METODOLÓGICO

2.1 Tipo y diseño de la investigación

La investigación según:

- El método de la investigación se considera mixta, ya que la información se concentra en variables estadísticas cuantitativas y cualitativas.
- El objetivo es aplicada, ya que a partir de los conocimientos adquiridos se darán soluciones para la clasificación del contenido del COS establecido en la Evaluación Nacional Forestal MAE – FAO mediante técnicas estadísticas y geo referenciales.
- El nivel de profundización en el objeto de estudio es exploratorio, debido a que se estudiará el COS en los diferentes tipos de suelos.
- La manipulación de variables es Cuasi Experimental, ya que se generará intervalos de clasificación para el carbono edáfico.
- El tipo de inferencia es inductiva deductivo, deductiva debido que la técnica a utilizar parte desde una información grande e inductiva debido a que los resultados serán exportados para estudios a nivel nacional.
- Según el periodo temporal es transversal, pues se trabajará con datos definidos en un período de tiempo de acuerdo a la disponibilidad de información de la base de datos de Evaluación Nacional Forestal MAE - FAO.

2.2 Identificación de las variables

Variable dependiente:

- Contenido de Carbono Orgánico de Suelo (COS)

Variables Independientes:

- Taxonomía de los suelos

- Textura de los suelos
- Pendiente
- Ecosistema
- Modelos de Elevación digital (DEM)
- Índice de Vegetación de Diferencia Normalizado (NDVI)
- Índice de Vegetación Ajustado al Suelo (SAVI)
- Índice de Resistencia Atmosféricamente Visible (VARI)
- Índice Diferencial de Agua Normalizada (NDWI)
- Índice de Área Calcinada (BI)
- Índice Normalizado de Áreas Quemadas 2 (NBR2)
- Índice de Vegetación Mejorado de Dos Bandas (EVI2)

2.3 Operacionalización de las variables

Tabla 1-2: Operacionalización de las variables

Variables	Descripción	Indicador	Instrumento
Contenido de carbono orgánico del suelo (COS)	Es un indicador de la calidad de los suelos minerales y sirve para detectar cambios de carbono en el tiempo	Contenido de COS	Software estadístico R y geográfico QGIS
Ecosistema	Es un sistema biológico constituido por una comunidad de organismos vivos y el medio físico donde se relacionan	Tipos de ecosistema	Software geográfico QGIS

Taxonomía	Es la clasificación de los suelos en función de varios parámetros y propiedades que se desarrolla en niveles	Tipos de suelo	Software geográfico QGIS
Textura	Indica el contenido relativo de partículas de diferente tamaño	Porcentaje de partículas según Arenas, limos y arcillas	Software geográfico QGIS
Pendiente	Es un declive del terreno y la inclinación, respecto a la horizontal de una vertiente	Altura sobre el nivel del mar	Software geográfico QGIS
DEM	Representación de ráster de una superficie continua, que en general hace referencia a la superficie de la tierra.	Altura sobre el nivel del mar	Software geográfico QGIS
NDVI	Índice de vegetación de diferencia normalizado, resalta la contribución de la vegetación a la respuesta espectral de la superficie.	Escala de valores en función a la respuesta espectral (-1, 1)	Bandas Multiespectrales
SAVI	Índice de vegetación ajustado al suelo, adecuado para trabajos semiáridas resalta la mayor distinción entre el suelo y la vegetación que al NDVI.	Escala de valores en función a la respuesta espectral (-1, 1)	Bandas Multiespectrales
VARI	Índice de resistencia atmosféricamente visible, resalta la vegetación en la parte visible del espectro, a la vez mitiga las	Escala de valores en función a la respuesta espectral (-1, 1)	Bandas Multiespectrales

	diferencias en la iluminación y los efectos atmosféricos.		
NDWI	Índice diferencial de agua normalizado, es muy empleado para evaluar cuerpos de agua cubierta, útil en aplicaciones de teledetección como en vigilancia de los cultivos de la salud	Escala de valores en función a la respuesta espectral (-1, 1)	Bandas Multiespectrales
BI	Índice de área calcinada, capaz de identificar superficies de suelo desnudo, en ambientes de baja intervención antrópica.	Escala de valores en función a la respuesta espectral (-1, 1)	Bandas Multiespectrales
NBR2	Índice normalizado de áreas quemadas 2, destaca las áreas de calcinación a la vez que mitiga las diferencias en la iluminación y los efectos atmosféricos.	Escala de valores en función a la respuesta espectral (-1, 1)	Bandas Multiespectrales
EVI2	Índice de vegetación mejorado de dos bandas, está orientado en optimizar la respuesta espectral de la vegetación con alta densidad, desarrollando el proceso de reducción de la influencia de la atmosfera	Escala de valores en función a la respuesta espectral (-1, 1)	Bandas Multiespectrales

Realizado por: Padilla S., Oscar R., 2020

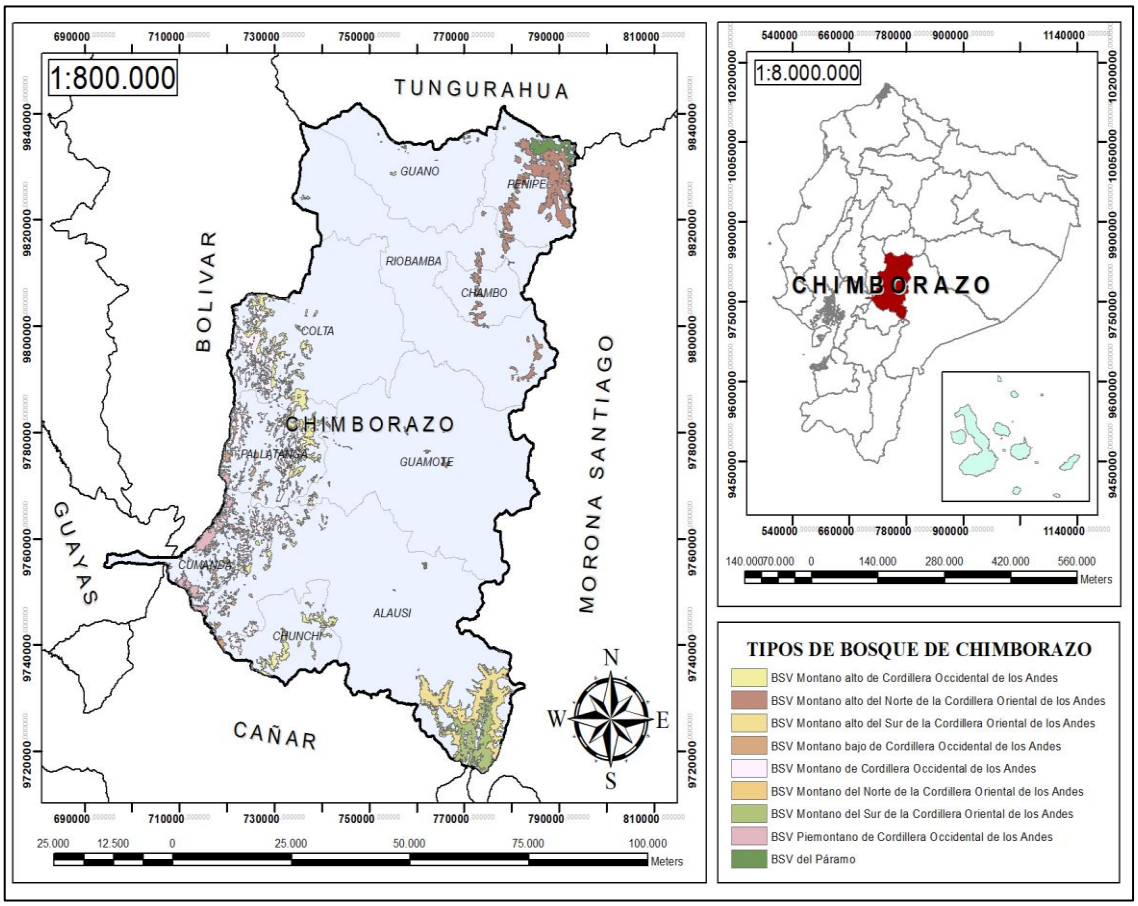
2.4 Características del lugar

2.4.1 Descripción del área de estudio

La Provincia de Chimborazo, está situada en el centro del sur del país, en la zona geográfica 17 Sur, en la región Interandina conocida como la provincia de las Altas Cumbre, debido a que se encuentra el volcán Chimborazo, el nevado más alto del Ecuador con una altura de 6.310 msnm, volcán que da nombre a la provincia. Limita al norte con Tungurahua, al sur con Cañar, por el

occidente con Bolívar, al suroeste con Guayas y al este con Morona Santiago. El área de la provincia de Chimborazo abarca 6500 km^2. Cuenta con 10 cantones, 61 parroquias de los cuales 45 son rurales y 16 son urbanas; y tiene una población de 524 004 habitantes según la proyección del Instituto Ecuatoriano de Estadísticas y Censo (INEC) del año 2020, siendo la novena provincia más poblada del Ecuador.

2.4.2 Localización del área de estudio



Gráfica 1-2: Ubicación de la provincia de Chimborazo y bosques existentes.
Fuente: (INEC, 2010).
Realizado por: Padilla S. Oscar R., 2020 – GIDAC.

2.4.3 Ubicación geográfica

Latitud: 1°40’0” S

Longitud: 78°39’0” W

Altitud: 3900 m.s.n.m.

2.5 Población de estudio

El estudio se realizó con los datos de carbono edáfico a una profundidad de 30 cm, registrada en la base de datos de Evaluación Nacional Forestal MAE – FAO gestionado por el Ministerio del Ambiente del Ecuador y datos del Proyecto Regional de Cooperación de Capacitación de Mapeo de Suelos de la FAO gestionada por el Ministerio de Agricultura y Ganadería.

2.6 Tamaño de la muestra

Se trabajó con la base de datos de Evaluación Nacional Forestal MAE – FAO que contenía un total de 590 datos y la base del Proyecto Regional de Cooperación de Capacitación de Mapeo de Suelos de la FAO que contenía un total de 644 datos, además se trabajó con los mapas digitales de Carbono orgánico de Suelo del Ecuador elaborada por el ministerio de Agricultura y Ganadería (MAG) y el mapa Global Soil Organic Carbon Map (GSOCmap) elaborada por la Organización de las Naciones unidas para la Alimentación y a la Agricultura (FAO) .

2.7 Método de muestreo

Para el presente trabajó de investigación no surgió la necesidad de utilizar técnicas de muestreo, debido a que se trabajó con la base de datos de Evaluación Nacional Forestal MAE – FAO y con la base del Proyecto Regional de Cooperación de Capacitación de Mapeo de Suelos de la FAO en el marco de la Alianza Regional por los Suelos.

2.8 Técnica de recolección de información

La recolección de muestras se efectuó a 30 cm de profundidad a través del método de barrenación, el cual consistió en recorrer el área e insertar el barreno en el suelo para obtener el porcentaje de muestra requerida. Para los datos del Proyecto Regional de Cooperación de Capacitación de Mapeo de Suelos de la FAO se aplicó el método Walkley Black el cual obtiene las variables de densidad aparente (g.cm-3) y materia orgánica (%), con un factor de 1.724 se trasforma materia orgánica en CO (%). Los datos de Evaluación Nacional Forestal MAE – FAO se obtuvo por medio del Analizador TOC el cual obtiene de manera directa el CO (%).

2.9 Modelo Estadístico

La clasificación del contenido de carbono edáfico se realizó mediante el análisis de árboles de decisión y se visualizó mediante el análisis Geo Referencial.

CAPITULO III

3. RESULTADOS Y DISCUSIÓN

Para el cumplimiento de los objetivos del presente trabajo, se utilizó como base la metodología CRISP-DM, debido a que es una de las más usadas para analizar grandes conjuntos de datos y descubrir información valiosa, está compuesta por seis fases o etapas: Definir y analizar el problema, Exploración de la data, Preprocesamiento de la data, Modelado, Evaluación e Implementación.

3.1 Etapa 1: Definir y analizar el problema

Para el desarrollo del trabajo se utilizaron datos de Evaluación Nacional Forestal proporcionadas el Ministerio del Ambiente del Ecuador (MAE), e información del Ministerio de Agricultura y Ganadería (MAG). siendo RStudio, la herramienta empleada en el desarrollo de la investigación. Los diferentes métodos y técnicas científicas sirvieron para cumplir con los objetivos planteados y responder a la pregunta de investigación: ¿La construcción y análisis de árboles de decisión, permitirá valorar adecuadamente el contenido de carbono edáfico en la provincia de Chimborazo, en los distintos tipos de suelo?

Partiendo de lo explicado, en esta etapa se consideran los siguientes puntos:

3.1.1 Descripción de las variables y su proceso de obtención

Se identificaron las variables útiles para el desarrollo del trabajo y se evaluó la validez del conjunto de datos.

3.1.1.1 Variable predictora

La base de datos proporcionada por el MAG contiene 644 puntos de muestreo en la provincia de Chimborazo, información utilizada en su totalidad, mientras que los datos proporcionados por el MAE contiene 590 puntos de muestreo en la región Interandina, de los cuales únicamente se utilizaron 434 datos debido a que el resto de puntos no contiene información en las variables de interés seleccionadas, esto se debió a que no se pudo realizar el proceso de recolección de muestras de suelo, puesto que para esos puntos no fue posible el acceso; ya sea por motivos de pendientes elevadas, rocas o porque no se tuvo el permiso respectivo por parte de los propietarios; como por ejemplo en el caso de comunidades indígenas.

3.1.1.2 Variables explicativas

Las variables explicativas que se tomó en consideración para el estudio fueron: ráster de Ecosistema, Taxonomía, Textura, Pendiente, DEM y los índices espectrales, los mismos que están relacionados con el carbono orgánico del suelo (COS).

La generación de los índices espectrales se obtuvo a partir de las bandas de la imagen Satelital Landsat 8, a continuación, se exponen las características de la imagen Landsat 8:

Tabla 1-3: Características generales para descargar las imágenes satelitales de Landsat 8.

Descripción Atributo	Valores Imagen Carchi y Pichincha	Valores Imagen Chimborazo y Pastaza	Valores Imagen Azuay	Valores Imagen Zamora Chinchipe
Identificador	LC08_L1TP_ 010060_2017092_ 20171012_01_T1	LC08_L1TP_ 010061_2017092_ 20171012_01_T1	LC08_L1TP_ 010062_2017092_ 20171012_01_T1	LC08_L1TP_ 010063_2017092_ 20171012_01_T1
Escena Landsat	LC80100602017263 LGNOO	LC80100612017263 LGNOO	LC80100622017263 LGNOO	LC80100632017263 LGNOO
Fecha de Adquisición	20/9/2017	20/9/2017	20/9/2017	20/9/2017
Época de adquisición	Verano	Verano	Verano	Verano
WRS Path/ Row	010 / 060	010 / 061	010 / 061	010 / 061
Cobertura de nubes	34.27%	32.44%	32.39%	19.62%
Nombre del sensor remoto	OLI / TIRS	OLI / TIRS	OLI / TIRS	OLI / TIRS
Formato de salida	GEOTIFF	GEOTIFF	GEOTIFF	GEOTIFF
Tamaño imagen pancromática	15 x 15 m	15 x 15 m	15 x 15 m	15 x 15 m
Tamaño imagen multibanda	30 x 30 m	30 x 30 m	30 x 30 m	30 x 30 m
Zona	WGS 84 - UTM 17	WGS 84 - UTM 17	WGS 84 - UTM 17	WGS 84 - UTM 17

Fuente: Landsat 8.
Realizado por: Padilla S., Oscar R., 2020.

En la Tabla 2-3, se muestran cada uno de los índices espectrales con su expresión en función del espectro electromagnético y en función de las bandas de Landsat 8.

Tabla 2-3: Expresión de los índices espectrales calculados en función del espectro electromagnético y de acuerdo con las bandas de Landsat 8.

Índice	Expresión matemática en función del espectro electromagnético	Expresión matemática en función de las bandas de Landsat 8
NDVI	$\frac{NIR - Red}{NIR + Red}$	$\frac{B5 - B4}{B5 + B4}$
SAVI	$\frac{NIR - Red}{NIR + Red + L}(1 + L)$	$\frac{B5 - B4}{B5 + B4 + L}(1 + L)$
VARI	$\frac{Green - Red}{Green + Red - Blue}$	$\frac{B3 - B4}{B3 + B4 - B2}$
NDWI	$\frac{Green - Swir1}{Green + Swir1}$	$\frac{B3 - B5}{B3 + B5}$
BI	$\frac{(SWIR1 + Red) - (NIR + Blue)}{(SWIR1 + Red) + (NIR + Blue)}$	$\frac{(B6 - B4) - (B5 - B2)}{(B6 - B4) + (B5 + B2)}$
NBR2	$\frac{SWIR1 - SWIR2}{SWIR1 + SWIR2}$	$\frac{B6 - B7}{B6 + B7}$
EVI2	$2.5 * \frac{NIR - Red}{NIR + 2.4Red + 1}$	$2.5 * \frac{B5 - B4}{B5 + 2.4B4 + 1}$

Realizado por: Padilla S., Oscar R., 2020.

Obtención de las variables explicativas

Con la ayuda de las coordenadas "X", "Y" de cada punto de muestreo y utilizando el Software geoestadístico QGIS V.3.14, se extrae información de los ráster de Ecosistema, Taxonomía, Textura, Pendiente, DEM y de los índices espectrales definidos anteriormente, obteniendo así doce variables explicativas.

Unificar categorías de las variables

Únicamente la variable Ecosistema requirió el proceso de unificar categorías lo cual se muestra a continuación con los siguientes niveles.

Tabla 3-3: Variable ecosistema unificada.

Ecosistema	Unificada
Intervención	Intervención
Bosque siempre verde pie montano de Cordillera Occidental de los Andes	BSV_Piemontano
Bosque siempre verde montano bajo de Cordillera Occidental de los Andes	BSV_Montano
Bosque siempre verde montano de Cordillera Occidental de los Andes	
Bosque siempre verde montano alto de Cordillera Occidental de los Andes	
Bosque siempre verde montano alto del Norte Cordillera Oriental de los Andes	
Arbustal siempre verde y Herbazal del Paramo	Arbustal
Arbustal siempre verde montano del norte de los Andes	
Herbazal del Paramo	Herbazal
Herbazal Húmedo montano alto superior del Paramo	

Realizado por: Padilla S., Oscar R., 2020.

3.1.1.3 Organizar el conjunto de datos

El conjunto de datos para la valoración de carbono edáfico contó con 12 variables explicativas, en la tabla siguiente se muestra las categorías de cada una de las variables explicativas.

Tabla 4-3: Valores y Niveles que pueden tomar las variables explicativas.

Nº	Variables predictoras	Tipo de variable	Tipos de datos	Niveles de Medición	Valores que puede tomar la variable
1	Ecosistema	Cualitativa	Politómica	Nominal	Intervención, BSV_Piemontano, BSV_Montano, Arbustal y Herbazal
2	Taxonomía	Cualitativa	Politómica	Nominal	Alfisol, Entisol, Histosol, Inceptisol, Mollisol, Vertisol
3	Textura	Cualitativa	Politómica	Ordinal	Fina. Media, Moderada y Gruesa
4	Pendiente	Cualitativa	Politómica	Nominal	Colinado, Escarpado, Montanoso, Ondulado y Plana

5	DEM	Cuantitativa	Discreta	Razón	Altura sobre el nivel el mar que se realiza el punto de muestreo
6	NDVI	Cuantitativa	Continua	Intervalo	Índice de vegetación de diferencia normalizada visible toma valores entre -1 y 1
7	SAVI	Cuantitativa	Continua	Intervalo	Índice de vegetación ajustado al suelo toma valores entre -1 y 1
8	VARI	Cuantitativa	Continua	Intervalo	Índice de resistencia atmosféricamente visible toma valores entre -1 y 1
9	NDWI	Cuantitativa	Continua	Intervalo	Índice diferencial de agua normalizada toma valores entre -1 y 1
10	BI	Cuantitativa	Continua	Intervalo	Índice de área calcinada puede tomar valores entre -1 y 1
11	NBR2	Cuantitativa	Continua	Intervalo	Índice normalizado de Áreas Quemadas 2 toma valores entre -1 y 1
12	EVI2	Cuantitativa	Continua	Intervalo	Índice de vegetación mejorado de dos bandas toma valores entre -1 y 1

Realizado por: Padilla S., Oscar R., 2020.

3.2 Etapa 2: Exploración de la data

En esta sección se analizó la data recopilada y se realizó un análisis exploratorio de los datos, atípicos, faltantes o ausentes (NA's), de manera de identificar una relación causal entre variables relevantes.

3.2.1 Análisis de datos faltantes (NA's)

La identificación de los datos ausentes se realizó en el software estadístico libre RStudio V1.3.1 bajo la plataforma de R V3.6.3, mediante la función *is.na()*, obteniendo los siguientes resultados:

Tabla 5-3: Número de datos faltantes

BD	DB Inicial	NA´s	BD Final
Provincia Chimborazo	644	53	591
Región Interandina	434	24	410

Realizado por: Padilla S., Oscar R., 2020

En la data de la provincia de Chimborazo se obtuvieron 53 valores NA, esto se debió a que en la imagen satelital mostró puntos con una alta presencia de nubes, y los valores de los índices espectrales no contenía información; por lo que se requiere eliminar esos valores; mientras que, en la data de la región Interandina, se encontraron 24 puntos en los que se presentaron complicaciones respecto a la generación de los índices.

3.2.2 Generación de las datas de estudio

Una vez definidos los tamaños de las datas, se generó 3 bases mediante información existente de dos mapas de carbono orgánico del suelo, el primero fue el Mapa Digital de Carbono Orgánico de Suelos del Ecuador elaborado por el ministerio de Agricultura y Ganadería (MAG); y el segundo mapa utilizado fue el digital Global Soil Organic Carbon Map (GSOCmap) elaborado por la Organización de las Naciones Unidas para la alimentación y la Agricultura (FAO)

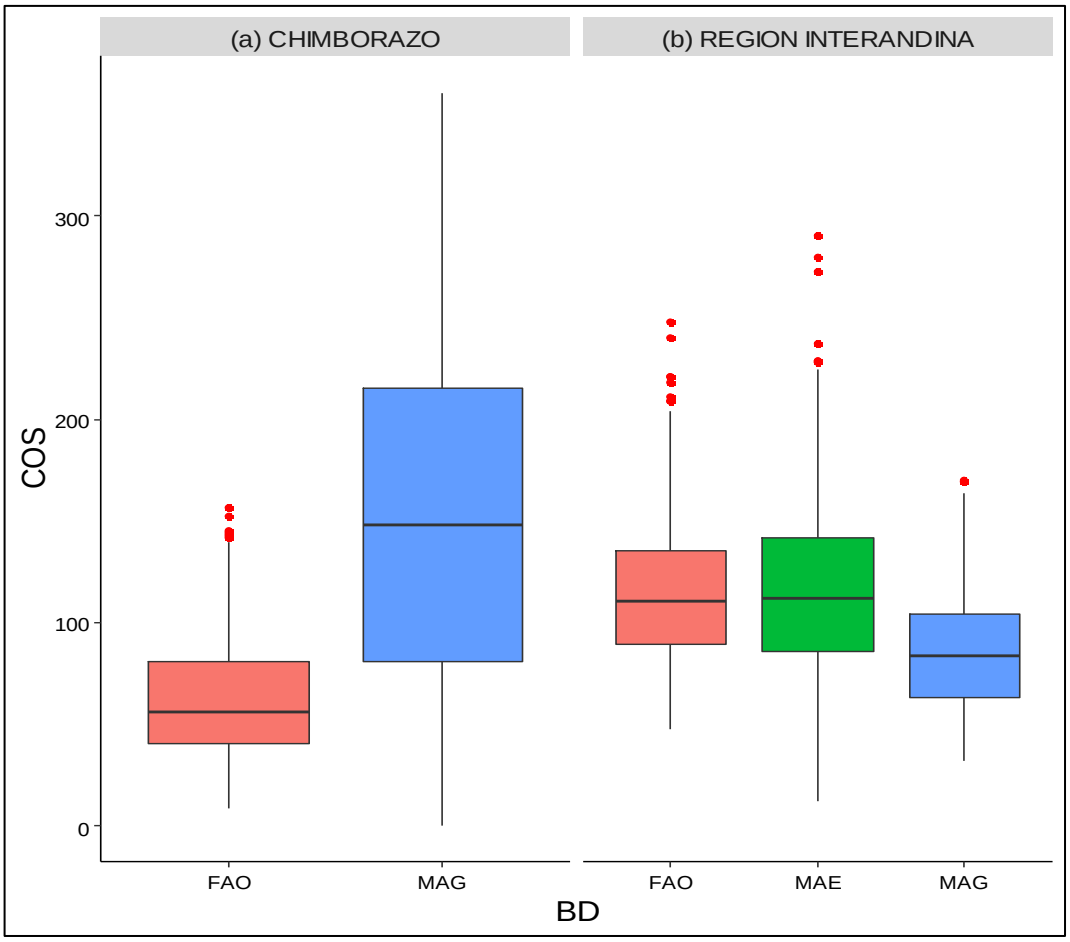
Para la provincia de Chimborazo se tiene una data proporcionada por el MAG y con las coordenadas de esos puntos se procedió a generar información de carbono orgánico del MAPA GSOCmap para obtener un data de nombre FAO, obteniendo así dos datas (MAG, FAO); para la Región Interandina se tiene una data proporcionada por el MAE y con las coordenadas de aquellos puntos de muestreo se extrajo información de COS de los mapas del MAG y FAO, obteniendo como resultado tres datas (MAE, MAG y FAO).

3.2.3 Análisis de datos atípicos

La existencia de los datos o posibles atípicos son un problema que alteran significativamente los resultados y las conclusiones que se derivan del análisis, en este caso se utilizó un diagrama de cajas con la función *boxplot()*, gráfico donde se observó la presencia de una mayor parte de los contenidos del COS como datos sospechosos, pero estos no se eliminaron debido a que son útiles para el estudio.

En la Gráfica 1-3, se visualizó los gráficos de boxplot de los datos de COS obtenidos en la provincia de Chimborazo (Grafico 1-3, (a)) en el que se observó datos sospechosos en la data

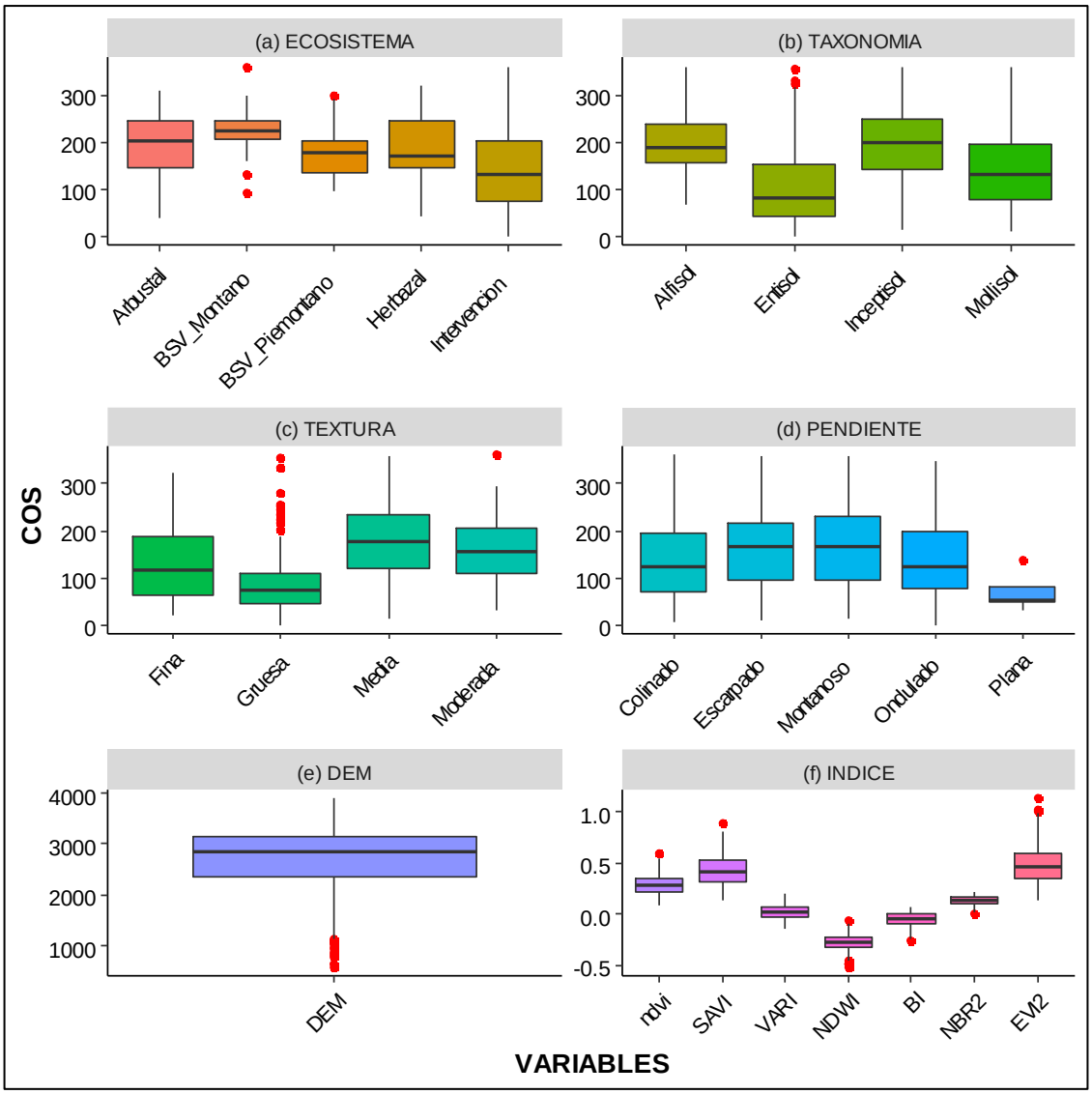
FAO a diferencia de los datos de COS obtenidas en la región interandina (Grafico 1-3, (b)); que se observó datos dudosos en las tres datas FAO, MAE y MAG.



Gráfica 1-3: Datos sospechosos en las datas de Chimborazo y Región Interandina.
Realizado por: Padilla S. Oscar R., 2020.

Provincia de Chimborazo

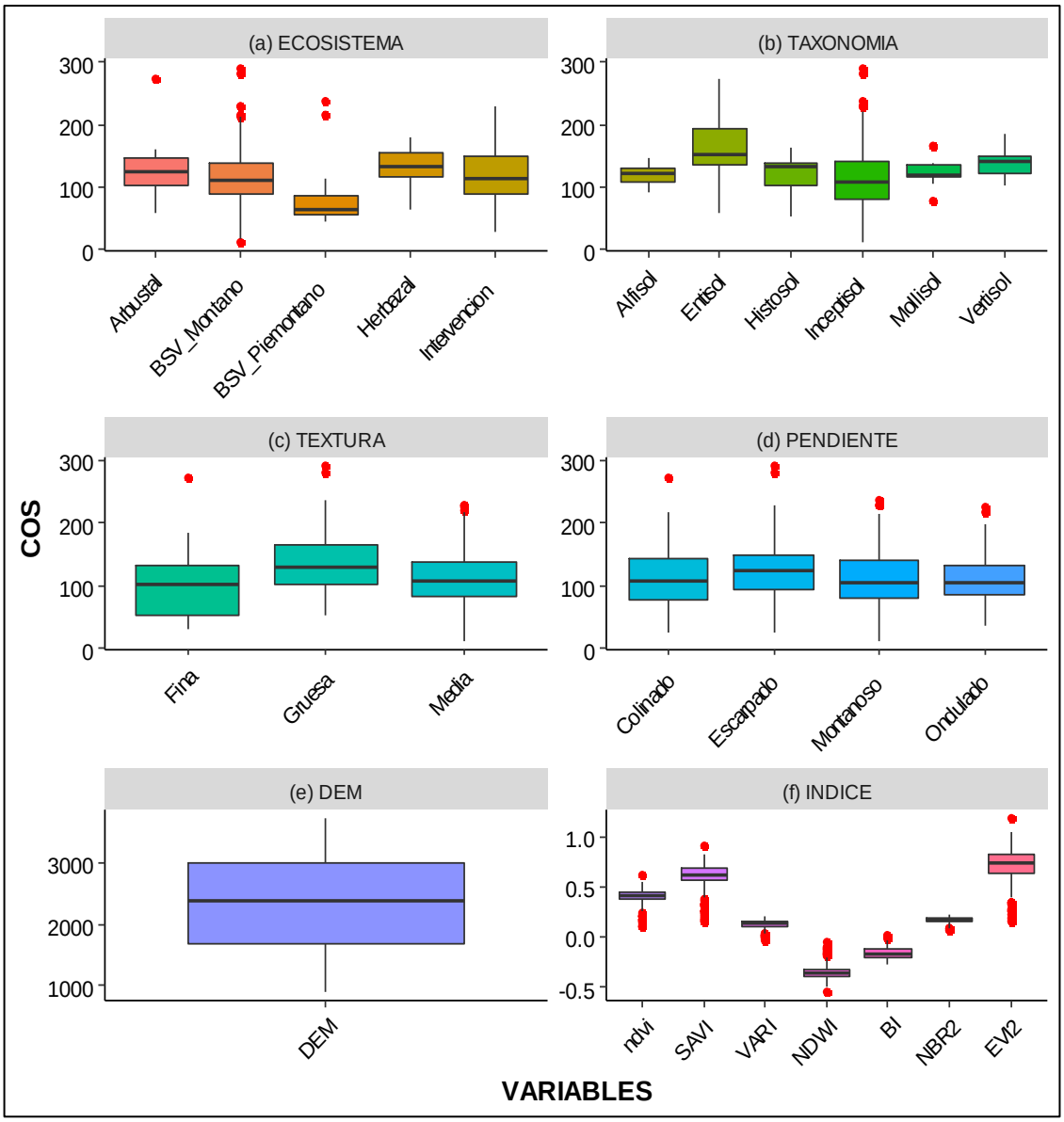
En los datos de contenido de carbono orgánico del suelo (COS) obtenidas por el MAG en la provincia de Chimborazo (Gráfico 2-3) se observó que existe la presencia de datos sospechosos en el ecosistema bosque siempre verde montano y bosque siempre verde piemontano (Gráfico 2-3, (a)), en la taxonomía entisol (Gráfico 2-3, (b)), en la textura del suelo existe datos sospechosos; en la textura moderada y con mayores datos sospechosos en la textura gruesa (Gráfico 2-3, (c)), en la pendiente plana (Grafico 2-3, (d)), en los niveles bajos de altitud sobre el nivel del mar se observó datos sospechosos (Gráfico 2-3, (e)), finalmente el índice VARI no presenta la existencia de éste tipo de datos (Gráfico 2-3, (f)).



Gráfica 2-3: Datos sospechosos en la data MAG de la provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020.

Callejón Interandino

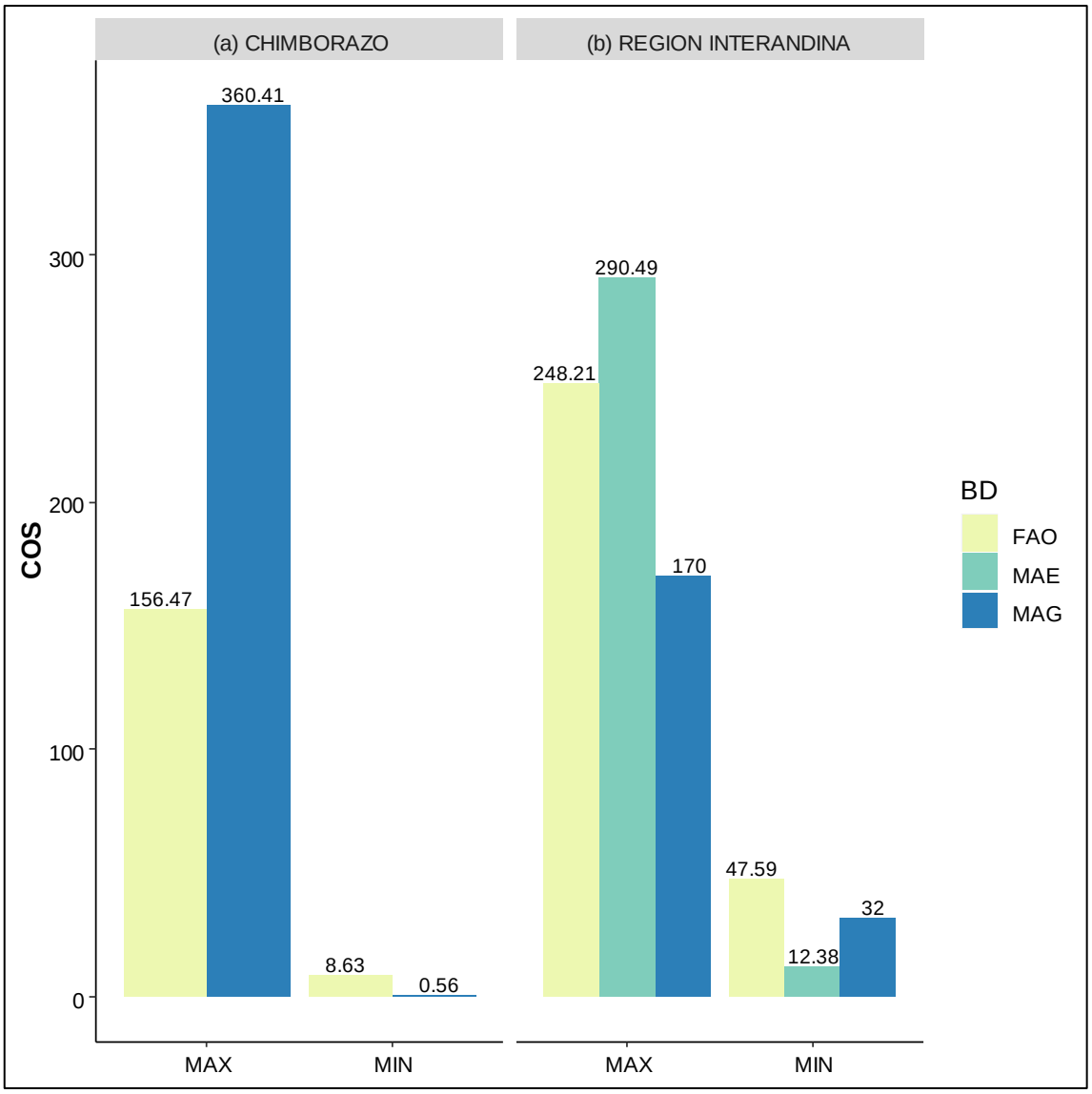
En los datos de contenido de carbono orgánico del suelo (COS) obtenidas por el MAE en la región interandina (Gráfico 3-3) se observó que existe la presencia de datos sospechosos en el ecosistema arbustal, bosque siempre verde montano y bosque siempre verde piemontano (Gráfico 3-3, (a)), en la taxonomía Inceptisol y Mollisol (Gráfico 3-3, (b)), en la textura del suelo fina, media y gruesa (Gráfico 3-3, (c)), además existe datos sospechosos en las pendientes colinado, escarpado, montanoso y ondulado (Grafico 3-3, (d)), de acuerdo a la altitud sobre el nivel del mar no existe datos sospechosos (Gráfico 3-3, (e)), finalmente observaron que todos los sietes índices presentan datos sospechosos (Gráfico 3-3, (f)).



Gráfica 3-3: Datos sospechosos en la data MAE de la Región Interandina.
Realizado por: Padilla S. Oscar R., 2020.

3.2.4 Análisis exploratorio de datos

El análisis exploratorio de datos se realizó de acuerdo a la provincia de Chimborazo y de la región interandina (Gráfico 4-3) en el que se observó que el mayor contenido de COS existente en el suelo en la provincia de Chimborazo (Gráfico 4-3, (a)) es de 360.42 Mg/ha y el menor COS es de 0.56 Mg/ha según los datos del MAG, mientras que la región interandina (Gráfico 4-3, (b)) se observó que el mayor contenido de carbono es de 290.49 Mg/ha y el menor COS es de 12.38 Mg/ha según los datos del MAE.

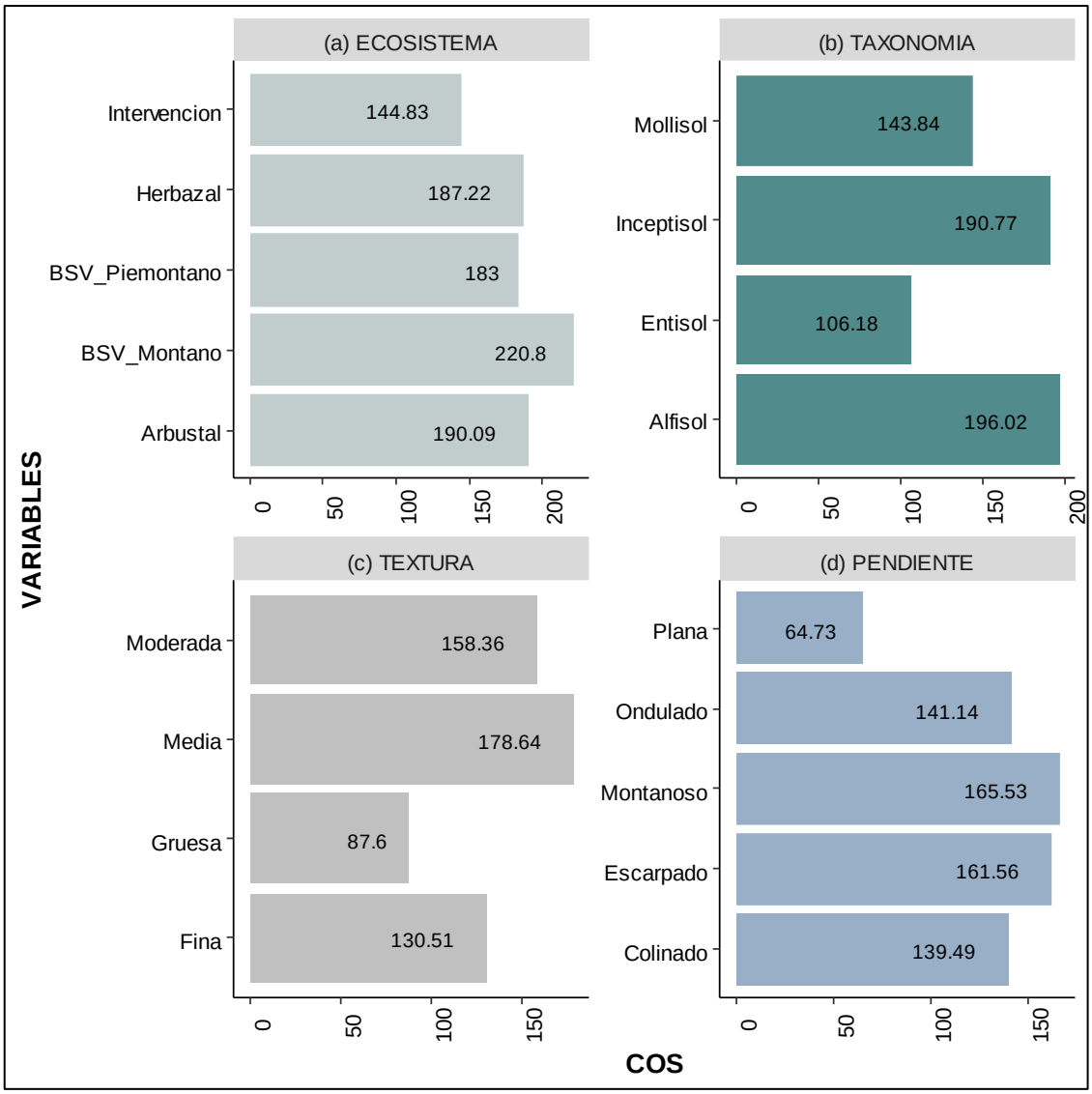


Gráfica 4-3: Contenidos de COS máximos y mínimos existentes en la provincia de Chimborazo y en la Región Interandina.

Realizado por: Padilla S. Oscar R., 2020.

Provincia de Chimborazo

De acuerdo a los datos del MAG (Gráfico 5-3) se observó que el mayor contenido de carbono orgánico del suelo (COS) se encuentra en el Ecosistema en Bosque siempre verde montano alto de la cordillera occidental de los Andes (Gráfico 5-3, (a)), en el suelo alfisol (Gráfico 5-3, (b)), textura media (Gráfico 5-3, (c)) y pendientes montañosa (Gráfico 5-3, (d)).



Gráfica 5-3: Contenidos de COS máximos existentes en los ecosistemas de la provincia de Chimborazo.

Realizado por: Padilla S. Oscar R., 2020.

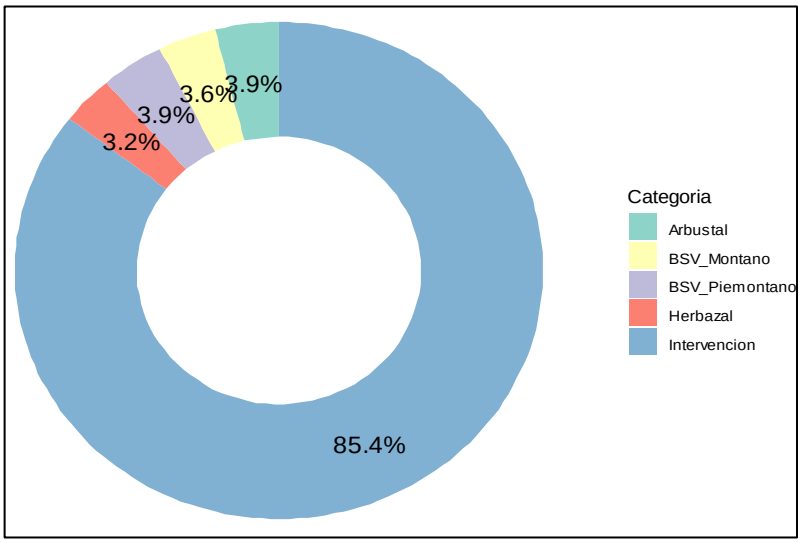
Distribución estadística de frecuencia de la variable Ecosistema

En la Tabla 6-3 y Gráfica 6-3, se observó que de acuerdo al tipo de ecosistema los contenidos de carbono edáfico de la provincia de Chimborazo están: 85.4 en intervención, 3.9% arbustal y bosque siempre verde piemontano (BSV_Piemontano), 3.6% en bosque siempre verde montano (BSV_Montano) y el 3.2% en Herbazal.

Tabla 6-3: Distribución estadística de frecuencia de la variable Ecosistema.

Tipos de Ecosistema	Frecuencia Absoluta (ni)	Frecuencia Relativa (fi)	Porcentaje Frecuencia relativa
Arbustal	23	0.039	3.9 %
BSV_Montano	21	0.036	3.6 %
BSV_Piemontano	23	0.039	3.9 %
Herbazal	19	0.032	3.2 %
Intervención	505	0.854	85.4 %

Realizado por: Padilla S., Oscar R., 2020.



Gráfica 6-3: Diagrama de pastel de la d.e.f. de la variable Ecosistema de la provincia de Chimborazo.

Realizado por: Padilla S. Oscar R., 2020.

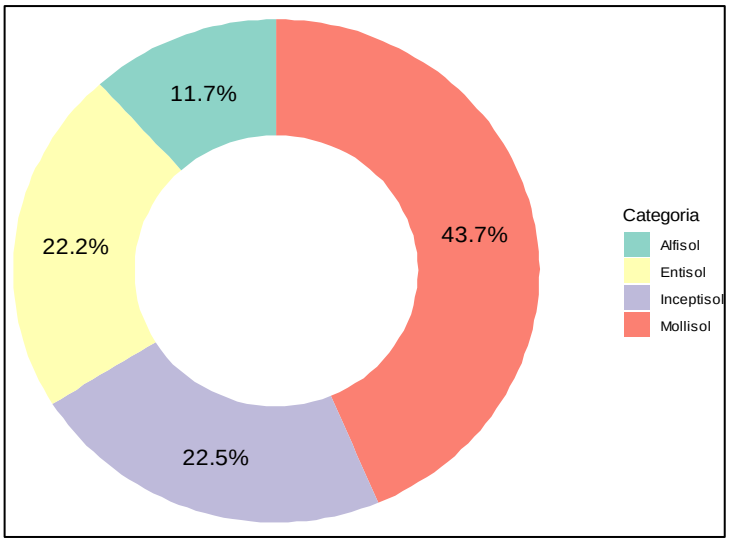
Distribución estadística de frecuencia de la variable Taxonomía

En la tabla 7-3 y Gráfica 7-3, se encontró que los contenidos de carbono edáfico de la provincia de Chimborazo según el tipo de suelo están: 43.7% en mollisol, 22.5% en inceptisol, 22.2% en entisol y el 11.7% en alfisol.

Tabla 7-3: Distribución estadística de frecuencia de la variable Taxonomía.

Tipos de suelo	Frecuencia Absoluta (ni)	Frecuencia Relativa (fi)	Porcentaje Frecuencia relativa
Alfisol	69	0.117	11.7 %
Entisol	131	0.222	22.2 %
Inceptisol	133	0.225	22.5 %
Mollisol	258	0.437	43.7 %

Realizado por: Padilla S., Oscar R., 2020.

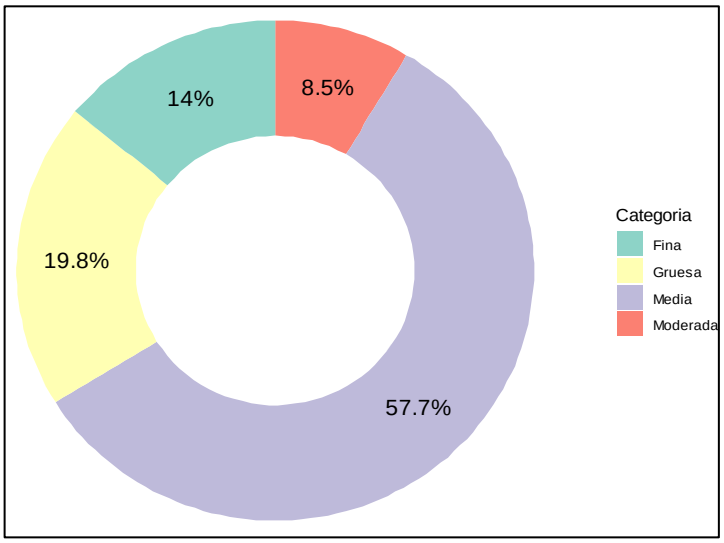


Gráfica 7-3: Diagrama de pastel de la d.e.f de la variable Taxonomía de la provincia de Chimborazo.

Realizado por: Padilla S. Oscar R., 2020.

Distribución estadística de frecuencia de la variable Textura

En la Tabla 8-3 y Gráfica 8-3, según los niveles de textura se encontró que los contenidos de carbono edáfico de la provincia de Chimborazo están: 57.7% en media, 19.8% en gruesa, 14% en fina y el 8.5% en moderada.



Gráfica 8-3: Diagrama de pastel de la d.e.f de la variable Textura de la provincia de Chimborazo.

Realizado por: Padilla S. Oscar R., 2020.

Tabla 8-3: Distribución estadística de frecuencia de la variable Textura.

Niveles de Textura	Frecuencia Absoluta (ni)	Frecuencia Relativa (fi)	Porcentaje Frecuencia relativa
Fina	83	0.140	14 %
Media	343	0.577	57.7 %
Moderada	50	0.085	8.5 %
Gruesa	117	0.198	19.8 %

Realizado por: Padilla S., Oscar R., 2020.

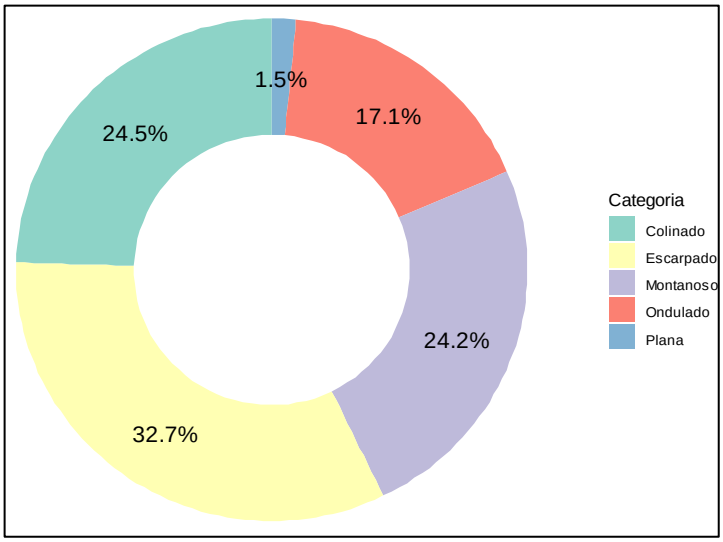
Distribución estadística de frecuencia de la variable Pendiente

En la Tabla 9-3 y Gráfica 9-3, dado el tipo de pendiente se encontró que los contenidos de carbono edáfico de la provincia de Chimborazo están: 32.7% en escarpado, 24.5% en colinado, 24.2% en montañoso y apenas el 1.5% en plana.

Tabla 9-3: Distribución estadística de frecuencia de la variable Pendiente.

Tipos de Pendiente	Frecuencia Absoluta (ni)	Frecuencia Relativa (fi)	Porcentaje Frecuencia relativa
Colinado	145	0.245	24.5 %
Escarpado	193	0.327	32.7 %
Montañoso	143	0.242	24.2 %
Ondulado	101	0.171	17.1 %
Plana	9	0.015	1.5 %

Realizado por: Padilla S., Oscar R., 2020.



Gráfica 9-3: Diagrama de pastel de la d.e.f de la variable Pendiente de la provincia de Chimborazo.

Realizado por: Padilla S. Oscar R., 2020.

Resumes Estadística Descriptiva de las variables cuantitativas

Provincia de Chimborazo

En la Tabla 10-3, se observó que el índice NDWI tiene una menor media, mediana y moda de -0.28, -0.27, -0.31 con menor dispersión entre los datos es decir que no están muy separadas de los valores de la distribución y la media aritmética.; y el índice EVI2 tienen una mayor media, mediana y moda de 0.48, 0.46, 0.39 con una mayor dispersión entre los datos es decir que no están muy separadas de los valores de la distribución y la media aritmética. Los índices NDVI, SAVI, VARI y EVI2 siguen una distribución asimétrica positiva y el NDWI, BI y NBR2 siguen una distribución asimétrica negativa es decir que se encuentra al lado izquierda de su media. El índice con menor coeficiente de asimetría (-0.46) es el BI y EVI2 con mayor coeficiente de asimetría (0.56). Además, todos los índices siguen una distribución leptocúrtica con mayor grado de concentración alrededor de su media. El índice con menor coeficiente de curtosis (-0.63) es el VARI y el NDWI con mayor coeficiente de curtosis (-0.04).

Tabla 10-3: Resumen estadístico de las variables cuantitativas provincia de Chimborazo.

MEDIDAS		**NDVI**	**SAVI**	**VARI**	**NDWI**	**BI**	**NBR2**	**EVI2**
Medidas de tendencia central	Media	0.28	0.43	0.02	-0.28	-0.06	0.13	0.48
	Mediana	0.28	0.42	0.02	-0.27	-0.05	0.13	0.46
	Moda	0.24	0.36	- 0.03	-0.31	-0.04	0.14	0.39
Medidas de posición	$Q_{(25\%)}$	0.21	0.32	-0.03	-0.32	-0.10	0.10	0.34
	$Q_{(50\%)}$	0.28	0.42	0.02	-0.28	-0.05	0.13	0.46
	$Q_{(75\%)}$	0.35	0.52	0.07	-0.23	0.00	0.16	0.6
Medias de dispersión	Desviación Típica	0.095	0.142	0.066	0.076	0.067	0.039	0.182
	Varianza	0.009	0.020	0.004	0.006	0.005	0.002	0.033
	Coeficiente de variación	0.19	0.28	-0.04	-0.35	-0.12	0.09	0.30
Medidas de Forma	Coeficiente de asimetría	0.35	0.37	0.09	-0.22	-0.46	-0.37	0.56
	Coeficiente de curtosis	-0.45	-0.43	-0.63	-0.04	-0.47	-0.40	-0.12

Realizado por: Padilla S., Oscar R., 2020.

Región Interandina

En la Tabla 11-3, se observó que el índice NDWI tiene una menor media, mediana y moda de -0.35, -0.37, -0,37 con una menor dispersión entre los datos es decir que no están muy separadas de los valores de la distribución y la media aritmética.; y el índice EVI2 tienen una mayor media, mediana y moda 0.71, 0.74, 0.74 con una mayor dispersión entre los datos es decir que no están muy separadas de los valores de la distribución y la media aritmética. Los índices NDWI y BI siguen una distribución asimétrica positiva y el NDVI, SAVI, VARI, NBR2 y EVI2 siguen una distribución asimétrica negativa es decir que se encuentra al lado izquierda de su media. VARI presentó un menor coeficiente de asimetría (-1.42) y NDWI un mayor coeficiente de asimetría (1.06). Además, todos los índices siguen una distribución leptocúrtica con mayor grado de concentración alrededor de su media. El índice con menor coeficiente de curtosis (0.44) es el BI y el VARI con mayor coeficiente de curtosis (2.46).

Tabla 11-3: Resumen estadístico de las variables cuantitativas provincia de Chimborazo.

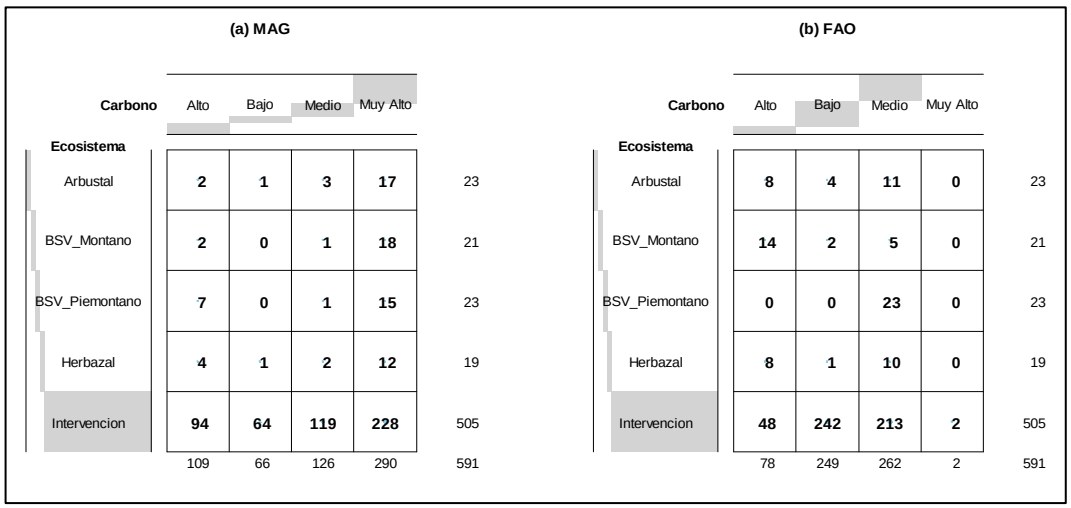
MEDIDAS		**NDVI**	**SAVI**	**VARI**	**NDWI**	**BI**	**NBR2**	**EVI2**
Medidas de tendencia central	Media	0.40	0.60	0.13	-0.35	-0.16	0.16	0.71
	Mediana	0.41	0.62	0.14	-0.37	-0.18	0.17	0.74
	Moda	0.42	0.63	0.15	-0.37	-0.19	0.17	0.74
Medidas de posición	$Q_{(25\%)}$	0.37	0.56	0.11	-0.40	-0.20	0.15	0.64
	$Q_{(50\%)}$	0.41	0.62	0.14	-0.37	-0.18	0.17	0.74
	$Q_{(75\%)}$	0.45	0.68	0.16	-0.32	-0.13	0.19	0.82
Medias de dispersión	Desviación Típica	0.082	0.122	0.041	0.074	0.053	0.031	0.166
	Varianza	0.007	0.015	0.002	0.006	0.003	0.001	0.027
	Coeficiente de variación	0.32	0.48	0.09	-0.43	-0.21	0.13	0.55
Medidas de Forma	Coeficiente de asimetría	-1.21	-1.21	-1.42	1.06	0.96	-0.96	-0.89
	Coeficiente de curtosis	-0.45	-0.43	-0.63	-0.04	-0.47	-0.40	-0.12

Realizado por: Padilla S., Oscar R., 2020.

3.2.5 Análisis de correspondencia de los datos

Análisis de Correspondencia del carbono en los tipos de Ecosistema

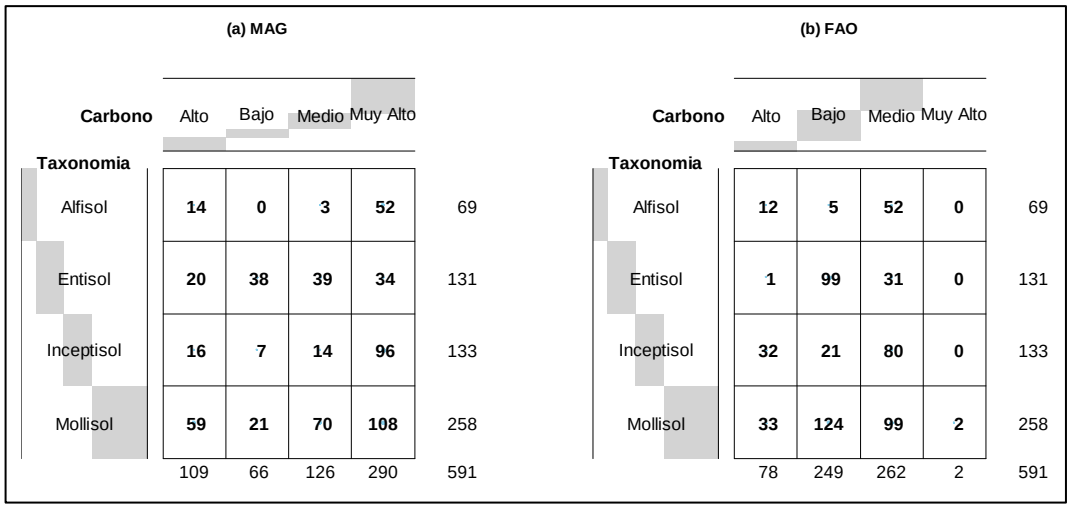
En la Gráfica 10-3, se observó que de un total de 591 muestras la mayor parte de COS se encontraron en el tipo de ecosistema intervención, los mismos que por el nivel de clasificación en los datos del MAG (Gráfica 10-3, (a)) 228 son Muy Altos, mientras que de los datos de la FAO (Gráfica 10-3, (b)) 242 son Bajos.



Gráfica 10-3: Carbono edáfico en los Ecosistemas de la Provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020.

Análisis de Correspondencia del carbono en los tipos de Suelo

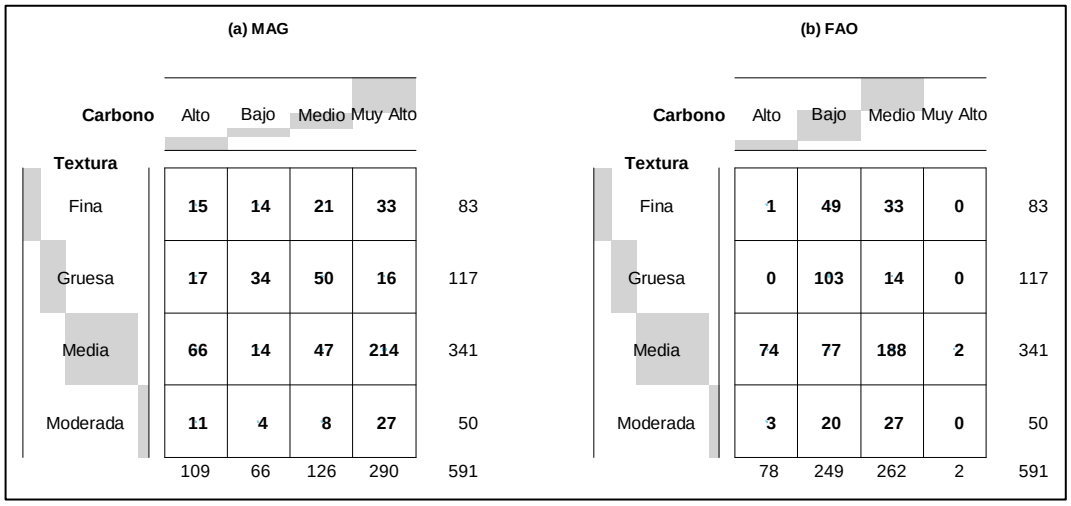
En la Gráfica 11-3, se observó que de un total de 591 muestras la mayor parte de COS se encontraron en el tipo de suelo Mollisol, los mismos que por el nivel de clasificación en los datos del MAG (Gráfica 11-3, (a)) 108 son Muy Altos, mientras que de los datos de la FAO (Gráfico 11-3, (b)) 124 son Bajos.



Gráfica 11-3: Carbono edáfico en los Suelos de la provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020.

Análisis de Correspondencia del carbono en la Textura del Suelo

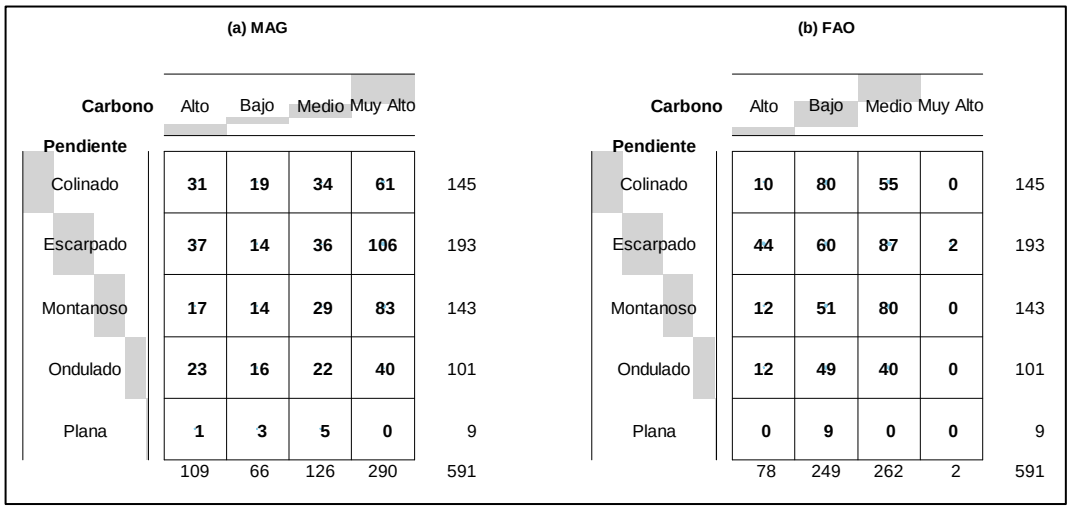
En la Gráfica 12-3, se observó que de un total de 591 muestras la mayor parte de COS se encontraron en la clase de textura media, los mismos que por el nivel de clasificación en los datos del MAG (Gráfica 12-3, (a)) 214 son muy altos, mientras que de los datos de la FAO (Gráfico 12-3, (b)) 188 son Medios.



Gráfica 12-3: Carbono edáfico en la Textura del suelo de la provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020.

Análisis de Correspondencia del carbono en los tipos de Pendiente

En la Gráfica 13-3, se observó que de un total de 591 muestras la mayor parte de COS se encontraron en la pendiente Escarpado, los mismos que por el nivel de clasificación en los datos del MAG (Gráfica 13-3, (a)) 106 son muy altos, mientras que de los datos de la FAO (Gráfico 13-3, (b)) 87 son Medios.



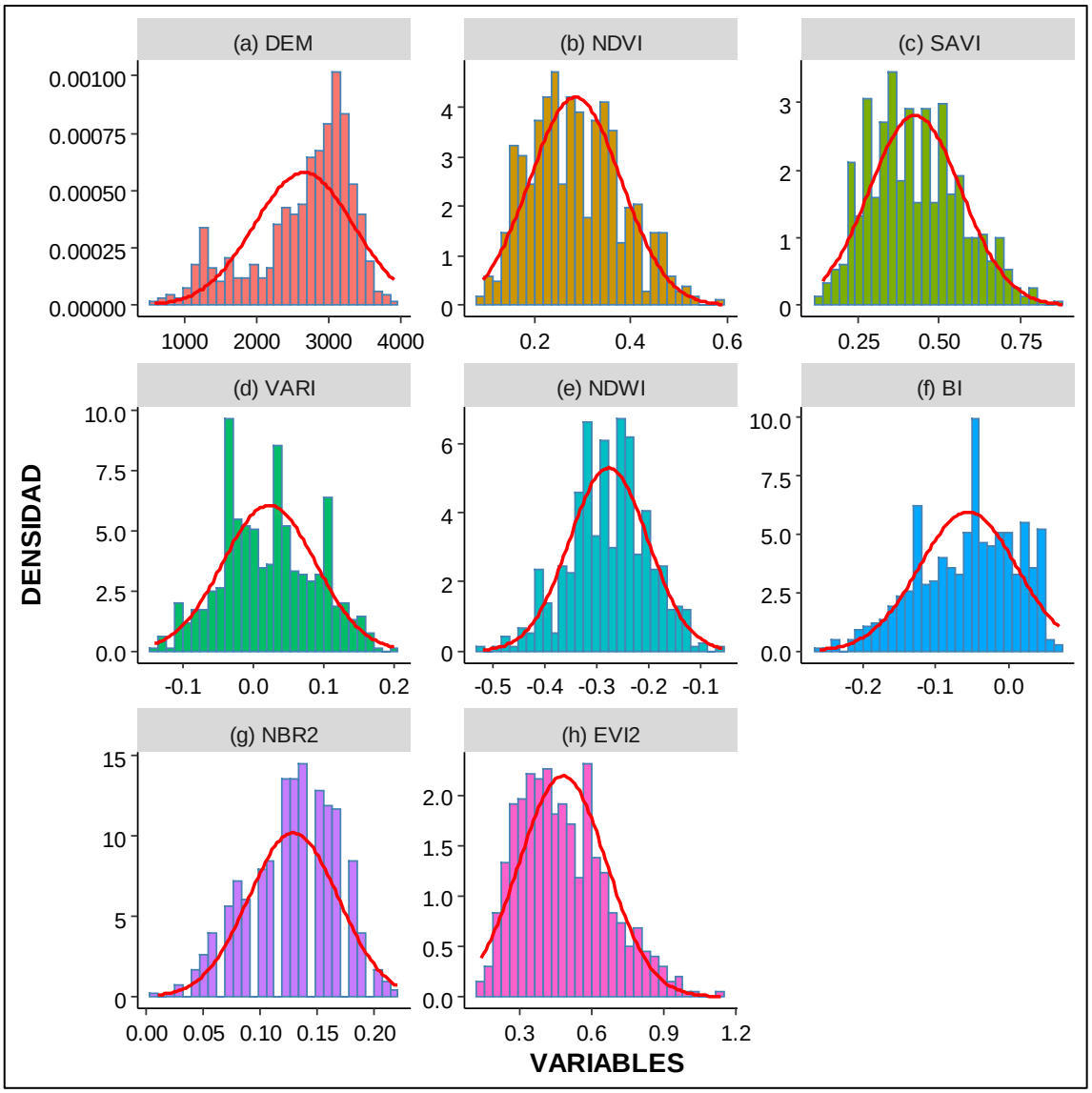
Gráfica 13-3: Carbono edáfico en las Pendientes de la provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020.

3.2.6 Pruebas de Normalidad

3.2.6.1 Métodos Gráficos

Provincia de Chimborazo

Los histogramas generados por las variables cuantitativas se presentaron en la Gráfica 14-3, los mismos indican que los datos no provienen de una distribución normal, debido a que no se ajusta a la curva de la distribución normal.

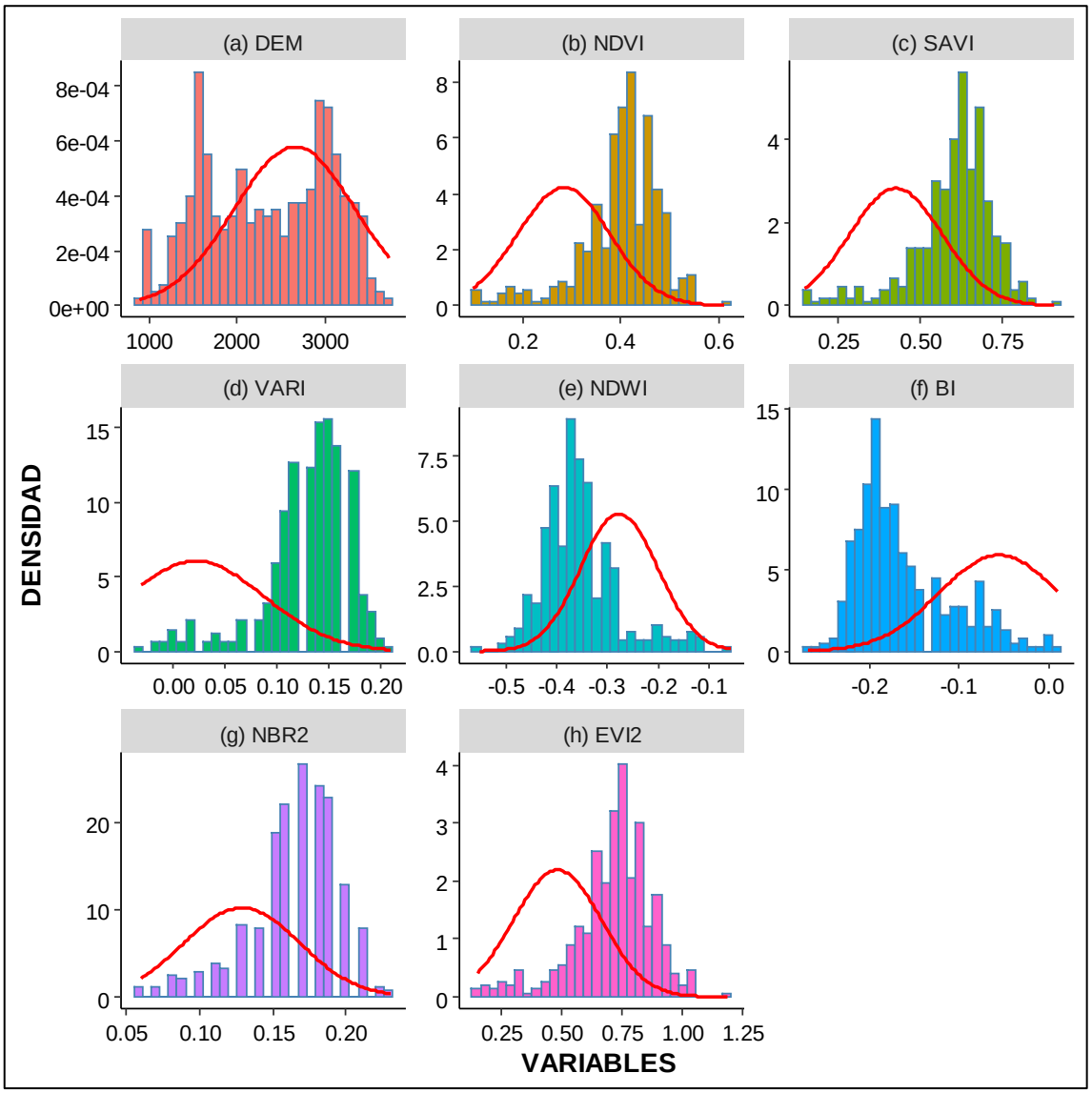


Gráfica 14-3: Histograma + Curva normal teórica de las variables cuantitativas de la provincia de Chimborazo.

Realizado por: Padilla S. Oscar R., 2020.

Región Interandina

Los histogramas generados por las variables cuantitativas se presentaron en la Gráfica 15-3, los mismo indican que los datos no provienen de una distribución normal, debido a que no se ajusta a la curva de la distribución normal.



Gráfica 15-3: Histograma + Curva normal teórica de las variables cuantitativas de la región interandina.

Realizado por: Padilla S. Oscar R., 2020.

3.2.6.2 Contraste de Hipótesis

Prueba de Hipótesis

i. Planteamiento de Hipótesis

H_0: Los datos provienen de una distribución normal

H_1: Los datos no provienen de una distribución normal

ii. Nivel de significancia

$\alpha = 0.05$

iii. Estadístico de Prueba

$$D = \max(|F_i - \emptyset * Z_i| \ - \ |F_{i-1} - \emptyset * Z_i| \ )$$

$$dode: F_i \rightarrow es\ la\ probabilidad\ y\ \emptyset Z_i \rightarrow es\ el\ percentil$$

iv. Regla de Decisión

Tabla 12-3: Prueba de Normalidad mediante contraste de Hipótesis.

DATOS	VARIABLES	ESTADISTICO	P.VAL	α	DECISION
CHIMBORAZO	DEM	0.1298	< 2.2e-16	0.05	Se Rechaza Ho
	NDVI	0.0695	3.627e-07	0.05	Se Rechaza Ho
	SAVI	0.0649	3.33e-06	0.05	Se Rechaza Ho
	VARI	0.077	7.032e-09	0.05	Se Rechaza Ho
	NDWI	0.0454	0.00551	0.05	Se Rechaza Ho
	BI	0.0847	7.381e-11	0.05	Se Rechaza Ho
	NBR2	0.09	2.335e-12	0.05	Se Rechaza Ho
	EVI2	0.0672	1.145e-06	0.05	Se Rechaza Ho
REGION INTERNDINA	DEM	0.1117	2.341e-13	0.05	Se Rechaza Ho
	NDVI	0.1262	< 2.2e-16	0.05	Se Rechaza Ho
	SAVI	0.1175	7.253e-15	0.05	Se Rechaza Ho
	VARI	0.1465	< 2.2e-16	0.05	Se Rechaza Ho
	NDWI	0.1309	< 2.2e-16	0.05	Se Rechaza Ho
	BI	0.1588	< 2.2e-16	0.05	Se Rechaza Ho
	NBR2	0.1346	< 2.2e-16	0.05	Se Rechaza Ho
	EVI2	0.0951	1.687e-09	0.05	Se Rechaza Ho

Realizado por: Padilla S., Oscar R., 2020

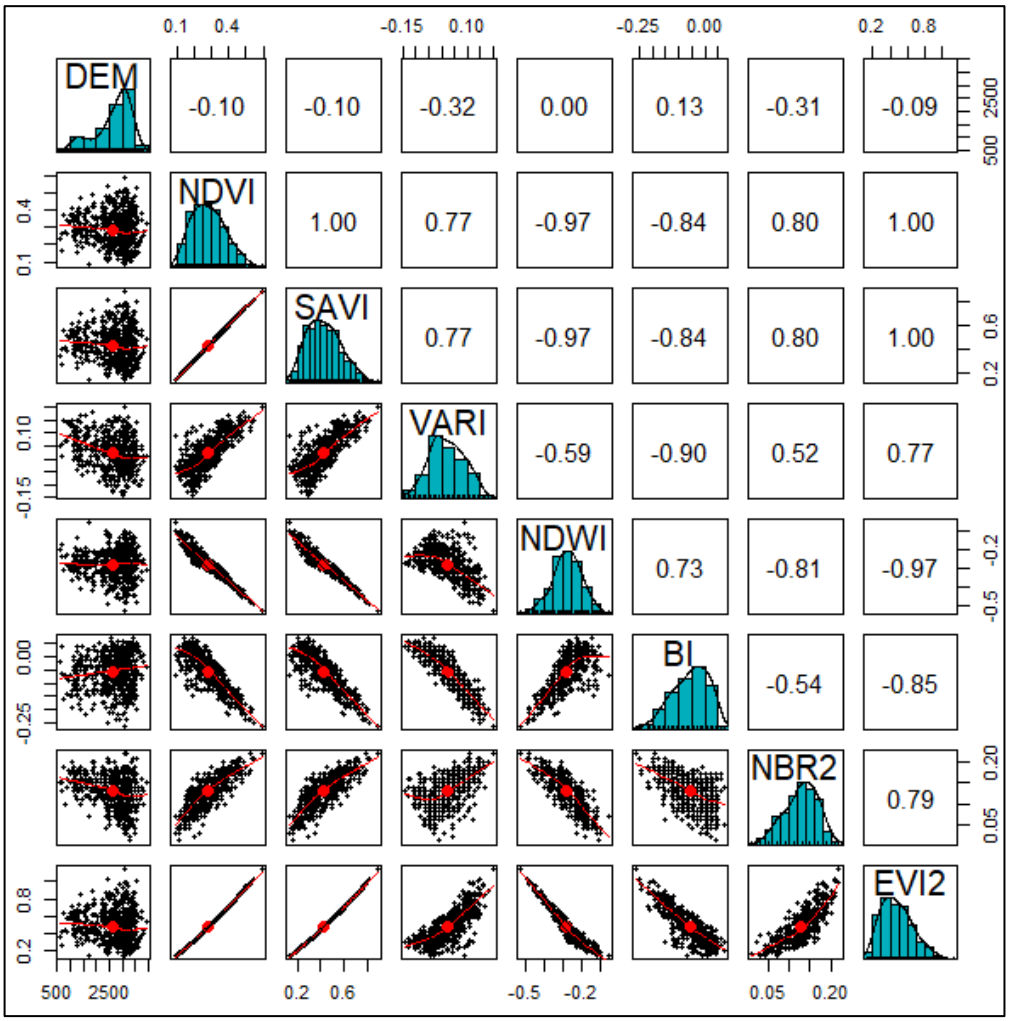
v. Conclusión

Mediante el estadístico de prueba de Kolmogorov Smirnov con la corrección de Lilliefors se obtuvo para todas las variables cuantitativas los valores p (p.val) son menores al nivel de significancia del 5% concluyendo así que la Altitud y los índices espectrales calculados para los datos de la provincia de Chimborazo y para datos de la región interandina no provienen de una distribución normal.

3.2.7 Coeficiente de Correlación

En este parte se procedió aplicar el coeficiente de correlación entre las variables independientes numéricas, con el objetivo de indicar cuan asociadas se encuentran dos variables entre sí. La correlación de *Karl Pearson* funciona bien con variables cuantitativas que tienen una distribución normal y debido a que en este estudio las variables no tienen una distribución no sería adecuado utilizar este método, pero en el libro de *Handbook of Biological Statistics* se menciona que sigue siendo bastante robusto a pesar de la falta de normalidad y es más sensible a los valores extremos, que los coeficientes de correlación de Spearman y Kendall que son utilizados en el caso no paramétrico pero con una previa transformación de los datos a rangos.

Provincia de Chimborazo

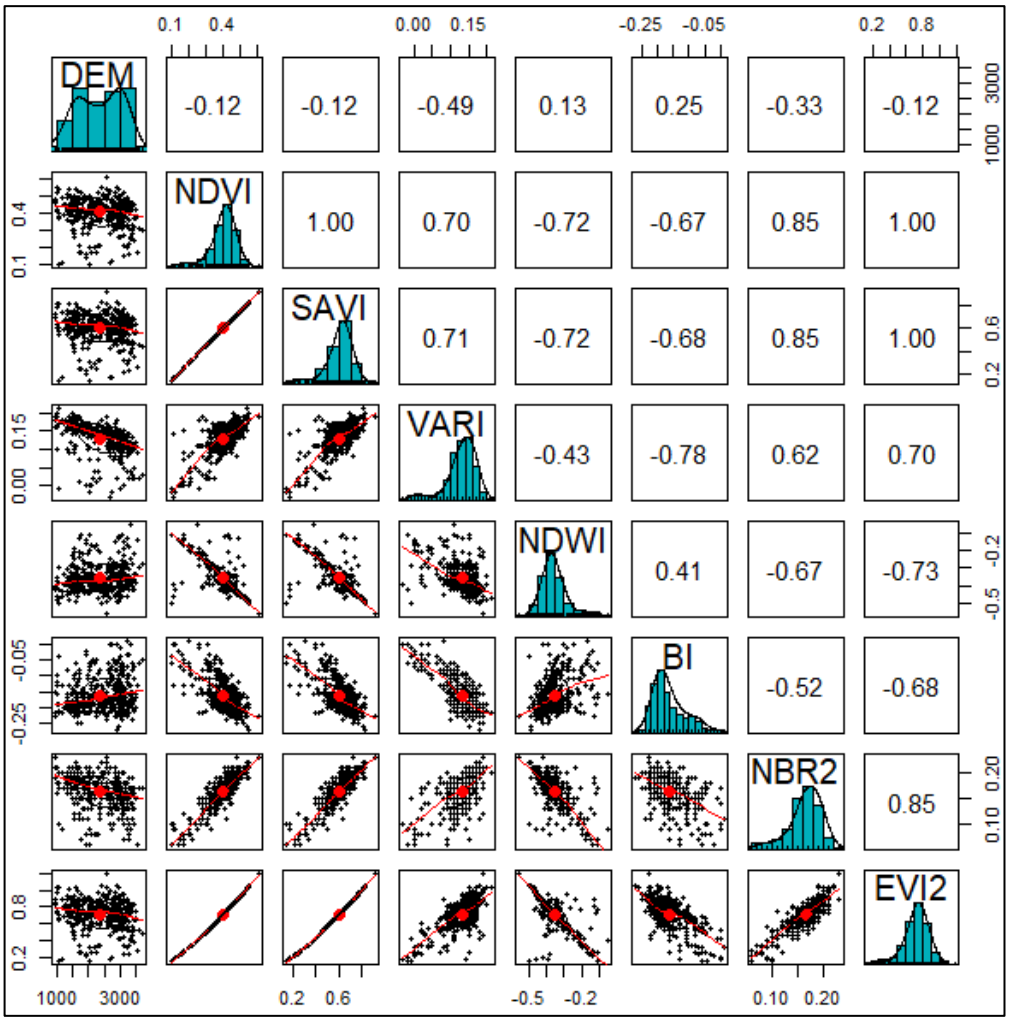


Gráfica 16-3: Correlación de Pearson de los datos de la provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020

En la Gráfica 16-3, se identificó a las variables NDVI, SAVI y EVI2 que tienen una relación lineal positiva perfecta; relación lineal positiva alta entre NDVI con SAVI y NBR2, SAVI con VARI y NBR2, VARI con EVI2, y entre NBR2 con EVI2; relación lineal negativa fuerte entre

NDVI con NDWI y BI, SAVI con NDWI y BI, VARI con BI, NDWI con NBR2 y EVI2, BI con EVI2; relación negativa baja de la variable DEM con NDVI, SAVI y EVI2; una relación negativa débil entre DEM con VARI y NBR2 y finalmente se observó una relación lineal positiva baja entre DEM con NDWI y BI.

Región interandina



Gráfica 17-3: Correlación de Pearson de los datos de la región interandina.
Realizado por: Padilla S. Oscar R., 2020.

En la Gráfica 17-3, se evidenció a las variables NDVI, SAVI y EVI2 que tienen una relación lineal positiva perfecta; relación lineal positiva fuerte entre NDVI, SAVI con NBR2, SAVI con VARI y NBR2, VARI con EVI2, y entre NBR2 con EVI2; relación lineal negativa fuerte entre NDVI con NDWI, SAVI con NDWI, VARI con BI, NDWI con EVI2; relación negativa moderada entre NDVI con BI, SAVI con BI, NDWI con NBR2, BI con EVI2; relación negativa débil de la variable DEM con VARI y NBR2, relación negativa baja entre DEM con NDVI, SAVI, y EVI2 y una relación lineal positiva baja entre DEM con NDWI y BI.

3.3 Etapa 3: Preprocesamiento de la Data

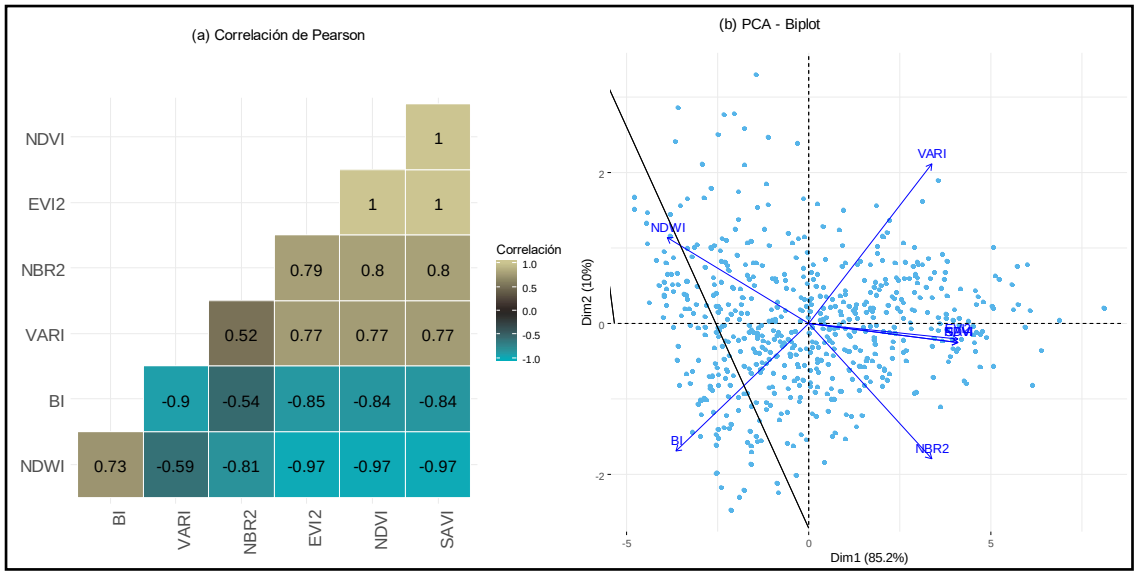
El foco principal de esta fase corresponde a la preparación y/o conversión de la data previamente seleccionada y comprendida para su uso; como entrada en la sección de modelamiento, para este proceso se trabajó mediante la utilización del software estadístico libre RStudio V1.3.1.

3.3.1 Prescindir de variables innecesarias.

3.3.1.1 Detección de problemas de multicolinealidad

Para identificar la existencia de relaciones lineales entre dos o más índices espectrales se procedió a verificar problemas de multicolinealidad.

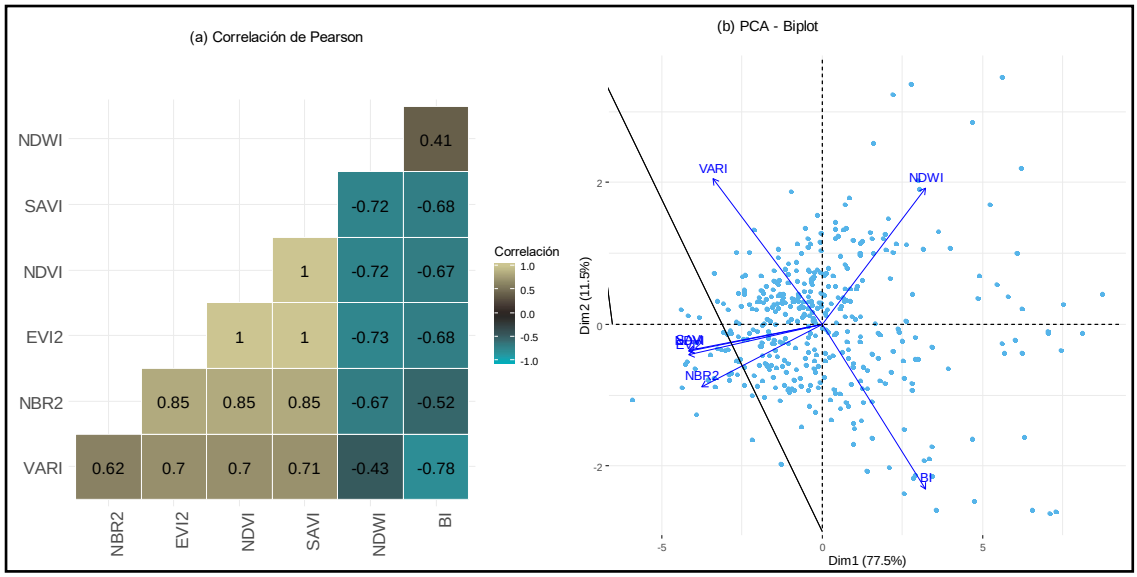
Provincia de Chimborazo



Gráfica 18-3: Problemas de multicolinealidad en los datos de la Provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020.

Los problemas de multicolinealidad para los datos de la provincia de Chimborazo se observó en la Gráfica 18-3, mediante la correlación de Pearson (Gráfica 25-3, (a)) existieron correlaciones de 1 en los índices espectrales NDVI, EVI2 y SAVI los mismos indican una relación perfecta directamente positiva, con lo que se evidenció claramente la existencia de relaciones lineales entre estos índices, es decir hay problemas de multicolinealidad perfecta; mediante la gráfica de *Biplot* generada por el análisis de componentes principales (Gráfica 25-3, (b)) se visualizó rotundamente que las flechas de los índices NDVI, SAVI y EVI2 se sobreponen es decir estos índices se encuentran en coordenadas similares en el plano lo que indica la existencia de patrones similares evidenciando así la existencia de problemas de multicolinealidad.

Región Interandina



Gráfica 19-3: Problemas de multicolinealidad en los datos de la región interandina.
Realizado por: Padilla S. Oscar R., 2020.

Los problemas de multicolinealidad para los datos de la región Interandina se observó en la Gráfica 19-3, dada la correlación de Pearson (Gráfica 26-3, (a)) se obtuvieron correlaciones de 1 en los índices espectrales NDVI, EVI2y SAVI que muestran una relación perfecta directamente positiva, con lo que se indicó claramente la existencia de relaciones lineales entre estos índices, es decir hay problemas de multicolinealidad perfecta; mediante la gráfica de *Biplot* generada por el análisis de componentes principales (Gráfica 26-3, (b)) se observó visiblemente que las flechas de los índices NDVI, SAVI y EVI2 se sobreponen es decir estos índices se encuentran en coordenadas similares en el plano lo que indica que tienen patrones similares de tal modo que existe problemas de multicolinealidad entre estos índices.

3.3.1.2 Solución de problemas de multicolinealidad

En esta parte se procedió a solucionar problemas de multicolinealidad determinando el factor de inflación de la varianza (VIF) a través de la regresión lineal, en el que se tomó en cuenta únicamente los índices con valores de VIF < 10, y todos aquellos que no cumplan esta condición no fueron considerados en el estudio.

El proceso para obtener todos aquellos índices que cumplían esta condición consistió en realizar los siguientes pasos:

- Generar una regresión lineal con los siete índices espectrales
- Determinar el VIF.

- Verificar la condición VIF < 10 para cada índice
- Si hay valores de VIF mayores a 10, descartar el índice que tenga el mayor valor de VIF
- Volver a calcular la regresión lineal omitiendo el índice que tenga el mayor VIF.
- Determinar el VIF y verificar la condición si no se cumple se volver a calcular la regresión con el índice que tenga el mayor VIF.

Provincia de Chimborazo

Tabla 13-3: Valores VIF proceso 1, data Chimborazo.

Índice	VIF	VIF < 10	Decisión	Índice para eliminar
NDVI	731.22	F	No Cumple	
SAVI	1006.67	F	No Cumple	x
VARI	60.97	F	No Cumple	
NDWI	357.34	F	No Cumple	
BI	8.34	V	Si Cumple	
NBR2	3.46	V	Si Cumple	
EVI2	404.88	F	No Cumple	

Realizado por: Padilla S. Oscar R., 2020.

La Tabla 13-3, mostró los valores del VIF a partir de la regresión lineal generados con todos los índices, se observó que el BI y NBR2 cumplen con la condición requerida por lo que se procede a eliminar el SAVI para continuar con el procedimiento.

Tabla 14-3: Valores VIF proceso 2, data Chimborazo.

Índice	VIF	VIF < 10	Decisión	Índice para eliminar
NDVI	360.98	F	No Cumple	
VARI	56.76	F	No Cumple	
NDWI	330.76	F	No Cumple	
BI	8.31	V	Si Cumple	
NBR2	3.45	V	Si Cumple	
EVI2	369.11	F	No Cumple	x

Realizado por: Padilla S. Oscar R., 2020.

En la Tabla 14-3, se observó los valores del VIF a partir de la regresión lineal generados con seis índices, en el que se notó que los valores del VIF de los índices bajaron, pero aún no cumplía con la condición por lo que se procedió a eliminar el EVI2 para continuar con el procedimiento.

Tabla 15-3: Valores VIF proceso 3, data Chimborazo.

Índice	VIF	VIF < 10	Decisión	Índice para eliminar
NDVI	332.15	F	No Cumple	x
VARI	39.32	F	No Cumple	
NDWI	211.58	F	No Cumple	
BI	8.23	V	Si Cumple	
NBR2	3.38	V	Si Cumple	

Realizado por: Padilla S. Oscar R., 2020.

En la Tabla 15-3, se observó los valores del VIF a partir de la regresión lineal generados con cinco índices, en el que se evidenció que los valores VIF de los índices bajaron, pero siguieron sin cumplir la condición, por lo que se procedió a eliminar el NDVI para continuar con el procedimiento.

Tabla 16-3: Valores VIF proceso 4, data Chimborazo.

Índice	VIF	VIF < 10	Decisión	Índice para eliminar
VARI	5.89	V	Si Cumple	
NDWI	5.19	V	Si Cumple	
BI	8.18	V	Si Cumple	
NBR2	3.38	V	Si Cumple	

Realizado por: Padilla S. Oscar R., 2020.

En la Tabla 16-3, se mostró los valores del VIF a partir de la regresión lineal generados con cuatro índices, los cuales presentó valores VIF de los índices menores a 10 con lo que indicó que el problema de multicolinealidad había sido resuelto debido a que los índices que quedaron cumplieron con la condición.

Región Interandina

En la Tabla 17-3, se observó los valores del VIF de los índices VARI, NDWI, BI y NBR2 generados a partir de la regresión lineal generados con cumplieron con la condición requerida por lo que se procede a eliminar el índice SAVI para continuar con el procedimiento.

Tabla 17-3: Valores VIF proceso 1, data región Interandina.

Índice	VIF	VIF < 10	Decisión	Índice para eliminar
NDVI	599.77	F	No Cumple	
SAVI	679.16	F	No Cumple	x
VARI	3.17	V	Si Cumple	
NDWI	2.42	V	Si Cumple	
BI	2.95	V	Si Cumple	
NBR2	4.02	V	Si Cumple	
EVI2	224.27	F	No Cumple	

Realizado por: Padilla S. Oscar R., 2020.

En la Tabla 18-3, se identificó el índice EVI2 con unos valores del VIF mayor por lo que no cumple con la condición requerida y se procedió a eliminar para continuar con el procedimiento.

Tabla 18-3: Valores VIG proceso 2, data región Interandina.

Índice	VIF	VIF < 10	Decisión	Índice para eliminar
NDVI	187.36	F	No Cumple	
VARI	3.15	V	Si Cumple	
NDWI	2.41	V	Si Cumple	
BI	2.95	V	Si Cumple	
NBR2	4.00	V	Si Cumple	
EVI2	190.69	F	No Cumple	x

Realizado por: Padilla S. Oscar R., 2020.

En la Tabla 19-3, se visualizó los valores del VIF a partir de la regresión lineal generados con cinco índices, son menores a 10 y se indicó que el problema de multicolinealidad fue resulto debido a que los índices que quedaron cumplen con la condición.

Tabla 19-3: Valores VIF proceso 3, data región Interandina.

Índice	VIF	VIF < 10	Decisión	Índice para eliminar
NDVI	6.10	V	Si Cumple	
VARI	3.10	V	Si Cumple	
NDWI	2.17	V	Si Cumple	
BI	2.89	V	Si Cumple	
NBR2	4.00	V	Si Cumple	

Realizado por: Padilla S. Oscar R., 2020.

3.3.2 Conversión variable objetivo en Categoría

Debido a que se pretendió obtener la precisión de cada uno de los algoritmos empleados para determinar el árbol de decisión de tipo clasificación, en este estudio, la variable dependiente de tipo continua fue categorizada. Y para clasificar los niveles del COS se realizó una revisión del reporte de valores de concentración de carbono edáfico en el suelo, con el fin de obtener una referencia general de lo que se considera como valores altos y bajos. Se obtuvo que no existe una referencia en donde se establezca dicha jerarquización de manera general; por lo que, para fines de este trabajo, se consideró una clasificación generada por Gilberto Vela Correa, Jorge López Blanco María de Lourdes y Rodríguez Gamiño; quienes en un estudio sobre Niveles de carbono orgánico total en el suelo de conservación del Distrito Federal centro de México, consideraron límites de los intervalos, en función de los valores máximos y mínimos de concentración de carbono obtenidos de los análisis de suelo definiendo 4 niveles (Vela Correa et al., 2012, p. 23). Muy Alto (> 150 Mg/ha), Alto (100 - 150 Mg/ha), Medio (50 - 100 Mg/ha) y Bajo (< 50 Mg/ha)

3.3.3 Organizar la data

Mediante un análisis y preprocesamiento de la data, de un total de 12 variables explicativas, se trabajó con 9 en la data de la provincia de Chimborazo y con 10 en la data de la región interandina, esto se debió a los altos valores de multicolinealidad previamente calculados.

Tabla 20-3: Variables para los datos de la provincia de Chimborazo.

N°	Variables	Valores que puede tomar la variable
1	Carbono orgánico del suelo (COS)	Bajo, Medio, Alto, Muy Alto
2	Ecosistema	Intervención, BSV_Piemontano, BSV_Montano, Arbustal y Herbazal
3	Taxonomía	Alfisol, Entisol, Inceptisol, Mollisol
4	Textura	Fina. Media Moderada y Gruesa
5	Pendiente	Colinado, Escarpado, Montañoso, Ondulado y Plana
6	DEM	Altura sobre el nivel el mar en metros
7	VARI	Índice de resistencia atmosféricamente visible con valores [-1, 1]
8	NDWI	Índice diferencial de agua normalizada coa valores [-1, 1]
9	BI	Índice de área calcinada con valores [-1, 1]
10	NBR2	Índice normalizado de Áreas Quemadas 2 con valores [-1 y 1]

Realizado por: Padilla S., Oscar R., 2020.

Tabla 21-3: Variables para los datos de la región interandina.

N°	Variables	Valores que puede tomar la variable
1	Carbono orgánico del suelo (COS)	Bajo, Medio, Alto, Muy Alto
2	Ecosistema	Intervención, BSV_Piemontano, BSV_Montano, Arbustal y Herbazal
3	Taxonomía	Alfisol, Entisol, Histosol, Inceptisol, Mollisol, Vertisol
4	Textura	Fina. Media y Gruesa
5	Pendiente	Colinado, Escarpado, Montanoso y Ondulado
6	DEM	Altura sobre el nivel el mar en metros
7	VARI	Índice de resistencia atmosféricamente visible con valores [-1, 1]
8	NDWI	Índice diferencial de agua normalizada con valores [-1, 1]
9	BI	Índice de área calcinada con valores [-1, 1]
10	NBR2	Índice normalizado de Áreas Quemadas 2 con valores [-1, 1]
11	EVI2	Índice de vegetación mejorado de dos bandas con valores [-1, 1]

Realizado por: Padilla S., Oscar R., 2020.

3.3.4 Dividir en conjuntos de entrenamiento y test

Para dividir la data en conjuntos de entrenamiento y test se utilizó el criterio de Houlduot el cual indica una repartición de 70% para el conjunto de entrenamiento y 30% para el conjunto de prueba, en la Tabla 22-3, se indicó el tamaño de unidades de cada una de las datas.

Tabla 22-3: Datas de entrenamiento y prueba.

BD	Total	Data Entrenamiento	Data Prueba
MAG	591	416	175
MAE	410	288	122

Realizado por: Padilla S., Oscar R., 2020.

3.4 Etapa 4: Modelamiento

Aplicación de los algoritmos de clasificación

Finalmente, con cada uno de los algoritmos se desarrolló el modelo a partir de las variables predictoras, las mismas que permiten predecir la variable objetivo del modelo. El modelo se construyó a partir de dos conjuntos de datos, uno para el entrenamiento y otro para la prueba. En el conjunto de datos de entrenamiento se utilizó el 70% del total de datos, mientras que para el conjunto de datos de prueba se empleó el 30% restante, estos porcentajes son utilizados de manera aleatoria siendo un total de 591 datos (entrenamiento + prueba) con observaciones de 10 variables predictoras.

Luego del proceso realizado con los datos correspondientes en el software RStudio, se obtuvieron los siguientes resultados en cada uno de los algoritmos seleccionados.

3.4.1 Bondad del Clasificador

Tabla 23-3: Resultados de los algoritmos mediante la Bondad del Clasificador

Resultados	Algoritmo C5.0		Algoritmo SVM		Algoritmo CART	
	Precisión	**Kappa**	**Precisión**	**Kappa**	**Precisión**	**Kappa**
Árbol 1	57.72	0.3434	59.44	0.340	63.09	0.4383
Árbol 2	56.16	0.2445	57.43	0.3701	62.75	0.3964
Árbol 3	57.42	0.3351	61.34	0.3217	60.39	0.3244
Árbol 4	53.72	0.3323	60.16	0.3651	59.42	0.3798
Árbol 5	56.25	0.3165	58.51	0.3478	62.84	0.4124
Árbol 6	62.58	0.4186	61.44	0.4014	60.69	0.3247
Árbol 7	58.28	0.3343	60.39	0.3084	59.26	0.4054
Árbol 8	54.39	0.3150	57.15	0.3993	64.14	0.3480
Árbol 9	56.58	0.3066	59.52	0.3682	61.78	0.3887
Árbol 10	57.33	0.3206	61.58	0.3465	57.75	0.3768

Realizado por: Padilla S. Oscar R., 2020.

A partir de los resultados obtenidos sobre los rendimientos de los modelos se observó en la Tabla 24-3 que todos los algoritmos obtienen entre 56% y 62% de instancias correctamente clasificados, por lo tanto, no se puede confiar o darles credibilidad a estos resultados, puesto que apenas un

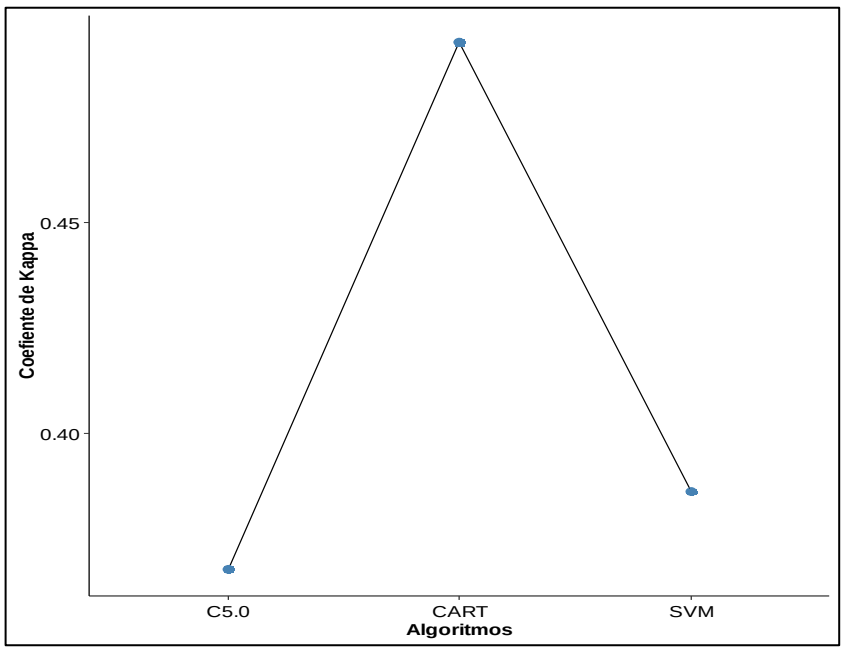
61% de la data presenta una clasificación como máximo, por lo que se pudo determinar que no se ha encontrado un algoritmo adecuado para este tipo de datos. Sin embargo, para la distinción entre cada uno, se tomó en cuenta la validación del coeficiente de kappa el cual indica que, si se tiene un valor cercano a 1 las variables están fuertemente relacionadas, y para valores próximos a -1 existe una pobre relación entre las variables. En el Grafico se aprecia el valor del coeficiente de Kappa de los tres algoritmos.

Tabla 24-3: Comparación de los algoritmos mediante la bondad del clasificador.

ALGORITMOS	PRESICION	ERROR	KAPPA	IC: 95%
C5.0	56.53	43.47	0.3267	(0.45, 0.61)
SVM	59.70	40.30	0.3567	(0.47, 0.64)
CART	61.21	38.79	0.3795	(0.54, 0.70)

Realizado por: Padilla S. Oscar R., 2020

En la Gráfica 20-3, se observó, el mejor rendimiento del algoritmo CART que una precisión de 61.21%, un índice de concordancia moderada de 0.38 y con un error de predicción de 38.79%, además su intervalo de precisión se encuentra entre [0.54 – 0.70].



Gráfica 20-3: Coeficiente de Kappa de cada los algoritmos mediante la bondad del clasificador.

Realizado por: Padilla S. Oscar R., 2020.

3.4.2 Validación cruzada

Tabla 25-3: Resultados de los algoritmos mediante una validación cruzada

Resultados	Algoritmo C5.0		Algoritmo SVM		Algoritmo CART	
	Precisión	**Kappa**	**Precisión**	**Kappa**	**Precisión**	**Kappa**
Árbol 1	59.52	0.3670	63.41	0.3600	66.67	0.4586
Árbol 2	57.50	0.3577	61.46	0.3460	69.05	0.5125
Árbol 3	58.45	0.3615	61.90	0.3313	68.29	0.5009
Árbol 4	57.14	0.3761	63.85	0.3778	64.29	0.4175
Árbol 5	60.30	0.3918	64.12	0.4311	67.44	0.4724
Árbol 6	59.85	0.3854	62.44	0.4154	68.54	0.4980
Árbol 7	54.76	0.3798	63.41	0.3825	64.29	0.4261
Árbol 8	56.10	0.3016	60.16	0.4110	65.12	0.4277
Árbol 9	60.98	0.3887	62.41	0.3776	59.74	0.5625
Árbol 10	61.00	0.3674	63.56	0.4305	63.80	0.6513

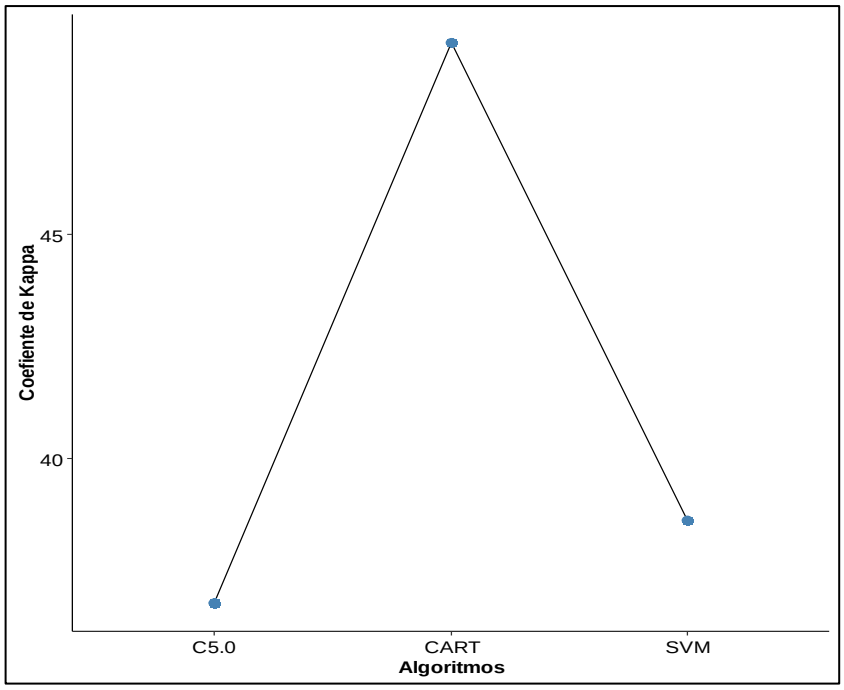
Realizado por: Padilla S. Oscar R., 2020.

A partir de los resultados obtenidos mediante una validación cruzada en la Tabla 26-3, se observó que todos los algoritmos obtienen entre 58% y 66% de instancias correctamente clasificados, sin embargo, aún no se puede confiar en estos resultados, puesto que nos está dando un 66% de clasificación como máximo, lo cual nos indica que todavía no hemos encontrado un algoritmo adecuado. Por lo tanto, considerando la validación del coeficiente de kappa que se puede observar en la Grafica nos indica claramente que el algoritmo que presenta un mejor rendimiento para los datos es el CART que además tiene su intervalo de precisión se encuentra entre [0.54 – 0.70], también se corroboró mediante la Grafica 21-3.

Tabla 26-3: Comparación de los algoritmos aplicados una validación cruzada.

ALGORITMOS	PRECISION	ERROR	KAPPA	IC: 95%
C5.0	58.56	41.44	0.3677	(0.45, 0.61)
SVM	62.67	37.33	0.3863	(0.47, 0.64)
CART	65.72	34.28	0.4928	(0.54, 0.70)

Realizado por: Padilla S. Oscar R., 2020.



Gráfica 21-3: Coeficiente de Kappa de cada los algoritmos con una validación cruzada.

Realizado por: Padilla S. Oscar R., 2020.

3.4.3 Comparación de modelos

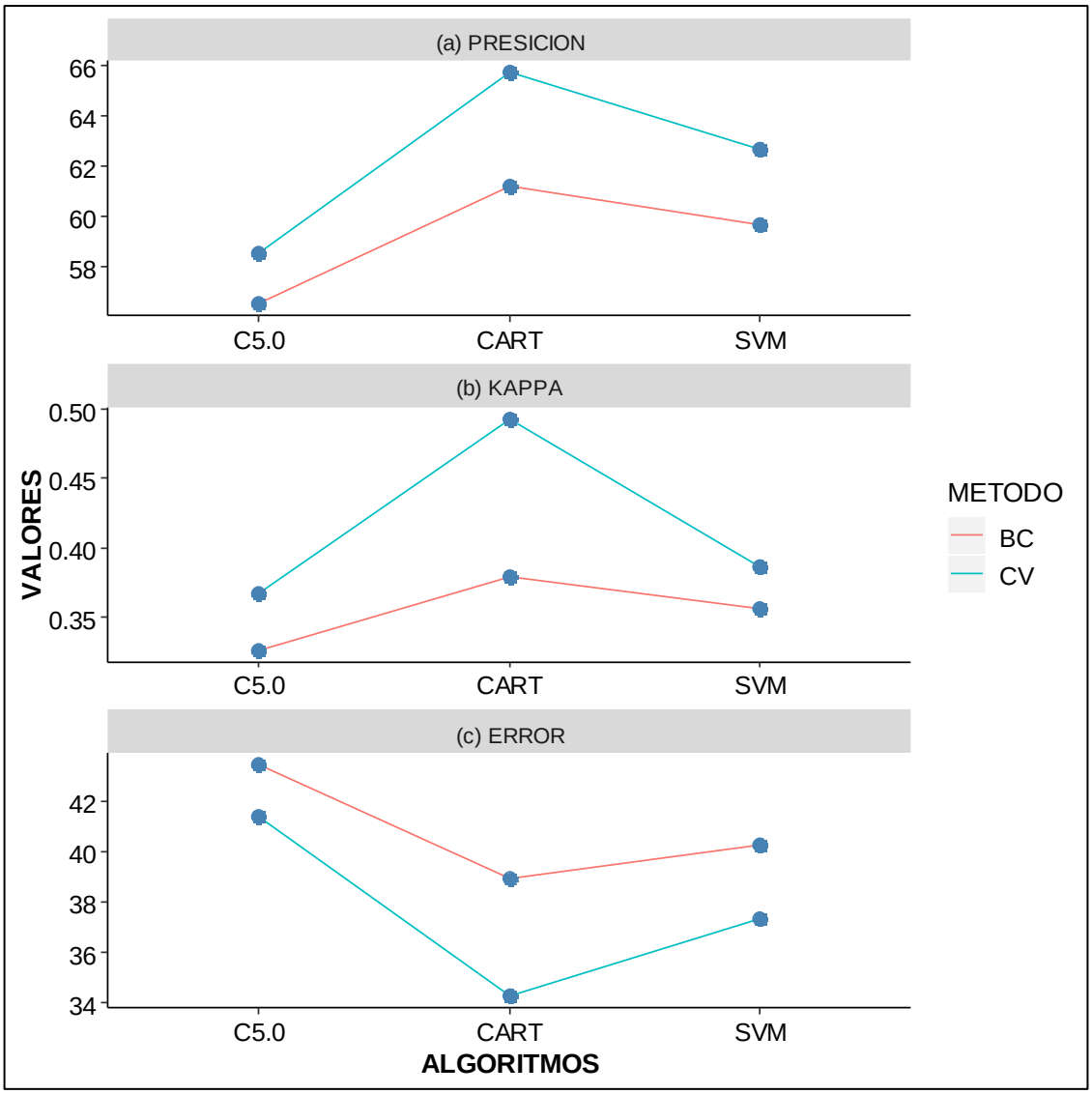
El resumen de los resultados obtenidos se expresó en la Tabla 29-3, donde se apreció que existe diferencias entre los algoritmos de acuerdo a sus propiedades, para esto se consideró los algoritmos mediante la bondad del clasificador y la utilización de una validación cruzada.

Tabla 27-3: Comparación de los Modelos

Algoritmos	Bondad del Clasificador (BC)			Validación Cruzada (CV)			IC: 95%
	Precisión	Error	Kappa	Precisión	Error	Kappa	
C5.0	56.53	43.47	0.3267	58.56	41.44	0.3677	(0.45, 0.61)
SVM	59.70	40.30	0.3567	62.67	37.33	0.3863	(0.47, 0.64)
CART	61.21	38.79	0.3795	65.72	34.28	0.4927	(0.54, 0.70)

Realizado por: Padilla S. Oscar R., 2020

En la Tabla 27-3 y la Grafica 22-3, se pudo concluir que el algoritmo que presenta un mejor rendimiento para los datos de carbono edáfico es el CART aplicado una validad cruzada, debido a que presentó una precisión de 65.72% (Grafica 22-3, (a), un coeficiente de concordancia moderada de 49.27 (Gráfica 22-3, (b)) y además su porcentaje de clasificación errónea de 34.28 el mismo que es menor que del resto (Gráfica 31-3, (c)).



Gráfica 22-3: Comparación de los modelos.
Realizado por: Padilla S. Oscar R., 2020

3.4.4 Validación del problema

Al observar la Tabla 29-3 y la Grafica 24-3, se pudo concluir que el algoritmo que presenta un mejor rendimiento para los datos de carbono edáfico es el CART donde se valida el problema general que dice "¿Cuál es el mejor algoritmo que permite clasificar el contenido de carbono edáfico en los diferentes tipos de suelo?".

Se evidenció en la Tabla 29-3 y la Grafica 24-3 (a), que el algoritmo CART aplicado una validación cruzada presenta una exactitud de 65.72%, donde se valida el problema específico 1 que dice "¿Cuál es la exactitud de los algoritmos empleados en los árboles de decisión que permita clasificar el contenido de carbono edáfico en los diferentes tipos de suelo?".

Finalmente se notó en la Tabla 27-3 y la Gráfica 22-3 (c) que el algoritmo CART aplicado una validación cruzada presenta un error de predicción de 34.28% validando así el problema específico 2 que dice "¿Cuál es el error mediante la técnica de árboles de decisión para predecir el contenido de carbono edáfico de acuerdo con el tipo de suelo?".

3.4.5 Validación de Hipótesis

Hipótesis especifica 1:

H_0: La técnica de árboles de decisión mediante el mejor algoritmo no permite una exactitud mayor al 70%.

H_1: La técnica de árboles de decisión mediante el mejor algoritmo permite una exactitud mayor al 70%.

Prueba de Hipótesis

1. **Planteamiento de Hipótesis**

   $H_0: \mu \leq 70$ $H_1: \mu > 70$

2. **Nivel de significancia**

   $$\alpha = 0.05$$

3. **Estadístico de Prueba**

   $$t = \frac{x - \mu}{\frac{s}{\sqrt{n}}}$$

   $$t = \frac{65.72 - 70}{2.68/\sqrt{10}} = -0.51$$

4. **Regla de decisión**

   $$t > t_\alpha \rightarrow Se\ Rechaza\ H_0$$

   Se comprueba que $t = -0.51$ es menor que $t_\alpha = 1.83$

5. **Conclusión**

   A un nivel de significancia del 5% no se rechaza la hipótesis nula (H_0) y se concluye que no existe evidencia suficiente para indicar que la técnica de árboles de decisión mediante el mejor algoritmo permite una exactitud mayor al 70%.

Hipótesis especifica 2:

H_0: La técnica de árboles de decisión no permite un error de predicción inferior al 30%.

H_1: La técnica de árboles de decisión permite un error de predicción inferior al 30%.

Prueba de Hipótesis

1. **Planteamiento de Hipótesis**

   $H_0: \mu \geq 30$ $H_1: \mu < 30$

2. **Nivel de significancia**

   $$\alpha = 0.05$$

3. **Estadístico de Prueba**

   $$t = \frac{34.28 - 30}{7.18/\sqrt{10}} = 0.19$$

4. **Regla de decisión**

   $$t > t_\alpha \rightarrow Se\ Rechaza\ H_0$$

   Se comprueba que $t = 0.19$ es menor que $t_\alpha = 1.83$

5. **Conclusión**

   A un nivel de significancia del 5% no se rechaza la hipótesis nula (H_0) y se concluye que no existe evidencia suficiente para indicar que la técnica de árboles de decisión permite un error de predicción inferior al 30%.

Hipótesis General

Habiéndose comprobado con respecto a los datos de carbono orgánico de los suelos proporcionados por el MAG , en la hipótesis especifica 1 que la técnica de los árboles de decisión mediante el mejor algoritmo de clasificación no permite una clasificación mayor al 70%, y también se ha comprobado en la hipótesis especifica 2 que la técnica de los árboles de decisión no permite un error de predicción inferior al 30%, se valida la hipótesis general que dice “La clasificación con árboles de decisión mediante el mejor algoritmo permitir catalogar los niveles de carbono edáfico en las distintas zonas de la provincia de Chimborazo con mayor exactitud.”

3.5 Etapa 5: Evaluación

El algoritmo que cumplió mejor estos requisitos fue el CART ya que presentó una precisión y un coeficiente de concordancia mayor que los otros. Por lo tanto, para la valoración de carbono edáfico se trabajó con el algoritmo CART; a continuación, se determinaron las variables de importancia y se mostró de manera gráfica todo el camino que sigue la variable estimada, a través de las variables explicativas; así como su nivel hasta su estimación.

A continuación, se presentan los resultados obtenidos al aplicar el algoritmo CART, para las datas de la provincia de Chimborazo y región Interandina:

3.5.1 Provincia de Chimborazo

3.5.1.1 MAG

En la Tabla 28-3, se presentó las variables independientes posicionadas en una serie ordenada de acuerdo con su importancia, indicando Textura como de mayor importancia y Ecosistema como una variable sin importancia.

Tabla 28-3: Importancia de las variables data MAG Chimborazo.

Variables Independientes	Importancia	Importancia Normalizada (%)
Textura	38.00	100
Índice normalizado de Áreas Quemadas 2 (NBR2)	32.49	85.50
Taxonomía	30.83	81.13
Índice de Resistencia Atmosféricamente Visible (VARI)	28.02	73.74
Índice Diferencial de Agua Normalizado (NDWI)	24.69	64.97
Modelo de Elevación Digital (DEM)	11.42	30.05
Pendiente	11.12	29.26
Índice de Área Calcinada (BI)	3.93	10.34
Ecosistema	0	0

Realizado por: Padilla S. Oscar R., 2020.

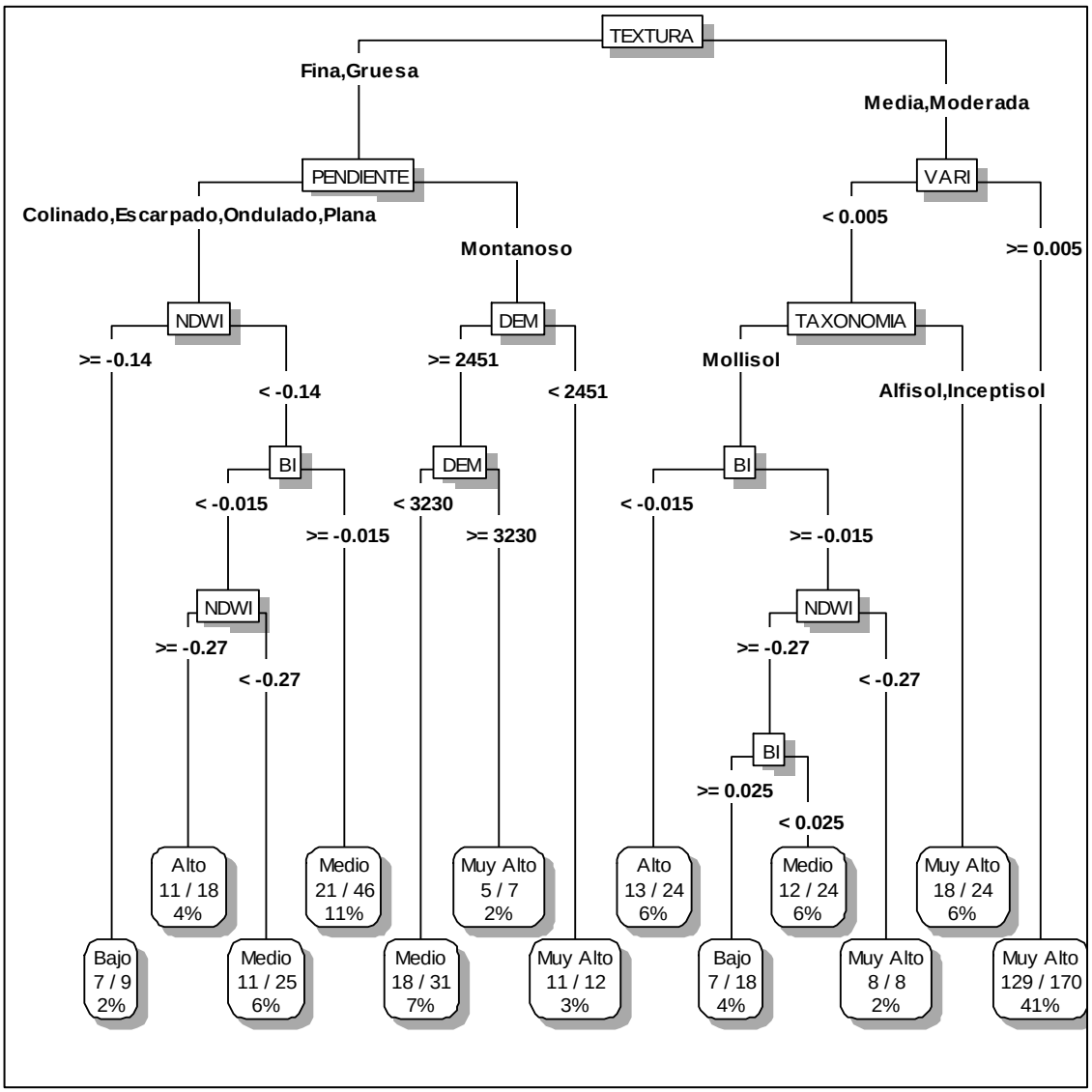
La Tabla 29-3, presentó los resultados sobre el rendimiento del algoritmo con los datos de MAG de la provincia de Chimborazo, el mismo que mostró un 65.72% de precisión del modelo, un error de predicción de 34.28% y un índice de concordancia moderada de 0.493.

Tabla 29-3: Rendimiento del algoritmo con la data MAG Chimborazo.

Medidas del rendimiento	Valores
PRECISION	65.72
ERROR	34.28
KAPPA	0.493
IC: 95%	(0.54, 0.70)

Realizado por: Padilla S. Oscar R., 2020.

En la Gráfica 23-3, se observó el árbol de decisión generado tras el entrenamiento del algoritmo CART con los datos del MAG de la provincia de Chimborazo, el modelo entrenado generó la Textura como variable predominante con 13 reglas de decisión, de las cuales 5 representan a carbono Muy Alto, 2 al carbono Alto, 4 al carbono Medio y 2 al carbono Bajo.



Gráfica 23-3: Árbol de decisión de la Provincia de Chimborazo data MAG.
Realizado por: Padilla S. Oscar R., 2020.

3.5.1.2 FAO

La Tabla 30-3, presentó las variables independientes posicionadas en una serie ordenada de acuerdo con su importancia, indicando que el Modelo de Elevación Digital (DEM) como de mayor importancia y el Índice de Área Calcinada (BI) como una variable sin importancia.

Tabla 30-3: Variables de importancia data FAO Chimborazo.

Variables Independientes	Importancia	Importancia Normalizada (%)
Modelo de Elevación Digital (DEM)	66.19	100
Textura	62.27	94.08
Índice Normalizado de Áreas Quemadas 2 (NBR2)	50.19	75.83
Taxonomía	38.85	58.69
Índice Diferencial de Agua Normalizada (NDWI)	33.64	50.82
Ecosistema	25.96	39.22
Pendiente	5.28	7.98
Índice de Resistencia Atmosféricamente Visible (VARI)	2.51	3.79
Índice de Área Calcinada (BI)	0	0

Realizado por: Padilla S. Oscar R., 2020.

En la Tabla 31-3, se presentó los resultados sobre el rendimiento del algoritmo con los datos de la FAO de la provincia de Chimborazo el mismo que muestra un 70.71% de precisión del modelo, un error de predicción de 29.29% y un índice de concordancia moderada de 0.528.

Tabla 31-3: Rendimiento del algoritmo con la data FAO Chimborazo.

Medidas del rendimiento	Valores
PRECISION	70.71
ERROR	29.29
KAPPA	0.528
IC: 95%	(0.62, 0.76)

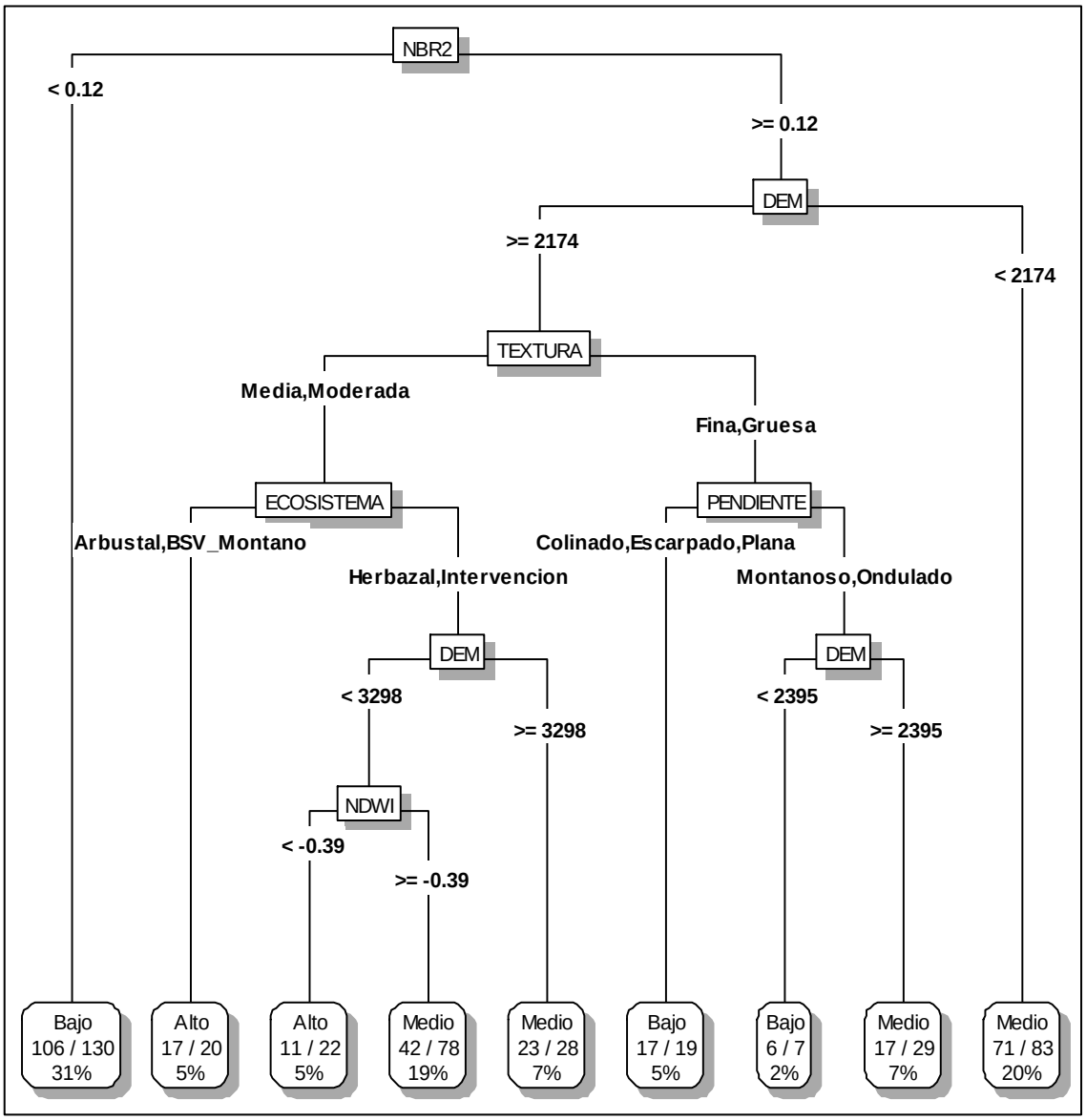
Realizado por: Padilla S. Oscar R., 2020.

En la Gráfica 24-3, se observó el árbol de decisión construido tras el entramiento del algoritmo CART con los datos de la FAO de la provincia de Chimborazo, el modelo entrenado generó el Índice NBR2 como variable predominante con 9 reglas de decisión, de las cuales 2 representa al carbono Alto, 4 al carbono Medio y 3 al carbono Bajo.

Dos de las reglas de decisión para la clasificación de carbono edáfico que arroja el modelo es:

- Si la muestrea obtenida tiene un índice normalizado de áreas quemadas 2 mayor o igual a 0.12 (NBR2 >= 0.12), y fue obtenida a una altura menor a 2174 msnm (DEM < 2174), entonces se indica que el carbono es Medio.

- Si la muestrea obtenida tiene un índice normalizado de áreas quemadas 2 mayor o igual a 0.12 (NBR2 >= 0.12), y fue obtenida a una altura mayor o igual a 2174 msnm (DEM >= 2174), además que se encuentra en la textura Fina o Media y a una Pendiente ya sea Colinado, Escarpado o Plana, entonces se indica que el carbono es Bajo.



Gráfica 24-3: Árbol de decisión de la Provincia de Chimborazo data FAO.
Realizado por: Padilla S. Oscar R., 2020.

3.5.2 Región Interandina

3.5.2.1 MAE

La Tabla 32-3, presentó las variables independientes posicionadas en una serie ordenada de acuerdo con su importancia, indicando que el Modelo de Elevación Digital (DEM) como de mayor importancia y como variables sin importancia para el modelo se indicó Taxonomía y el Índice Normalizado de Áreas Quemadas 2 (NBR2).

Tabla 32-3: Importancia de las variables data MAE región Interandina.

Variables Independientes	Importancia	Importancia Normalizada (%)
Modelo de Elevación Digital (DEM)	26.33	100
Índice de Área calcinada (BI)	20.11	76.38
Pendiente	14.81	56.25
Índice de Resistencia Atmosféricamente Visible (VARI)	12.65	48.04
Índice de Vegetación de Diferencia Normalizada (NDVI)	12.57	47.74
Índice Diferencial de Agua Normalizada (NDWI)	9.45	35.89
Ecosistema	4.32	16.41
Textura	4.03	15.31
Taxonomía	0	0
Índice Normalizado de Áreas Quemadas 2 (NBR2)	0	0

Realizado por: Padilla S. Oscar R., 2020.

La Tabla 33-3, presentó los resultados sobre el rendimiento del algoritmo con los datos del MAE de la región interandina, el mismo que mostró un 59.5% de precisión del modelo, un error de predicción de 40.5% y un índice de concordancia débil de 0.359.

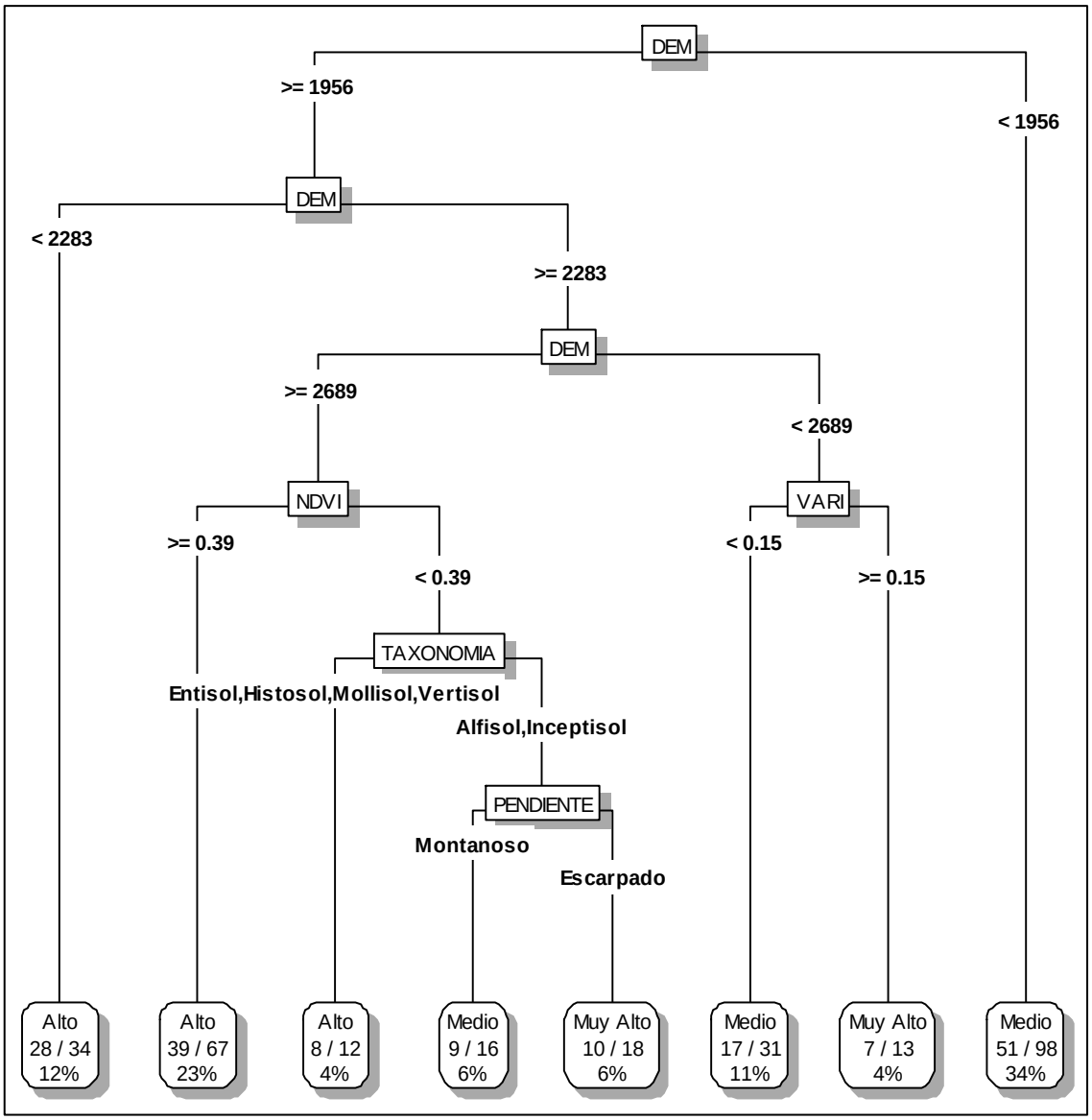
Tabla 33-3: Rendimiento del algoritmo con la data MAE región Interandina.

Medidas del rendimiento	Valores
PRECISION	59.50
ERROR	40.50
KAPPA	0.359
IC: 95%	(0.49, 0.66)

Realizado por: Padilla S. Oscar R., 2020.

En la Gráfica 25-3, se observó el árbol de decisión generado tras el entrenamiento del algoritmo CART con los datos del MAE de la región Interandina, el modelo entrenado generó el DEM (Modelo de elevación digital) como variable predominante con 8 reglas, de las cuales 2 representan a carbono Muy Alto, 3 al carbono Alto y 3 al carbono Medio.

Una de las reglas de decisión para la clasificación de carbono edáfico que arroja el modelo es: Si la muestra obtenida se encuentra a una altura mayor o igual a 1956 msnm (DEM >= 1956), también a una altura mayor o igual a 2283 msnm (DEM >= 2283), finalmente a una altura menor a 2689 msnm (DEM < 2689) y un índice de resistencia atmosféricamente visible mayor o igual a 0.15 (VARI >= 0.15), entonces se indica que el carbono es Muy Alto.



Gráfica 25-3: Árbol de decisión de la región interandina data MAE.
Realizado por: Padilla S. Oscar R., 2020

3.5.2.2 MAG

En la Tabla 34-3, se presentó las variables independientes posicionadas en una serie ordenada de acuerdo son su importancia, indicando Pendiente como de mayor importancia y el Índice de resistencia Atmosféricamente Visible (VARI) sin importancia para el modelo.

Tabla 34-3: Importancia de las variables data MAG región Interandina.

Variables Independientes	Importancia	Importancia Normalizada (%)
Pendiente	19.62	100
Textura	14.80	75.43
Taxonomía	14.72	75.03
Modelo de Elevación Digital (DEM)	13.74	70.03
Índice de Vegetación de Diferencia Normalizada (NDVI)	13.41	68.34
Índice Normalizado de Áreas Quemadas 2 (NBR2)	12,47	63.56
Índice Diferencial de Agua Normalizada (NDWI)	6.14	31.29
Ecosistema	3.26	16.62
Índice de Área Calcinada (BI)	3.19	16.26
Índice de Resistencia Atmosféricamente Visible (VARI)	0	0

Realizado por: Padilla S. Oscar R., 2020.

La Tabla 35-3, presentó los resultados sobre el rendimiento del algoritmo con los datos de MAG de la región Interandina, el mismo que muestra el 70.49% de precisión del modelo, un error de predicción de 29.51% y un índice de concordancia moderada de 0.469.

Tabla 35-3: Rendimiento del algoritmo con la data MAG región Interandina.

Medidas de rendimiento	Valores
PRECISION	70.49
ERROR	29.51
KAPPA	0.509
IC: 95%	(0.62, 0.78)

Realizado por: Padilla S. Oscar R., 2020.

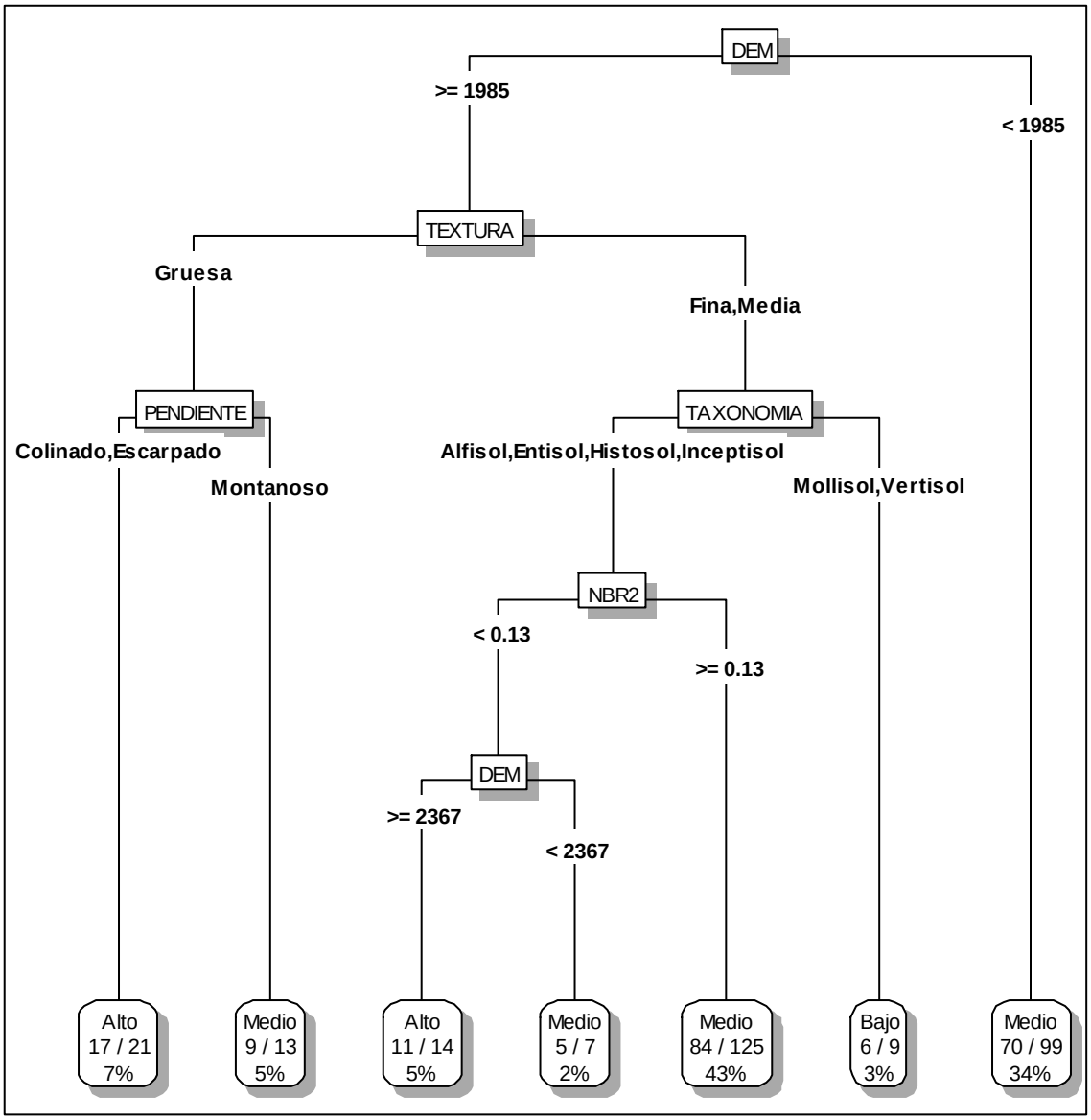
En la Gráfica 26-3, se observó el árbol de decisión construido tras el entramiento del algoritmo CART con los datos del MAG de la región Interandina, el modelo entrenado generó el DEM

(Modelo de Elevación Digital) como variable predominante con 7 reglas de decisión, de las cuales 2 representa al carbono Alto, 4 al carbono Medio y 1 al carbono Bajo.

Dos de las reglas de decisión para la clasificación de carbono edáfico que arroja el modelo es:

- Si la muestra obtenida se encuentra a una altura menor a 1985 msnm (DEM < 1985) se indica que el carbono es Medio.

- Si la muestra obtenida se encuentra a una altura mayor o igual a 1985 msnm, en la textura del suelo Fina o Gruesa (Textura = Fina, Gruesa), en el tipo de suelo Mollisol o Vertisol (Taxonomía = Mollisol, Vertisol), entonces se indica que el carbono es Bajo.



Gráfica 26-3: Árbol de decisión de la región interandina data MAG.
Realizado por: Padilla S. Oscar R., 2020

3.5.2.3 FAO

En la Tabla 36-3, se presentó las variables independientes posicionadas en una serie ordenada de acuerdo a la importancia, indicando el Índice de Vegetación de Diferencia Normalizado (NDVI) como de mayor importancia y Textura como de menor importancia.

Tabla 36-3: Importancia de las variables data FAO región Interandina

Variables Independientes	Importancia	Importancia Normalizada (%)
Índice de Vegetación de Diferencia Normalizado (NDVI)	17.88	100
Modelo de Elevación Digital (DEM)	17.34	96.98
Índice de Resistencia Atmosféricamente Visible (VARI)	17.00	95.08
Índice de Área Calcinada (BI)	15.77	88.20
Ecosistema	13.57	77.01
Índice Diferencial de Agua Normalizado (NDWI)	9.61	53.75
Taxonomía	8.13	45.47
Índice Normalizado de Áreas Quemadas 2 (NBR2)	5.37	30.03
Pendiente	3.56	19.91
Textura	2.24	12.53

Realizado por: Padilla S. Oscar R., 2020

La Tabla 37-3, presentó los resultados sobre el rendimiento del algoritmo con los datos de la FAO de la región interandina el mismo que muestra el 63.93% de precisión del modelo, un error de predicción de 36.07% y un índice de concordancia moderada de 0.464.

Tabla 37-3: Rendimiento del algoritmo data FAO región Interandina

Medidas de rendimiento	Valores
PRECISION	63.93
ERROR	36.07
KAPPA	0.464
IC: 95%	(0.57, 0.68)

Realizado por: Padilla S. Oscar R., 2020

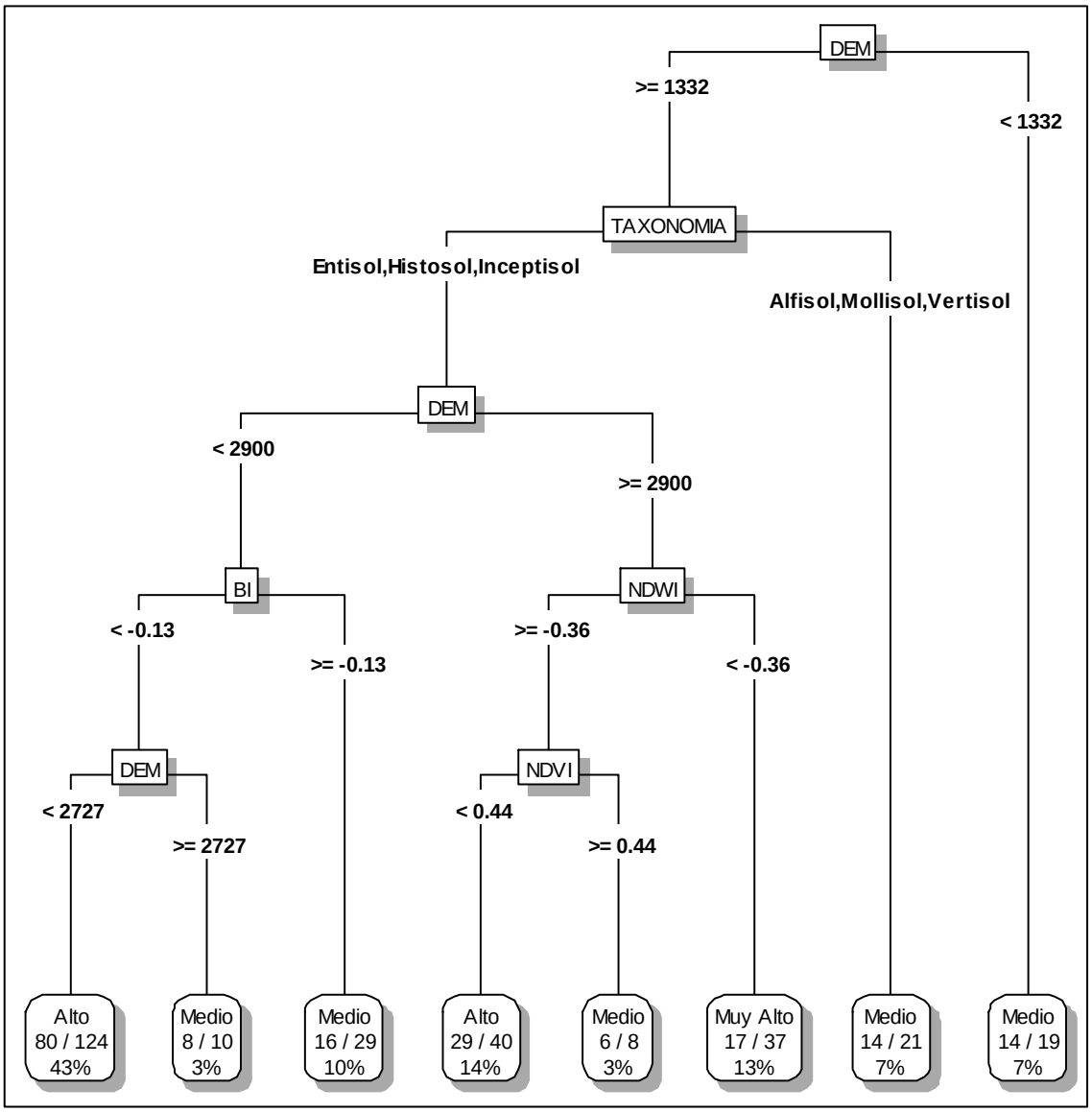
En la Gráfica 27-3, se observó el árbol de decisión construido tras el entramiento del algoritmo CART con los datos de la FAO de la región Interandina, el modelo entrenado generó el DEM

(Modelo de Elevación Digital) como variable predominante con 8 reglas de decisión, de las cuales 1 representa al carbono Muy Alto, 2 al carbono Alto y 5 al carbono Medio.

Dos de las reglas de decisión para la clasificación de carbono edáfico que arroja el modelo es:

- Si la muestra obtenida se encuentra a una altura menor a 1332 msnm (DEM < 1332) se indica que el carbono es Medio.

- Si la muestra obtenida se encuentra a una altura mayor o igual a 1332 msnm (DEM >= 1332), en el tipo de suelo Entisol, Histosol o Inceptisol, con una altura mayor o igual a 2900 msnm (DEM >= 2900) y el Índice Diferencial de Agua Normalizado menor a -0.36 se indica que el carbono es Muy Alto.



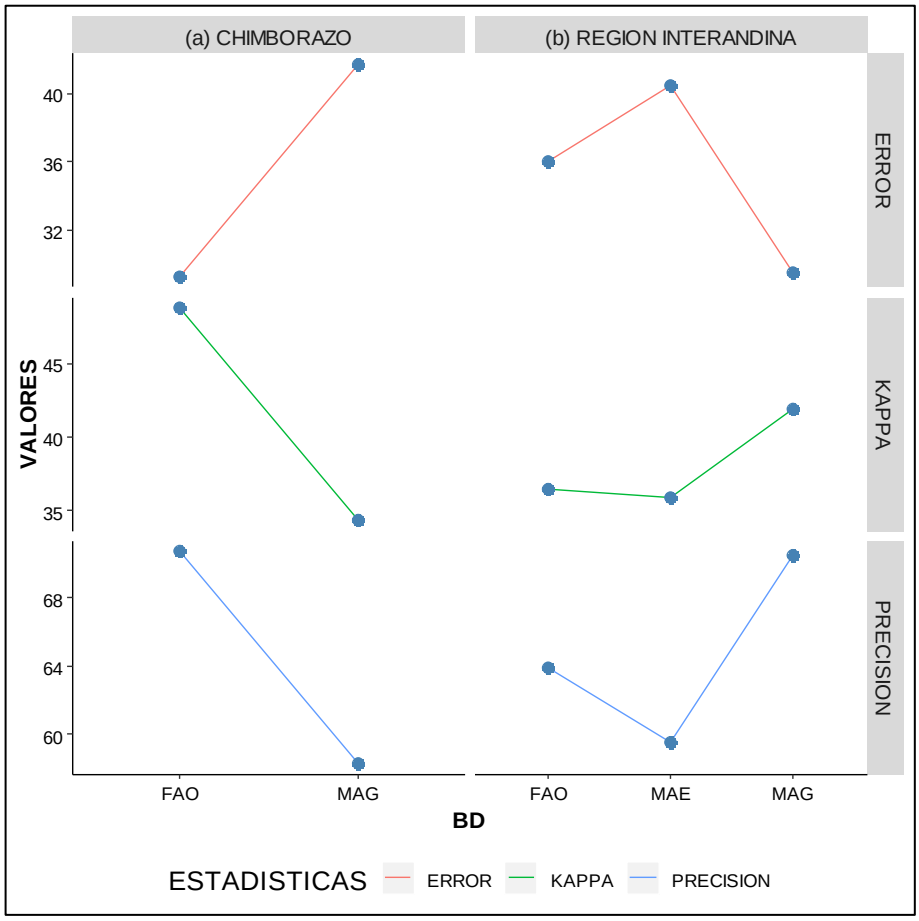
Gráfica 27-3: Árbol de decisión de la región interandina data FAO.
Realizado por: Padilla S. Oscar R., 2020

3.5.3 Comparación de los resultados

Tabla 38-3: Comparación de los Resultados.

MEDIDAS DE RENDIMIENTO	CHIMBORAZO		REGION INTERANDINA		
	MAG	FAO	MAE	MAG	FAO
PRECISION	65.22	70.71	59.50	70.49	63.93
ERROR	34.28	29.29	40.50	29.51	36.07
KAPPA	0.493	0.528	0.359	0.509	0.464
IC: 95%	(0.54, 0.70)	(062, 0.76)	(0.49, 0.66)	(0.62, 0.78)	(0.57, 0.68)

Realizado por: Padilla S. Oscar R., 2020.



Gráfica 28-3: Comparación de los resultados.
Realizado por: Padilla S. Oscar R., 2020

El comportamiento de los resultados obtenidos en los árboles de decisión, generados a partir del algoritmo CART, se observó en la Tabla 38-3, a nivel de la provincia de Chimborazo el algoritmo utilizado tiene una mejor clasificación con los datos de la FAO obtenidos del Digital Global Soil Organic Carbon Map (GSOCmap) elaborado por la Organización de las Naciones Unidas para la

Alimentación y la Agricultura (FAO), puesto que presentó el 70.71% de precisión y un error de 29.29% para predecir el contenido de carbono edáfico; mientras que en el callejón Interandino o región Interandina se pudo observar que el algoritmo tiene una mejor clasificación con los datos del MAG obtenidos del Mapa Digital de Carbono Orgánico de Suelos de Ecuador elaborado por el Ministerio de Agricultura y Ganadería (MAG), ya que presenta el 70.49% de precisión y un error de 29.51% para predecir el carbono orgánico del suelo. El comportamiento de los resultados se pudo observar en la Gráfica 28-3.

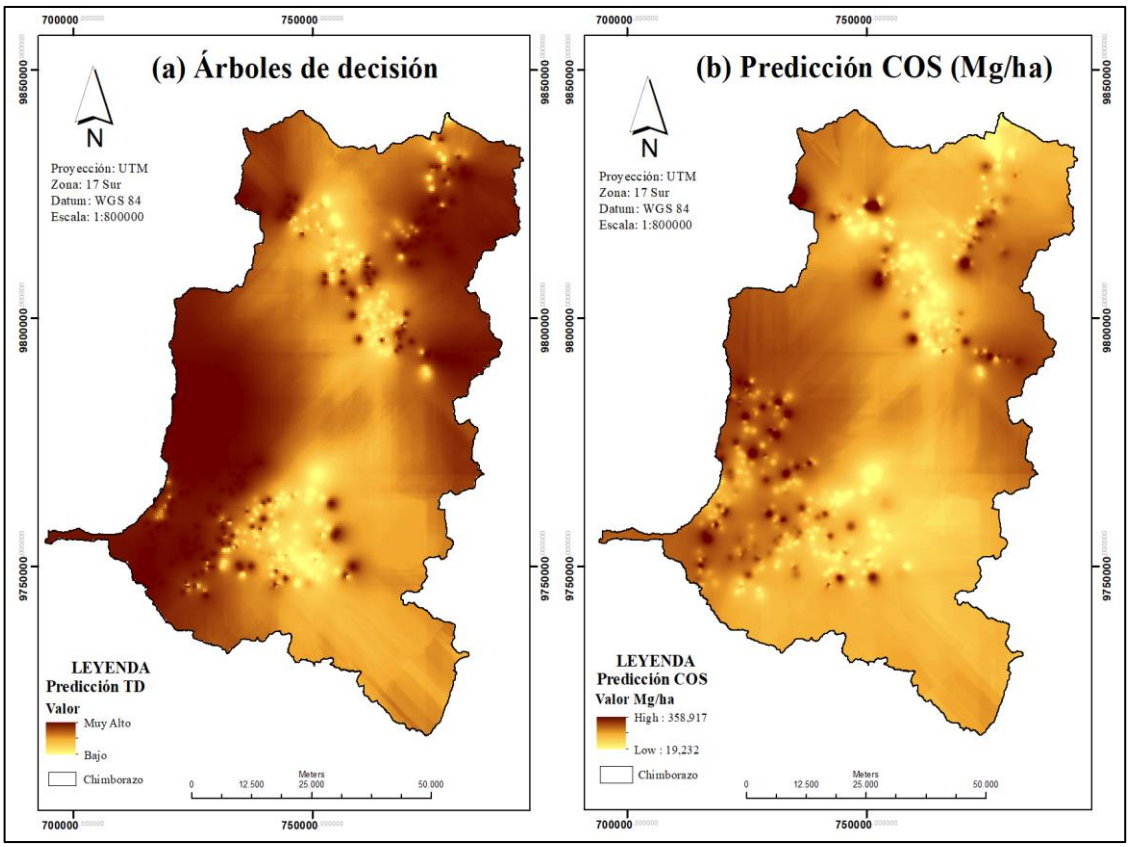
3.6 Etapa 6: Implementación

Esta sección corresponde a la parte proyectiva de este trabajo en que se analizan los resultados obtenidos y se muestran las optimizaciones realizadas a los modelos de manera de obtener el mejor resultado posible.

3.6.1 Provincia de Chimborazo

3.6.1.1 Predicción

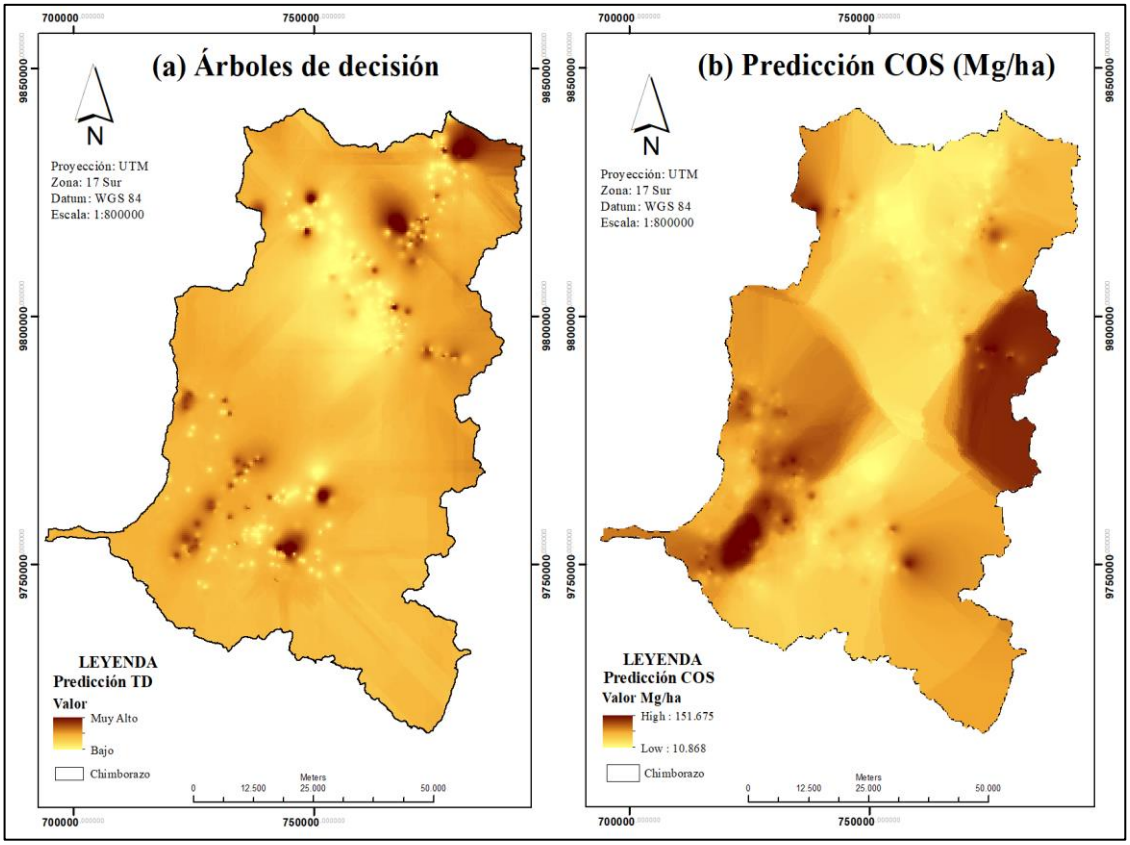
MAG



Gráfica 29-3: Predicción COS MAG Chimborazo
Realizado por: Padilla S. Oscar R., 2020 – GIDAC.

De acuerdo a los datos del Proyecto Regional de Cooperación de Capacitación de Mapeo de Suelos de la Fao elaboradas por el MAG, se pudo observar que la predicción de carbono edáfico generada a través de la técnica de árboles de decisión (Gráfica 29-3) presenta un mapa con los niveles del COS, el cual indica que el color más oscuro representa al carbono orgánico del suelo de clasificación Muy Alto y el color claro representa la clasificación de COS Bajo; mediante la predicción del COS en toneladas por hectárea (Gráfica 29-3, (b)) se obtiene un mapa con los valores del carbono edáfico estableciendo un máximo de 358.92 Mg/ha y un mínimo 19.23 Mg/ha.

FAO

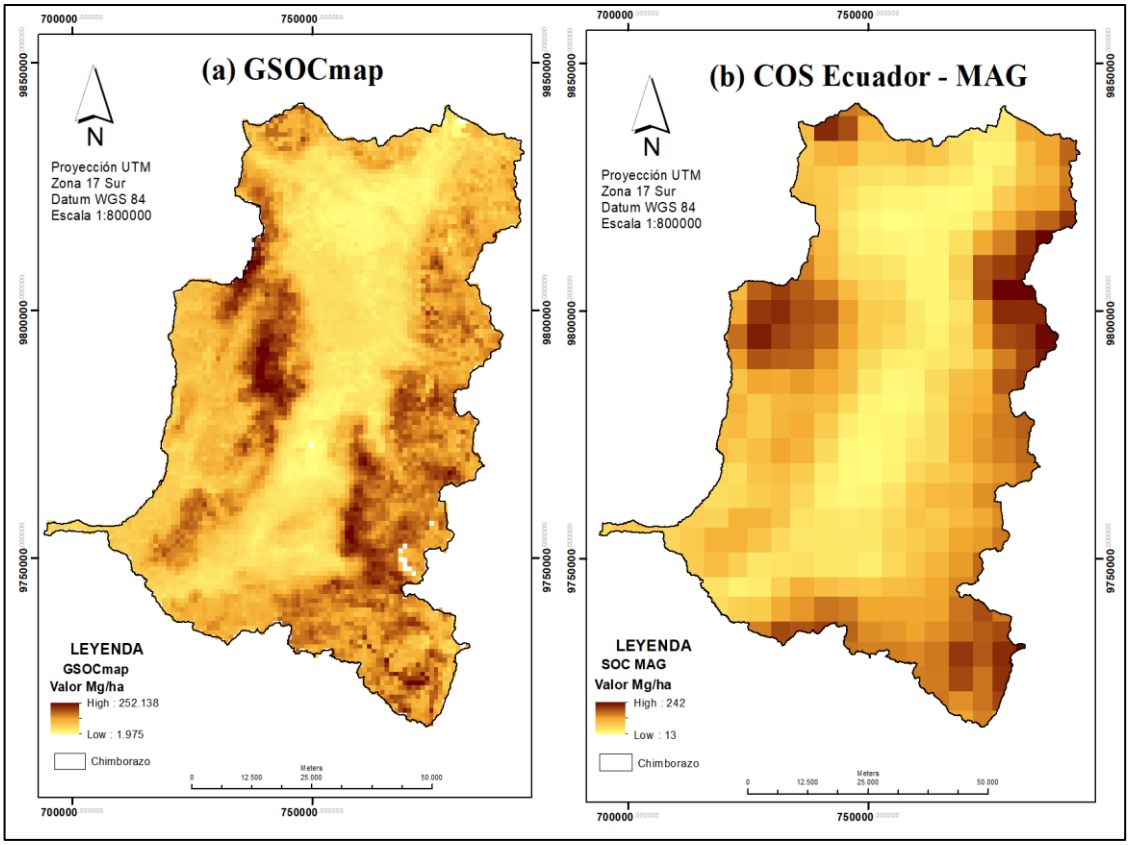


Gráfica 30-3: Predicción COS FAO Chimborazo.
Realizado por: Padilla S. Oscar R., 2020 – GIDAC.

De acuerdo a los datos del GOSCmap elaborados por la FAO, se pudo observar que la predicción de carbono edáfico generada a través de la técnica de árboles de decisión (Gráfica 30-3, (a)) presenta un mapa con los niveles del COS, el cual indica que el color más oscuro representa al carbono orgánico del suelo de clasificación Muy Alto y el color claro representa la clasificación de COS Bajo; mediante la predicción del COS en toneladas por hectárea (Gráfica 30-3, (b)) se obtiene un mapa con los valores de COS estableciendo un máximo de 151.68 Mg/ha y un mínimo 10.87 Mg/ha.

3.6.1.2 Recortes COS

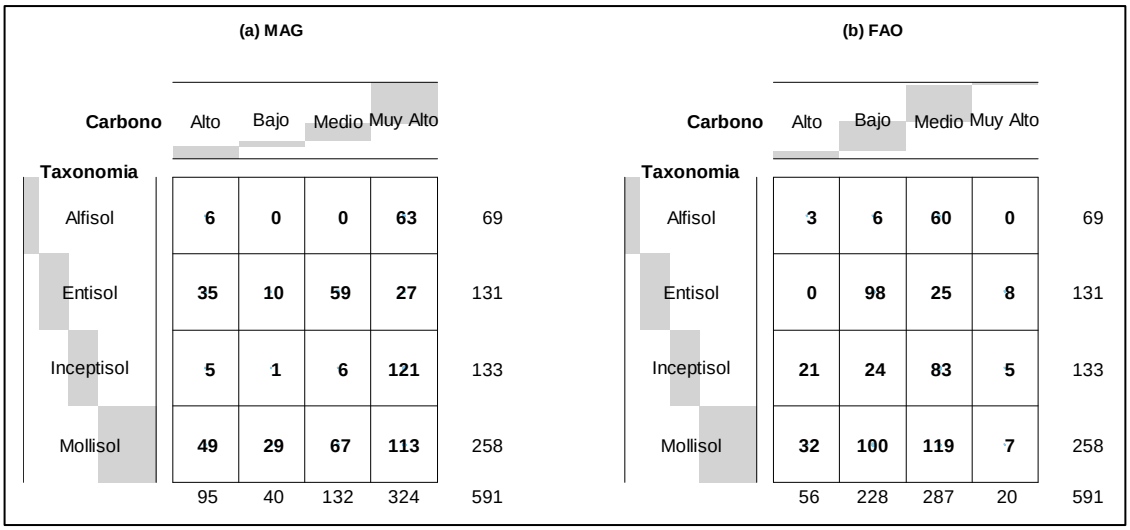
De acuerdo a los recortes de COS (Gráfica 31-3) del mapa Digital Global Soil Organic Carbon Map (GSOCmap) elaboradas por la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) (Gráfica 31-3, (a)) y del mapa de Carbono Orgánico del Suelo de Ecuador elaboradas por el Ministerio de Agricultura y Ganadería (MAG) (Gráfica 31-3, (b)) se pudo observar que tanto la predicción de carbono edáfico mediante árboles de decisión y en toneladas por hectárea se asemejan en gran parte al GSOCmap debido que este se presentó casi la misma forma de colores en base al suelo y que además fue construido a 1Km de pixeles.



Gráfica 31-3: Recortes GSOCmap y COS – Ecuador para la provincia de Chimborazo
Realizado por: Padilla S. Oscar R., 2020 – GIDAC.

3.6.1.3 Niveles de COS

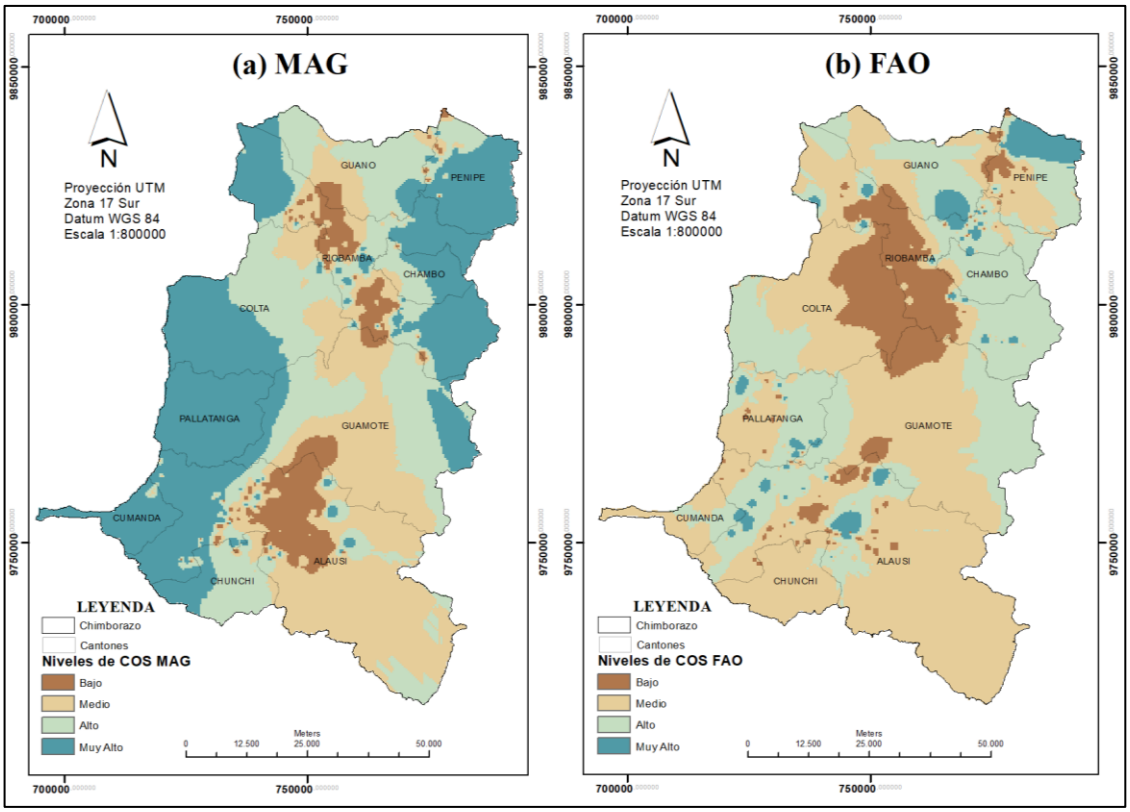
Mediante la nueva clasificación de COS generada por el árbol de decisión, se observó en la Gráfica 32-4, que de un total de 591 muestras la mayor parte se encuentra se encuentra en el tipo de suelo Inceptisol clasificados como carbono Muy Altos (Gráfica 32-3, (a)); mientras que en los datos de la FAO la mayor parte de muestras está en el suelo Mollisol clasificados como carbono Medio (Gráfico 32-3, (b)).



Gráfica 32-3: Niveles de COS por el tipo de suelo de la provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020.

3.6.1.4 Áreas con los niveles de COS

En la Grafica 33-3, se observó la clasificación de COS, los cuales fueron diferenciados por cuatro colores y según la predicción realizada se encontró: para MAG mayores áreas que representan el Carbono Muy Alto (Gráfica 33-3 (a)) mientras que para FAO se obtienen mayores áreas que indican el Carbono Medio (Gráfica 33-3 (b)).



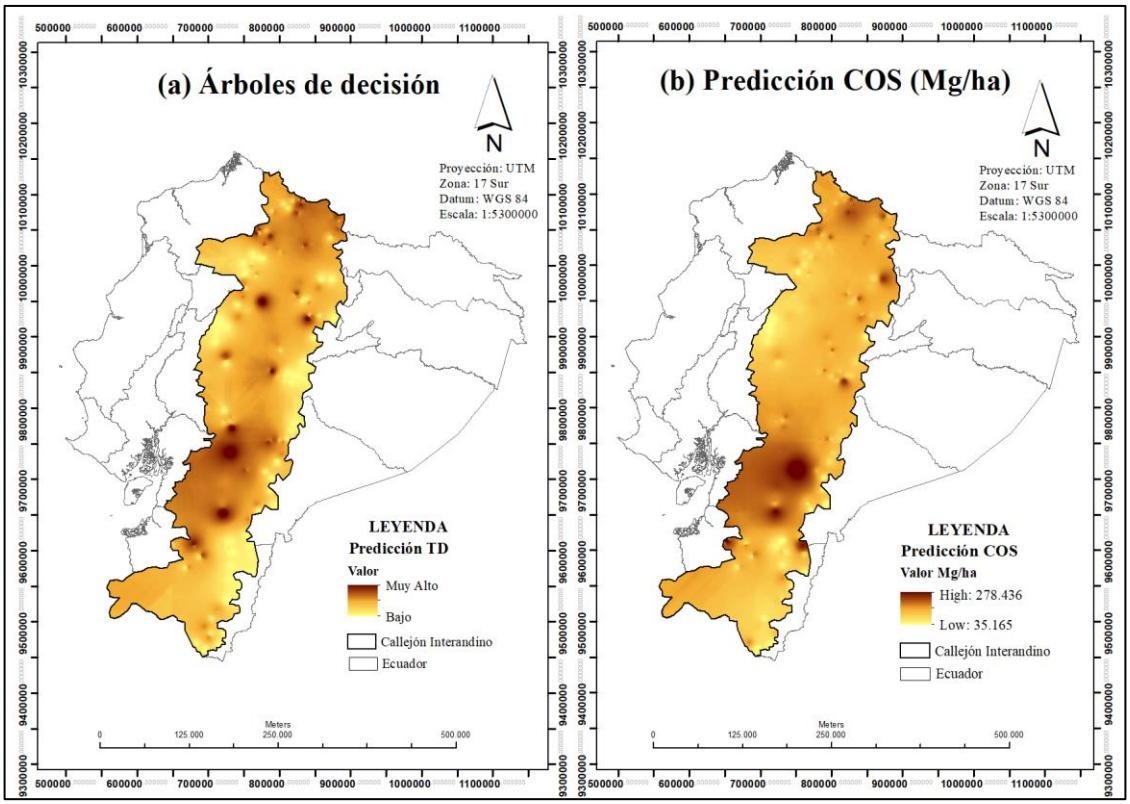
Gráfica 33-3: Áreas según la clasificación de COS en la Provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

3.6.2 Región Interandina

3.6.2.1 Predicción

MAE

De acuerdo a los datos de Evaluación Nacional Forestal MAE - FAO elaborados por el Ministerio del Ambiente del Ecuador (MAE), se pudo observar que la predicción de carbono edáfico generada a través de la técnica de árboles de decisión (Gráfica 34-3, (a)) presenta un mapa con los niveles del COS, el cual indica que el color más oscuro representa al carbono orgánico del suelo de clasificación Muy Alto y el color claro representa la clasificación de COS Bajo; mediante la predicción del COS en toneladas por hectárea (Gráfica 34-3, (b)) se obtiene un mapa con los valores de COS estableciendo un máximo de 278.436 Mg/ha y un mínimo 35.165 Mg/ha.

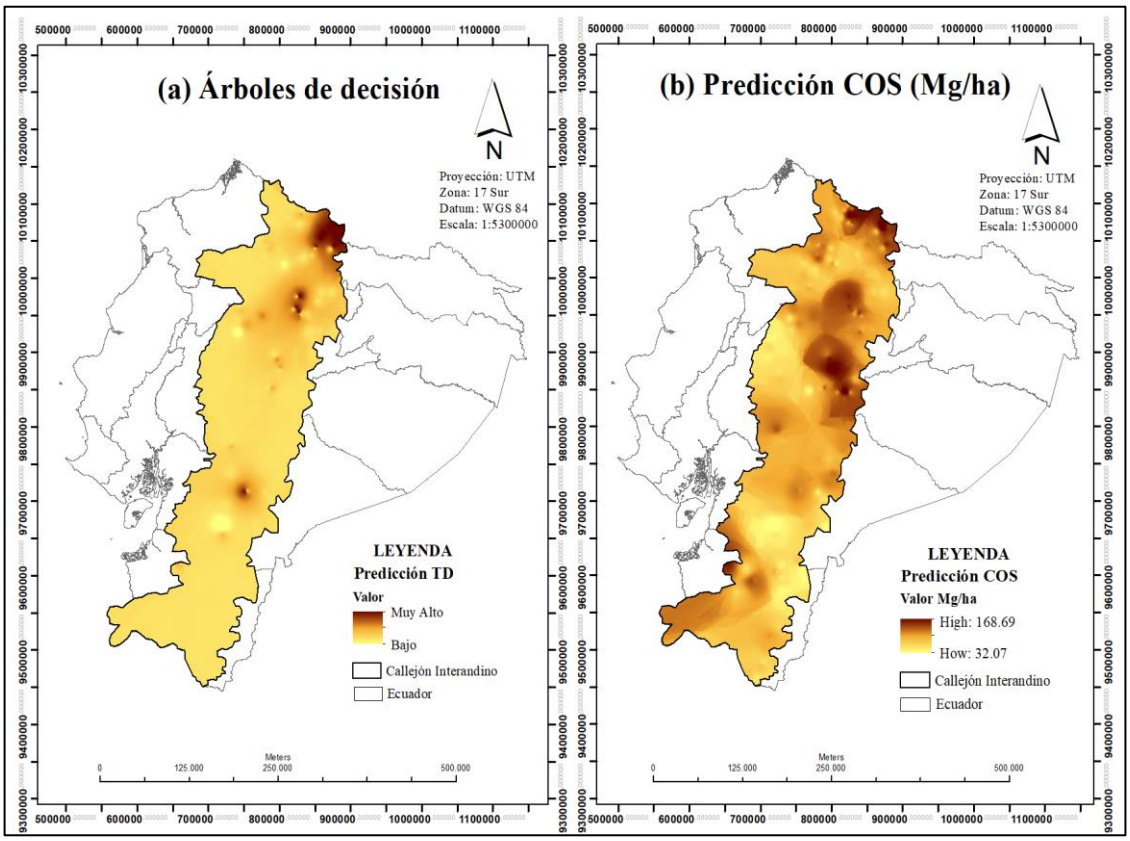


Gráfica 34-3: Predicción COS MAE región Interandina
Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

MAG

De acuerdo a los datos del COS a nivel de Ecuador elaboradas por el Ministerio de Agricultura y Ganadería (MAG), se pudo observar que la predicción de carbono edáfico generada a través de la técnica de árboles de decisión (Gráfica 35-3, (a)) presenta un mapa con los niveles del COS, el cual indica que el color más oscuro representa al carbono orgánico del suelo de clasificación Muy Alto y el color claro representa la clasificación de COS Bajo; mediante la predicción del COS en

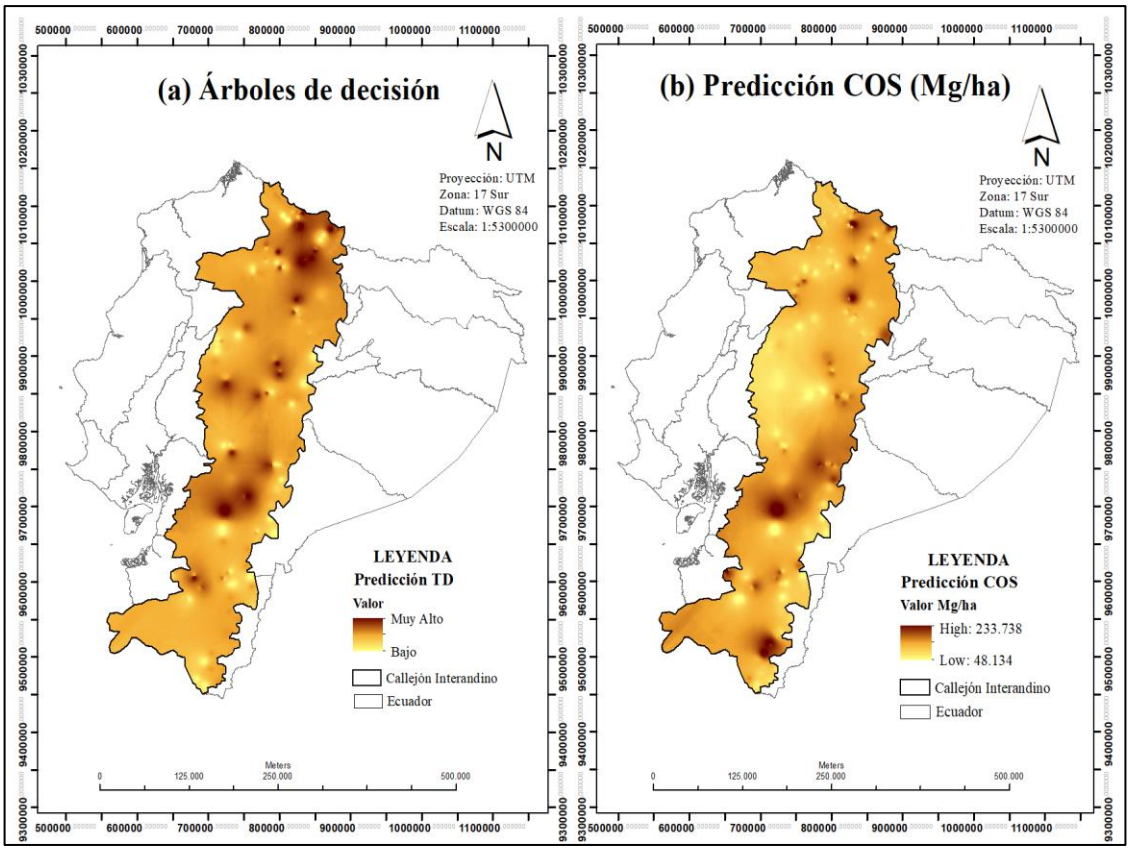
toneladas por hectárea (Gráfica 35-3, (b)) se obtiene un mapa con los valores de COS estableciendo un máximo de 168.69 Mg/ha y un mínimo 32.07 Mg/ha.



Gráfica 35-3: Predicción COS MAG región Interandina.
Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

FAO

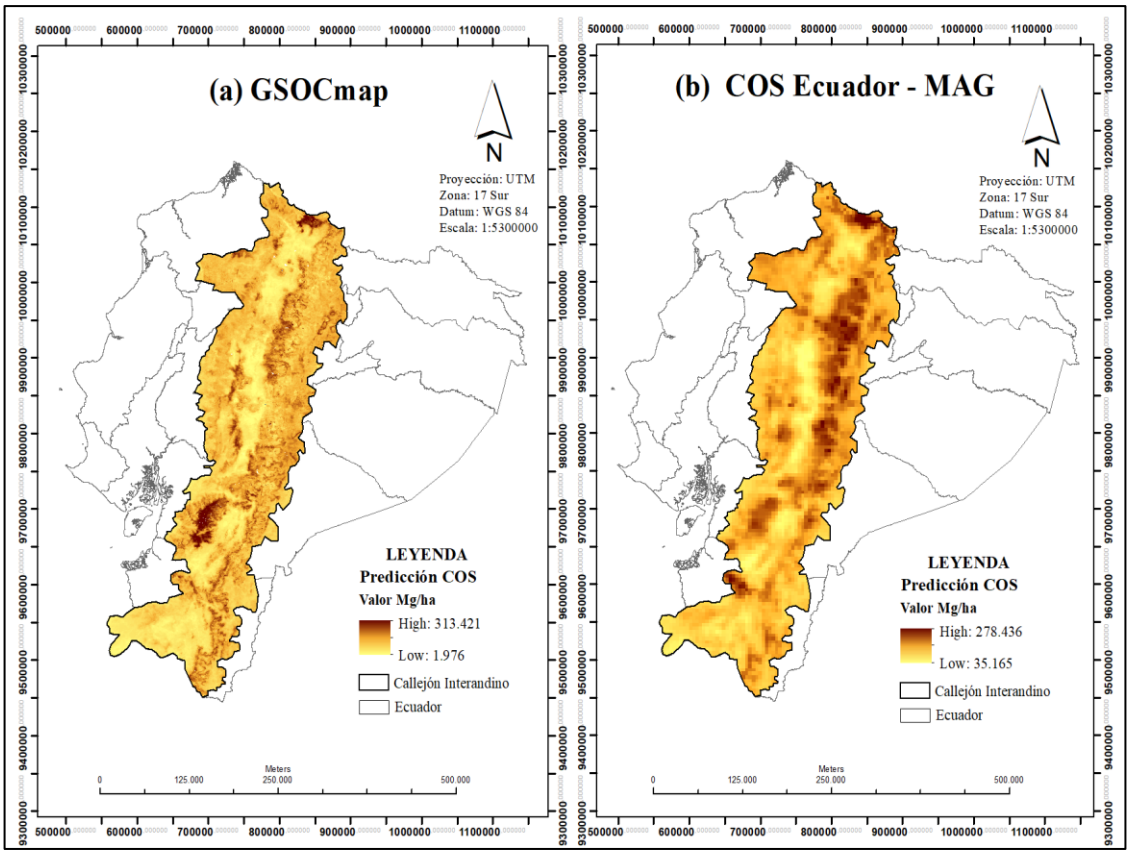
De acuerdo a los datos del GOSCmap elaborados por la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), se pudo observar que la predicción de carbono edáfico generada a través de la técnica de árboles de decisión (Gráfica 36-3, (a)) presenta un mapa con los niveles del COS, el cual indica que el color más oscuro representa al carbono orgánico del suelo de clasificación Muy Alto y el color claro representa la clasificación de COS Bajo; mediante la predicción del COS en toneladas por hectárea (Gráfica 36-3, (b)) se obtiene un mapa con los valores del carbono edáfico estableciendo un máximo de 233.738 Mg/ha y un mínimo de 48.134 Mg/ha.



Gráfica 36-3: Predicción COS FAO región Interandina.
Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

3.6.2.2 Recortes COS

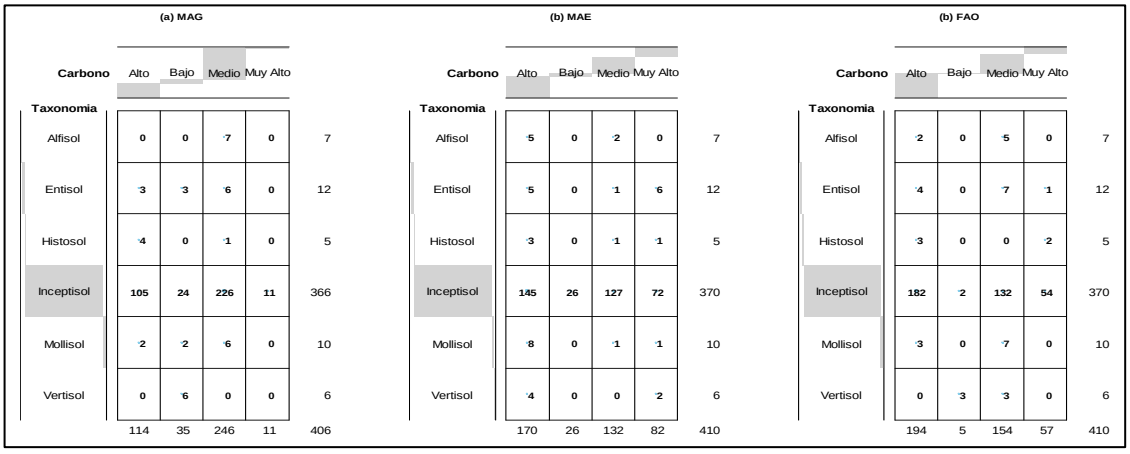
De acuerdo a los recortes de COS (Gráfica 37-3) del mapa Digital Global Soil Organic Carbon Map (GSOCmap) elaboradas por la Organización de Naciones Unidas para las Alimentación y la Agricultura (FAO) (Gráfica 37-3, (a)) y del mapa de Carbono orgánico del Suelo de Ecuador elaboradas por el Ministerio de Agricultura y Ganadería (MAG) (Gráfica 37-3, (b)) se pudo observar que tanto la predicción de carbono edáfico mediante árboles de decisión y en toneladas por hectárea se asemejan en gran parte al GSOCmap debido que este se presentó casi la misma forma de colores en base al suelo y que además fue construido a 1Km de pixeles.



Gráfica 37-3: Recortes GSOCmap y COS – Ecuador para la región Interandina.
Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

3.6.2.3 Niveles de COS

Mediante la nueva clasificación de COS generada por el árbol de decisión, se observó en la Grafica 38-4, que de un total de 591 muestras la mayor parte se encuentra se encuentra en el tipo de suelo Inceptisol clasificados como carbono Muy Altos (Gráfica 38-3, (a)); mientras que en los datos de la FAO la mayor parte de muestras está en el suelo Mollisol clasificados como carbono Medio (Gráfico 38-3, (b)).



(a) MAG

Taxonomia \ Carbono	Alto	Bajo	Medio	Muy Alto	
Alfisol	0	0	7	0	7
Entisol	3	3	6	0	12
Histosol	4	0	1	0	5
Inceptisol	105	24	226	11	366
Mollisol	2	2	6	0	10
Vertisol	0	6	0	0	6
	114	35	246	11	406

(b) MAE

Taxonomia \ Carbono	Alto	Bajo	Medio	Muy Alto	
Alfisol	5	0	2	0	7
Entisol	5	0	1	6	12
Histosol	3	0	1	1	5
Inceptisol	145	26	127	72	370
Mollisol	8	0	1	1	10
Vertisol	4	0	0	2	6
	170	26	132	82	410

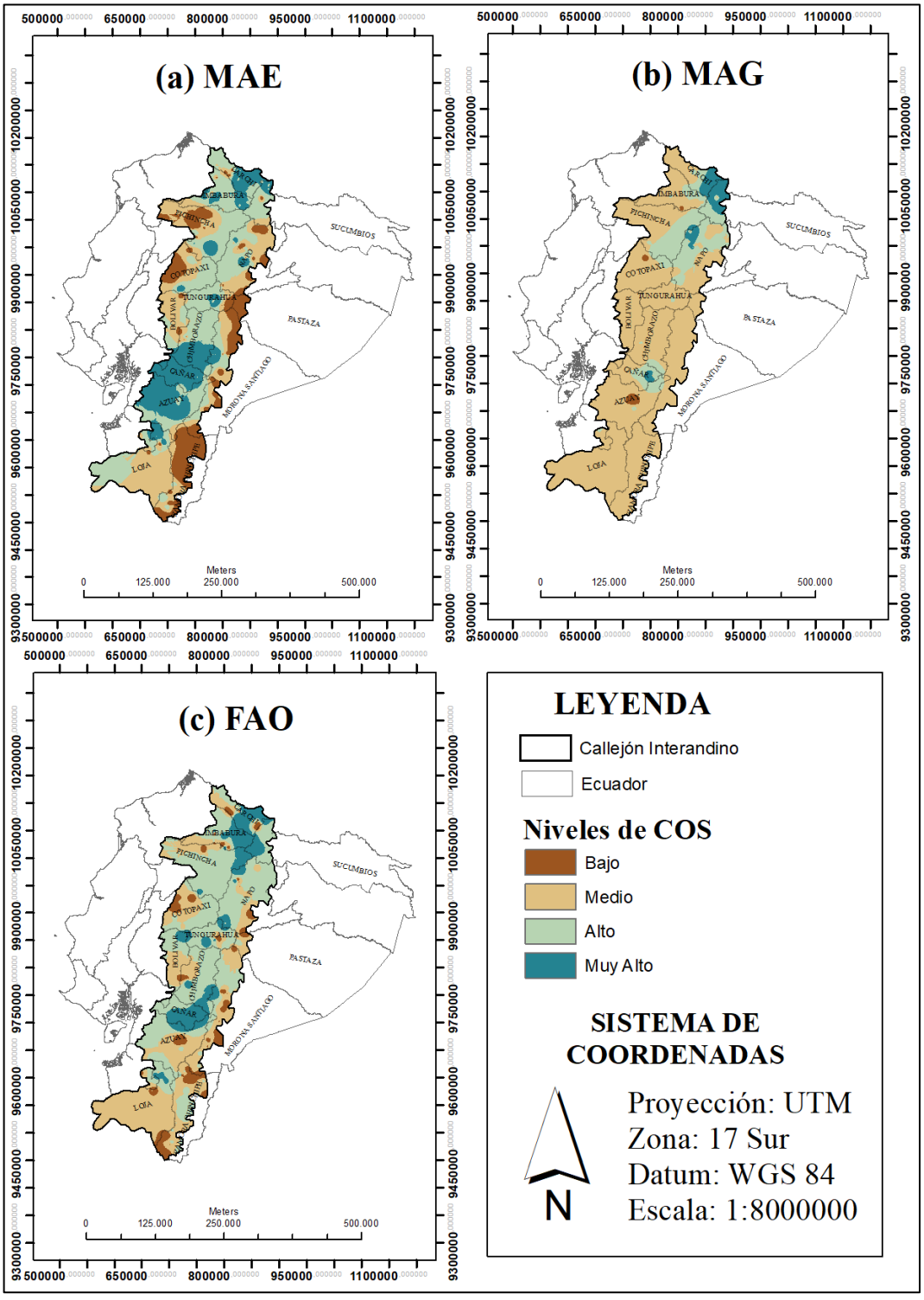
(b) FAO

Taxonomia \ Carbono	Alto	Bajo	Medio	Muy Alto	
Alfisol	2	0	5	0	7
Entisol	4	0	7	1	12
Histosol	3	0	0	2	5
Inceptisol	182	2	132	54	370
Mollisol	3	0	7	0	10
Vertisol	0	3	3	0	6
	194	5	154	57	410

Gráfica 38-3: Niveles de COS por el tipo de suelo de la región Interandina
Realizado por: Padilla S. Oscar R., 2020

3.6.2.4 Áreas con los niveles de COS

En la Gráfica 39-3, se observó la clasificación de COS los mismos que están diferenciados por diferentes colores los cuatro niveles de carbono Bajo, Medio, Alto y Muy Alto; según la predicción se encontró: para MAE y FAO mayores área de Carbono Alto (Gráfica 39-3 (a)) (Gráfica 39-3 (c)) respectivamente, para MAG mayor área de Carbono Medio (Gráfica 39-3 (b)).



Gráfica 39-3: Áreas según la clasificación de COS en la Provincia de Chimborazo.
Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

CONCLUSIONES

En el análisis estadístico se pudo evidenciar que en la provincia de Chimborazo el mayor contenido de COS (360 Mg/ha) se presentó con los datos del MAG, obtenidos a través del Proyecto regional de Cooperación de Capacitación de Mapeo de Suelo de la FAO; fue en una zona de bosque siempre verde montano en la cordillera occidental de los Andes con suelo Alfisol, textura media (0.25mm – 0.50mm), pendiente montañoso; en la región Interandina el mayor COS (290 Mg/ha) se presentó en los datos del MAE obtenidos a través de la Evaluación Nacional Forestal MAE – FAO, y fue en el ecosistema Arbustal, suelo Entisol, con una textura del suelo gruesa (1.0mm – 2.0mm), pendiente de tipo escarpado.

La vegetación que presentó mayores muestras de Carbono Orgánico del Suelo (COS) fueron los ecosistemas intervenidos, en el tipo de suelo mollisol y en la textura media (0.25mm – 0.5mm), obteniendo un nivel Muy Alto en una pendiente de tipo montañoso para la provincia de Chimborazo y un nivel Medio en un pendiente escarpado para la región Interandina.

La variable dependiente Carbono Orgánico del Suelo (COS) de tipo continua fue categorizada de acuerdo al estudio realizado en los suelos del Distrito Federal de México, en 4 niveles: Muy Alto (> 150 Mg/ha), Alto (100 - 150 Mg/ha), Medio (50 - 100 Mg/ha) y Bajo (< 50 Mg/ha), debido a que se pretendió obtener la precisión de cada uno de los algoritmos empleados para determinar el árbol de decisión de tipo clasificación.

La técnica de árboles de decisión mediante el algoritmo de clasificación CART, presentó un mejor rendimiento debido a que los resultados fueron relevantes tanto en su aplicación con el conjunto de prueba y al aplicarlo con los nuevos datos, mostrando una precisión de 65,72%, con un error de predicción de 34.28% y un índice de concordancia moderada de 0.49.

El mapeo digital de COS se generó con las predicciones del carbono orgánico del suelo, mismas que se obtuvieron mediante la técnica de árboles de decisión y permitió revelar de forma aceptable los niveles del contenido de carbono edáfico existentes en los suelos de la provincia de Chimborazo y del callejón Interandino.

RECOMENDACIONES

En base a los resultados encontrados mediante la técnica de árboles de decisión generadas a través del algoritmo CART, es recomendable continuar con estudios relacionados a este tema y extenderlos a zonas de interés nacional, pero se sugiere considerar la variable altitud debido a que la frontera agrícola va desde la parte baja hacia arriba por lo que es necesario conservar las zonas altas.

Es importante incrementar el número de muestras de carbono orgánico del suelo en los distintos niveles, con la finalidad de equilibrar las categorías de la variable clasificadora y elevar la precisión del algoritmo de clasificación basado en la técnica árbol de decisión.

Se deben investigar otras técnicas de aprendizaje supervisado enmarcadas dentro de la minería de datos, necesarias para la construcción de modelos de clasificación del contenido de carbono orgánico en los suelos, pues la minería de datos una herramienta aplicable en múltiples áreas y con varios fines.

Es necesario incrementar el número de muestras de carbono orgánico del suelo en zonas nativas o endémicas, debido a que el estudio realizado se efectuó con mayores muestras de zonas intervenidas; esto permitirá conocer el comportamiento del COS a medida que trascurran los años, y se logrará identificar el aumento o pérdida de la cantidad de Carbono edáfico cuando los bosques se transforman en zonas de cultivos.

Para futuras investigaciones con el fin de evitar una sobre estimación es necesario determinar la concentración de carbono edáfico de manera directa, para lo cual se debe aplicar el método del Analizador TOC; por ello sería importante equipar los laboratorios con lo que se requieran este tipo de estudios.

Se recomienda continuar con investigaciones bajo esta línea, mismas que permitirán identificar el potencial de la técnica de árboles de decisión en diferentes campos de la ciencia; para que puedan ser aplicados es situaciones de interés nacional.

GLOSARIO

Inteligencia Artificial: Tiene como principal objetivo imitar el comportamiento humano con el fin de obtener el mejor resultado esperado, permite dar solución a problemas referentes al análisis de la información con el fin de optimizar el aprendizaje y la toma de decisiones (Gómez Victoria, 2014, p. 35).

Aprendizaje automático: Son métodos de análisis y modelización de datos que se basan en la Inteligencia Artificial. El planteamiento de estos métodos es imitar hasta un cierto punto a la Inteligencia Natural, donde el aprendizaje suele ser base de la presentación de ejemplos, contraejemplos y excepciones (Coello Blanco et al., 2018, p. 1421).

Aprendizaje supervisado: Se utilizan en problemas en los cuales se conoce a priori el número de clases y los reconocimientos de patrones representantes de cada clase. Tienen como objetivo determinar cuál es la clase a que pertenece una nueva muestra sin clase, en base a las clases de las que ya se tiene conocimiento, así como patrones de entrada y salida (Villanueva Morales et al., 2015, p. 264).

Arboles de Decisión: Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la IA, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema (Quintero y Amézquita Collazos, 2003, p. 9).

Minería de datos: Es una solución para el análisis de fenómenos no explicativos en bases de datos y la búsqueda de patrones ocultos entre los datos, para posteriormente ser usados en la predicción de comportamientos a futuro, es decir mediante la aplicación de técnicas de Inteligencia Artificial y de Aprendizaje Automático, facilitando así la toma de decisiones (Escobar Terán et al., 2016, p. 506).

Clasificación: El objetivo de la clasificación es definir una serie de clases, que pueden ser jerárquicas, dentro de las cuales se pueden colocar; por ejemplo, los diferentes tipos de clientes que tiene una empresa (Hernández Y., 2015).

Correlación: Valor entre uno y menos uno que implica el grado de relación entre dos variables. Una correlación positiva indica que existe una relación directamente proporcional entre las variables, y una correlación negativa indica que existe una relación inversamente proporcional entre las variables.

Exactitud: Es la proximidad de un resultado al valor verdadero. Se calcula dividiendo el número total de registros correctamente clasificados, por el número total de registros o de referencia y expresándolo como porcentaje (Fawcett, 2006, p. 865).

Precisión: Es la medida de cuan cerca o dispersos están los resultados unos de otros, y se expresa normalmente como la desviación estándar o desviación estándar relativa, ya que se acepta la varianza como el mejor indicador de la dispersión; menor varianza, mayor precisión (Crubellati et al., 2009, p. 25).

Sensibilidad: Es la propiedad del método que demuestra la variación de respuesta en función de la concentración del analito. Puede ser expresada por la pendiente de la recta de regresión de calibración (Crubellati et al., 2009, p. 26).

Coeficiente Kappa: Es un estadístico que se emplea para cuantificar el grado de acuerdo entre os observadores, corrige el factor azar. Cosiste en el estudio de fiabilidad por equivalencia o concordancia. Cuando el valor obtenido es menor que -1 se dice que las variables tienen poca relación mientras si el valor es cercano a 1, se dice que existe una fuerte relación entre las variables (López de Ullibarri y Pita Fernández, 2010, p. 171).

Predicción: A partir de un conjunto de datos históricos con resultado conocido, se pretende modelizar estos datos para poder saber resultados futuros. Un modelo predictivo tiene diversas variables de entrada que han sido seleccionadas por su alta correlación con el resultado histórico, y la salida es el resultado en sí.

Segmentación: Es la división (o partición) de la totalidad de los datos en segmentos. Una segmentación es a menudo un paso previo a la modelización, dado que es más fácil crear un modelo para un segmento. Además, es uno de los métodos que sirve para poner datos en grupos.

BIBLIOGRAFÍA

ACOSTA, J.C.; et al. Determinación de perfiles de rendimiento académico en la UNNE con minería de datos educacional. [En línea]. Argentina. s.n. 2018. pp. 1078-1082. [Consulta: 25 marzo 2020]. ISBN 978-987-3619-27-4. Disponible en: http://sedici.unlp.edu.ar/handle/10915/68389. Ciencias Informáticas

AZEVEDO, A.I.R.L.; & SANTOS, M.F. KDD, SEMMA and CRISP-DM: a parallel overview. *IADS - DM* [En línea], 2008, vol. 7, nº. 3, pp. 181-186. [Consulta: 17 septiembre 2019]. ISSN 1530-9704. Disponible en: https://recipp.ipp.pt/handle/10400.22/136.

BLANCO, E.J.; & SANZ, H. Algoritmos de clustering y aprendizaje automático aplicados a Twitter. [En línea], 2016, vol. 56. [Consulta: 3 enero 2020]. Disponible en: https://upcommons.upc.edu/bitstream/handle/2117/82434/113257.pdf?sequence=1&isAll.

BORTOLINI, J.L.; et al. Arboles de clasificacion de Potimirim mexicana (Decapoda: Caridea), organismo hermafrodita protandrico secuencial. *Latin American Journal of Aquatic Research* [En línea], 2013, vol. 41, nº. 4, pp. 739-745. [Consulta: 28 mayo 2020]. ISSN 0718560X, 0718560X. DOI 10.3856/vol41-issue4-fulltext-10. Disponible en: http://www.lajar.cl/pdf/imar/v41n4/Articulo_41_4_10.pdf.

BOUZA HERRENA, C.N.; et al. *Modelación matemática de fenómenos del medio ambiente y a salud* [En línea]. Secretaría de salud del estado de Tabasco - Mexico: UGR. 2014. ISBN 84-616-7997-0. [Consulta: 16 diciembre 2019]. Disponible en: http://rgdoi.net/10.13140/2.1.1170.9126.

BRAÑA, J.P.; et al. Modelo de Sentiment Analysis para la clasificación de noticias en tiempo real en el Mercado de Valores de Buenos Aire. [En línea], 2016, vol. 18, pp. 333-337. [Consulta: 15 febrero 2020]. ISSN 978-950-698-377-2. Disponible en: https://digital.cic.gba.gob.ar/handle/11746/3198.

BRAVO MORALES, Nilo .Frank. *Teledeteccion Espacial LandSat,sentinel2, aster L1T y Modis* [En línea]. 1. Huánuco - Perú: s.n. 2017. [Consulta: 8 noviembre 2019]. Disponible en: https://acolita.com/wp-content/uploads/2018/01/Teledeteccion_espacial_ArcGeek.pdf.

BUZZI, M.A.; et al. Múltiples índices espectrales para predecir la variabilidad de atributos estructurales y funcionales en zonas áridas. [En línea], 2017, pp. 55-62. [Consulta: 15 mayo 2020]. Disponible en: https://www.academia.edu/31754883/M%C3%BAltiples_%C3%A

Dndices_espectrales_para_predecir_la_variabilidad_de_atributos_estructurales_y_funcionales_en_zonas_%C3%A1ridas.

CALANCHA ZUNIGA, N.A.; et al. *Breve aproximación a la técnica de árbol de decisiones*. 2010. S.l. s.n. 2010.

CAMBORDA ZAMUDIO, M.G. *Aplicación de árboles de decisión para la predicción del rendimiento académico de los estudiantes de los primeros ciclos de la Carrera de Ingeniería Civil de la Universidad Continental* [En línea]. (Trabajo de Titulación). (Maestría). Universidad Nacional del Centro del Perú. Huancayo - Perú. 2014. pp. 1-157. [Consulta: 25 abril 2019]. Disponible en: http://repositorio.uncp.edu.pe/handle/UNCP/1477.

CAMMARANO, D.; et al. Assessing the Robustness of Vegetation Indices to Estimate Wheat N in Mediterranean Environments. *Remote Sensing* [En línea], 2014, vol. 6, nº. 4, pp. 2827-2844. [Consulta: 7 noviembre 2019]. ISSN 2072-4292. DOI 10.3390/rs6042827. Disponible en: https://www.mdpi.com/2072-4292/6/4/2827.

CARVACHO BART, L.; & MARCELA, S.M. Comparación de índices de vegetación a partir de imágenes MODIS en la región del Libertador Bernardo O´Higgins, Chile,en el período 2001-2005. , 2010, pp. 728-737. ISSN 978-84-472-1294-1.

CHUVIECO SALINERO, E. *Teledetección ambiental, la observación de la Tierra desde el espacio* [En línea]. 3. Barcelona - España: Ariel, S. A. 2008. ISBN 978-84-344-8073-3. [Consulta: 6 noviembre 2019]. Disponible en: https://siglibreuruguay.wordpress.com/2016/09/01/libro-gratuito-teledeteccion-ambiental-la-observacion-de-la-tierra-desde-el-espacio-de-emilio-chuvieco/.

COELLO BLANCO, L.; et al. Uso de técnicas de minería de datos en la enseñanza del algebra lineal. [En línea], 2018, pp. 1420-1427. [Consulta: 25 enero 2020]. Disponible en: http://funes.uniandes.edu.co/11884/.

CORRALES GASCA, P.J.; et al. Métodos de clasificación: Análisis de fertilidad. [En línea], 2015, vol. 35, nº. 111, pp. 43-57. [Consulta: 25 marzo 2020]. ISSN 1405-1249. Disponible en: http://itcelaya.edu.mx/ojs/index.php/pistas/article/view/356.

CORTÉS MARTÍNEZ, F.; et al. Rules for predicting compliance with the quality of wastewater in a treatment plant applying data mining. [En línea], 2018, vol. 21, nº. 62, pp. 13-24. [Consulta: 5 marzo 2020]. ISSN 1988-3064. DOI 10.4114. Disponible en: http://journal.iberamia.org/index.php/intartif/article/view/168.

COSS BU, R. *Análisis y evaluación de proyectos de inversión* [En línea]. 2. México: Editorial Limusa. 2005. ISBN 978-968-18-1327-8. [Consulta: 20 diciembre 2019]. Disponible en: https://books.google.es/books?hl=es&lr=&id=XfVvR-TwcbEC&oi=fnd&pg=PA15&dq=Analisis+y+

evaluaci%C3%B3n+de+proyectos+de+inversi%C3%B3n&ots=avV7emWn5c&sig=PWMfnksL1rLy-pE25zqmlEXBq_A#v=onepage&q=Analisis%20y%20evaluaci%C3%B3n%20de% 20proyectos%20de%20inversi%C3%B3n&f=false.

CRISTIANINI, N.; et al. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* [En línea]. S.l.: Cambridge University Press. 2000. ISBN 978-0-521-78019-3. Disponible en: https://books.google.es/books?hl=es&lr=&id=_PXJn_cxv0AC&oi=fnd&pg=PR9&dq=Nello+Cristianini+and+John+Shawe-Taylor.+An+Introduction+to+Support+Vector+Machines.+Cambridge+University+Press,+2000.&ots=xSVe2EWo2a&sig=l5KVzd7SY8Tjha6SUCznYi2LP5M#v=onepage&q&f=false.

CRUBELLATI, R.; et al. *Aspectos prácticos de la validación e incertidumbre en medidas químicas.* [En línea]. S.l.: CYTED. Área de Desarrollo Sostenible. 2009. ISBN 978-987-96413-8-5. [Consulta: 4 julio 2020]. Disponible en: http://jadimike.unachi.ac.pa/handle/123456789/149.

DUPOUY BERRIOS, C.G. *Aplicación de árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa de interés en Chile* [En línea]. (Trabajo de Titulación). (Maestría). Universidad de Chile. Santiago - Chile. 2014. pp. 1-91. [Consulta: 3 abril 2019]. Disponible en: http://repositorio.uchile.cl/handle/2250/117556.

ECKERT, K.B.; & SUÉNAGA, R. Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. *Formación universitaria* [En línea], 2015, vol. 8, nº. 5, pp. 03-12. [Consulta: 6 febrero 2020]. ISSN 0718-5006. DOI 10.4067/S0718-50062015000500002. Disponible en: https://scielo.conicyt.cl/scielo.php?script=sci_abstract&pid=S0718-50062015000500002&lng=es&nrm=iso&tlng=en.

ESCOBAR TERÁN, H.E.; et al. Aplicaciones de Minería de Datos en Marketing. [En línea], 2016, vol. 3, nº. 8, pp. 503-512. [Consulta: 17 febrero 2020]. ISSN 1390-9304. Disponible en: https://revistapublicando.org/revista/index.php/crv/article/view/169.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters* [En línea], 2006, vol. 27, nº. 8, pp. 861-874. [Consulta: 18 noviembre 2019]. ISSN 0167-8655. DOI 10.1016/j.patrec.2005.10.010. Disponible en: http://www.sciencedirect.com/science/article/pii/S016786550500303X.

FERNÁNDEZ JERI, L.; & SALINAS FLORES, J. Evaluación de la decisión de obtener el título profesional con la elaboración de la tesis mediante técnicas multivariantes: Caso Universidad Nacional Agraria La Molina. *Anales Científicos* [En línea], 2017, vol. 78, nº. 2, pp. 92-99. [Consulta: 17 marzo 2020]. ISSN 2519-7398. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=6232132.

FERNÁNDEZ MENÉNDEZ, S.; et al. Incendios Forestales y Transferencia de Carbono Biomasa-Suelo en areas de montaña de Clima Atlántico. [En línea], 2011, [Consulta: 25 febrero 2020]. DOI 10.13140/RG.2.1.2996.4647. Disponible en: http://rgdoi.net/10.13140/RG.2.1.2996.4647.

GALA GARCÍA, Y. *Algoritmos SVM para problemas sobre big data* [En línea]. (Trabajo de Titulación). (Maestría). Univesidad Autónoma de Madrid. Madrid - España. 2013. pp. 1-68. [Consulta: 16 octubre 2019]. Disponible en: https://repositorio.uam.es/bitstream/handle/10486/14108/66152_Yvonne_Gala_Garcia.pdf?sequence=1&isAllowed=y.

GARCÍA, E.; & OTTO, M. Caracterización ecohidrológica de humedaes alto andinoss usando imágenes de satélite multitemporales en la cabecera de cuenca del río Santa Ancash, Perú. *Ecología Aplicada* [En línea], 2015, vol. 14, nº. 1-2, pp. 115-125. [Consulta: 5 noviembre 2019]. ISSN 1993-9507. DOI 10.21704/rea.v14i1-2.88. Disponible en: http://revistas.lamolina.edu.pe/index.php/eau/article/view/88.

GASTALDI, C.; et al. Teoría de la Decisión: Contribuciones de Von Neumann. *Divulgaciones Matemáticas* [En línea], 1998, vol. 6, nº. 1, pp. 37-42. [Consulta: 2 junio 2019]. Disponible en: http://emis.ams.org/journals/DM/v61/art5.pdf.

GIUDICI, P.; & FIGINI, S. *Applied Data Mining for Business and Industry* [En línea]. 2. S.l.: John Wiley & Sons, Ltd. 2009. ISBN 978-0-470-74583-0. [Consulta: 15 noviembre 2019]. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470745830.index.

GODOY VIERA, A.F. Técnicas de aprendizaje de máquina utilizadas para la minería de texto. [En línea], 2017, vol. 31, nº. 71, pp. 103-126. [Consulta: 5 marzo 2020]. ISSN 2448-8321. Disponible en: http://www.scielo.org.mx/scielo.php?pid=S0187-358X2017000100103&script=sci_arttext&tlng=en.

HARO RIVERA, S.M. Árbol De Decisión, Aplicación Con Datos Meteorológicos/Decision Tree, Application With Meteorological Data. *KnE Engineering* [En línea], 2020, pp. 37–46. [Consulta: 8 julio 2020]. ISSN 2518-6841. DOI 10.18502/keg.v5i2.6217. Disponible en: https://knepublishing.com/index.php/KnE-Engineering/article/view/6217.

HERNÁNDEZ Y., S. Herramientas para la toma de decisiones: Equipo 2: Árbol de Decisión. *Herramientas para la toma de decisiones* [En línea]. 2015. [Consulta: 20 julio 2019]. Disponible en: http://construccion3f.blogspot.com/2015/07/equipo-2-arbol-de-decision.html.

HUETE, A.; et al. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment* [En línea], 2002, vol. 83, nº. 1, pp. 195-213.

[Consulta: 10 diciembre 2019]. ISSN 0034-4257. DOI 10.1016/S0034-4257(02)00096-2. Disponible en: http://www.sciencedirect.com/science/article/pii/S0034425702000962.

HUETE, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment* [En línea], 1988, vol. 25, nº. 3, pp. 295-309. [Consulta: 4 noviembre 2019]. ISSN 0034-4257. DOI 10.1016/0034-4257(88)90106-X. Disponible en: http://www.sciencedirect.com/science/article/pii/003442578890106X.

IBARGUREN, I.; et al. BFPART: Best - FirSt PART. *Information Sciencies* [En línea], 2016, vol. 367-368, pp. 927-952. [Consulta: 8 agosto 2019]. ISSN 0020-0255. Disponible en: https://www.sciencedirect.com/science/article/pii/S0020025516305060.

INEC. Instituto Nacional de Estadística y Censos. [En línea]. 2010. [Consulta: 23 junio 2019]. Disponible en: https://www.ecuadorencifras.gob.ec/institucional/home/.

INTAVER INSTITUTE INC. Proyect Schedules and Decision Trees. *Intaver Institute* [En línea]. [Consulta: 10 noviembre 2019]. Disponible en: http://www.intaver.com/Articles/Article_DecisionTree.pdf.

JIANG, Z.; et al. Development of a two-band enhanced vegetation index without a blue band. [En línea], 2018, vol. 112, nº. 10, pp. 3833-3845. [Consulta: 7 noviembre 2019]. ISSN 0034-4257. DOI 10.1016/j.rse.2008.06.006. Disponible en: https://www.researchgate.net/publication/223925282_Development_of_a_two-band_enhanced_vegetation_index_without_a_blue_band.

KRAJEWSKI, L.J.; & RITZMAN, L.P. *Administración de operaciones: estrategia y análisis* [En línea]. México: Pearson Educación. 2000. ISBN 978-968-444-411-7. [Consulta: 12 julio 2019]. Disponible en: https://books.google.es/books?hl=es&lr=&id=B6LAqCoPSeoC&oi=fnd&pg=PA1&dq=Krajewski,+L.+J.,+y+Ritzman,+L.+P.+(2000).+Administraci%C3%B3n+de+Operaciones:+estrategia+y+analisis.+Pearson+Educaci%C3%B3n&ots=vP87CapIJZ&sig=jMmdwwP39_WEKR1QhCDNNHaKNOM#v=onepage&q&f=false.

LEYVA VÁZQUEZ, M.; et al. Facebook como herramienta para el aprendizaje colaborativo de la inteligencia artificial. [En línea], 2018, vol. 9, nº. 1, pp. 27-36. [Consulta: 25 enero 2020]. ISSN 2224-2643. Disponible en: http://refcale.uleam.edu.ec/index.php/didascalia/article/view/2565.

LOPERA, M.E. *Los árboles de decisión como herramienta para el análisis de riesgos de los proyectos* [En línea]. (Trabajo de Titulación). (Maestría). Universidad EAFIT - Escuela de Administración de Negocios. Medellín - Colombia. 2018. pp. 1-85. [Consulta: 20 abril 2019]. Disponible en: https://repository.eafit.edu.co/handle/10784/12980.

LÓPEZ DE ULLIBARRI, G.I.; & PITA FERNÁNDEZ, S. Medidas de concordancia: el índice Kappa. *Fisterra* [En línea], 2010, vol. 1, nº. 6, pp. 169-173. [Consulta: 21 marzo 2020]. Disponible en: https://www.fisterra.com/mbe/investiga/kappa/kappa.asp.

MARLON GÓMEZ, V. *Investigación del problema inverso de reconstrucción tomográfica en óptica adaptativa para astronomía a través de técnicas de minería de datos e inteligencia artificial* [En línea]. (Trabajo de Titulación). (Doctorado). Universidad de Oviedo Departamento de Exploración y Prospección de Minas. España. 2014. pp. 1-98. [Consulta: 26 enero 2020]. Disponible en: https://dialnet.unirioja.es/servlet/tesis?codigo=96621.

MATSUDO, N.L. *Árboles de decisión, una tecnica de data meaning* [En línea]. (Trabajo de Titulación). (Maestría). Universidad Nacional de Buenos Aires - Facultad de Ciencias exactas y Naturales- Departamento de Computación. Buenos Aires - Argentina. 1991. pp. 1-140. [Consulta: 19 mayo 2019]. Disponible en: https://www.google.com/search?q=Matsudo,+N.+L.+(1991).+Arboles+de+decisi%C3%B3n,+una+t%C3%A9cnica+de+data+meaning.+Buenos+Aires+,+Argentina&sxsrf=ACYBGNSeVC75NGlMgWh7Uuw-pBInH2GnTQ:1578866955151&source=lnms&tbm=vid&sa=X&ved=2ahUKEwinnYngif_mAhUFmlkKHa0NBfUQ_AUoBHoECAwQB.

MCFEETERS, S.K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing* [En línea], 1996, vol. 17, nº. 7, pp. 1425-1432. [Consulta: 6 noviembre 2019]. ISSN 0143-1161. DOI 10.1080/01431169608948714. Disponible en: https://doi.org/10.1080/01431169608948714.

MENESES TOVAR, C.L. El índice normalizado diferencial de la vegetación como indicador de la degradación del bosque. [En línea], 2012, vol. 62, nº. 238, pp. 39-46. [Consulta: 16 octubre 2020]. ISSN 0251-1584. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=3925443.

MOHAMMADI, A.; et al. Application of time series of remotely sensed normalized difference water, vegetation and moisture indices in characterizing flood dynamics of large-scale arid zone floodplains. *Remote Sensing of Environment* [En línea], 2017, vol. 190, pp. 70-82. [Consulta: 4 noviembre 2019]. ISSN 0034-4257. DOI 10.1016/j.rse.2016.12.003. Disponible en: http://www.sciencedirect.com/science/article/pii/S003442571630476X.

MONTERO, P.E. *Aprendizaje por refuerzo en espacios continuos: algoritmos y aplicación al tratamiento de la anemia renal* [En línea]. (Trabajo de Titulación). (Doctorado). Universitat de València. S.l. 2014. [Consulta: 16 febrero 2020]. Disponible en: https://dialnet.unirioja.es/servlet/tesis?codigo=90545.

NGUYEN CONG, B.; et al. Aprendizaje supervisado de funciones de distancia: estado del arte. *Revista Cubana de Ciencias Informáticas* [En línea], 2015, vol. 9, nº. 2, pp. 14-28. [Consulta: 9

enero 2020]. ISSN 2227-1899. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S2227-18992015000200002&lng=es&nrm=iso&tlng=es.

NIETO A., M.; et al. Delimitación y aálisis del Incendio forestal de Sierra de Gata (Cáceres) mediante Imágenes de los satélites Landsat 8 y Sentinel 2. *VII Congreso Forestal Español Gestión del monte: servicios ambientales y bioeconomía* [En línea]. S.l. Sociedad Española de Ciencias Forestales. 2017. pp. 1-13. ISBN 978-84-941695-2-6. Disponible en: http://secforestales.org/publicaciones/index.php/congresos_forestales/article/view/19470.

NÚÑEZ REYES, A.; et al. Agrupamiento de textos cortos en dominios cruzados. [En línea], 2016, pp. 133-145. [Consulta: 7 marzo 2020]. Disponible en: https://pdfs.semanticscholar.org/961e/d8d9af9c5cd2c786cf43fcd43183b328b704.pdf.

ORTEGA GUTIÉRREZ, C.E. *Respuesta espectral del cultivo de arroz (Oryza sativa L.) en dos fases fenológicas durante el periodo invernal 2014.* [En línea]. (Trabajo de Titulación). (Pregrado). Universidad Central de Ecuador. Quito - Ecuador. 2015. [Consulta: 16 noviembre 2019]. Disponible en: http://www.dspace.uce.edu.ec/handle/25000/7257.

ORTIZ LOZANO, J.M.; et al. Aplicación de Árboles de Clasificación a la detección precoz de abandono en los estudios universitarios de Administración y Dirección de Empresas. *Revista: Rect@. Revista Electronica de Comunicacion Y Trabajos de Asepuma, Periodo: 12, Volumen: 18, Número: , Página inicial: 177, Página final: 201* [En línea], 2017, vol. 18, pp. 177-201. [Consulta: 6 enero 2020]. ISSN 1575-605X. DOI 10.24309/recta.2017.18.2.05. Disponible en: https://repositorio.comillas.edu/xmlui/handle/11531/26523.

OTTO, M.; et al. Hydrological differentiation and spatial distribution of high altitude wetlands in a semi-arid Andean region derived from satellite data. *Hydrology and Earth System Sciences* [En línea], 2017, vol. 15, nº. 5, pp. 1713-1727. [Consulta: 5 noviembre 2019]. ISSN 1027-5606. DOI https://doi.org/10.5194/hess-15-1713-2011. Disponible en: https://www.hydrol-earth-syst-sci.net/15/1713/2011/.

PÉREZ LÓPEZ, C. *Técnicas de segmentación conceptos, herramientas y aplicaciones* [En línea]. Madrid - España: Garceta Grupo Editorial. 2011. ISBN 978-84-92812-19-6. [Consulta: 10 diciembre 2019]. Disponible en: https://www.marcialpons.es/libros/tecnicas-de-segmentacion/9788492812196/.

QUINLAN, J.R. Induction of decision trees. *Machine Learning* [En línea], 1986, vol. 1, nº. 1, pp. 81-106. [Consulta: 10 agosto 2019]. ISSN 1573-0565. DOI 10.1007/BF00116251. Disponible en: https://doi.org/10.1007/BF00116251.

QUINLAN, J.R. ***C4.5:*** *Programs for Machine Learning* [En línea]. San Mateo - California: Morgan Kaufmann Publishers. 1993. ISBN 1-55860-238-0. [Consulta: 5 agosto 2019]. Disponible en: https://books.google.com.ec/books?hl=es&lr=&id=b3ujBQAAQBAJ&oi=fnd&pg=PP1&dq=Quinlan,+J.R.+%E2%80%9CC4.5:+Programs+for+Machine+Learning%E2%80%9D.+Morgan+Kaufmann+Publishers,+1993&ots=sQ8vRMBsB2&sig=CS9fhKYbc6PkSzDrycC5VjNbkqc#v=onepage&q=Quinlan%2C%20J.R.%20%E2%80%9CC4.5%3A%20Programs%20for%20Machine%20Learning%E2%80%9D.%20Morgan%20Kaufmann%20Publishers%2C%201993&f=false.

QUINTERO, M.; & AMÉZQUITA COLLAZOS, E. Guía para el uso de «Árboles de decisión»: Alternativas de uso de la tierra para los Llanos Orientales de Colombia: Estudio de caso: Puerto López, Meta: Herramienta para la toma de decisiones. [En línea], 2003, pp. 1-46. [Consulta: 15 mayo 2019]. Disponible en: https://cgspace.cgiar.org/handle/10568/69621.

RIVERA CAMACHO, R.; et al. Modelado y propagación de valores de sentimiento en relaciones de usuario - Modelling and Propagation of Sentiment Values in Relations between Users. [En línea], 2015, vol. 107, pp. 9-17. [Consulta: 6 marzo 2020]. Disponible en: https://pdfs.semanticscholar.org/d5fe/e384773d5dcdfea7c35b5d204d6c24458cef.pdf.

ROCHE, A. *Ärboles de decisión y Series de tiempo* [En línea]. (Trabajo de Titulación). (Maestría). Universidad de la República, Faculta de Ingeniería. Montevideo - Uruguay. 2009. pp. 1-83. [Consulta: 1 abril 2019]. Disponible en: http://premat.fing.edu.uy/ingenieriamatematica/archivos/tesis_ariel_roche.pdf.

RODRÍGUEZ MORENO, V.M.; & BULLOCK, S.H. Comparación espacial y temporal de índices de la vegetación para verdor y humedad y aplicación para estimar LAI en el Desierto Sonorense. *Revista mexicana de ciencias agrícolas* [En línea], 2013, vol. 4, nº. 4, pp. 611-623. [Consulta: 15 marzo 2020]. ISSN 2007-0934. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S2007-09342013000400010&lng=es&nrm=iso&tlng=es.

RODRÍGUEZ TAPIA, S. Los métodos de aprendizaje automático supervisado en la clasificación textual según el grado de especialización. [En línea], 2018, vol. 30, pp. 131-149. [Consulta: 5 marzo 2020]. ISSN 0214-9141. DOI 10.21001. Disponible en: https://www.researchgate.net/publication/326403259_Los_metodos_de_aprendizaje_automatico_supervisado_en_la_clasificacion_textual_segun_el_grado_de_especializacion.

ROMERO, G.; & PAREDES, A. Análisis de la deserción estudiantil en la USB, facultad Ingeniería de Sistemas, con técnicas de minería de datos. [En línea], 2013, vol. 4, nº. 1, pp. 13-18. [Consulta: 25 marzo 2020]. Disponible en: http://revistas.unisimon.edu.co/index.php/identic/article/view/2484.

ROMERO ROMERO, C.A. *Estudio comparativo de algoritmos de inteligencia artificial y minería de datos enfocados a la toma de decisiones empresariales de eleccion de personal* [En línea]. (Trabajo de Titulación). (Pregrado). UDEC - Universidad de Cundinamarca - Facultad de Ingeniería - Ingeniería de Sistemas. Bogotá - Colombia. 2018. pp. 1-146. [Consulta: 25 febrero 2020]. Disponible en: http://repositorio.ucundinamarca.edu.co/handle/20.500.12558/1086. Ingeniería de Sistemas

ROUSE, J.W.; et al. Monitoring vegetation systems in the Great Plains with ERTS. [En línea], 1974, vol. 1, nº. 1, pp. 309-317. [Consulta: 6 noviembre 2019]. ISSN 19740022614. Disponible en: https://ntrs.nasa.gov/search.jsp?R=19740022614.

R-PROJECT. R: What is R? [En línea]. [Consulta: 15 enero 2020]. Disponible en: https://www.r-project.org/about.html.

RSTUDIO. RStudio. [En línea]. [Consulta: 15 enero 2020]. Disponible en: https://rstudio.com/products/rstudio/.

SÁNCHEZ MARTINEZ, I.; et al. Árboles de decisión ID3 para el diagnóstico de apendicitis aguda en ninos. [En línea], pp. 37-51. [Consulta: 11 abril 2019]. Disponible en: https://www.rcs.cic.ipn.mx/2016_113/Arboles%20de%20decision%20ID3para%20el%20diagnostico%20de%20apendicitis%20aguda%20en%20ninos.pdf.

SÁNCHEZ RODRÍGUEZ, E.; et al. Comparación del NDVI con el PVI y el SAVI como indicadores para la asignación de modelos de combustible para la estimación del riesgo de incendios en Andalucía. *Tecnologías geográficas para el desarrollo sostenible : IX Congreso del Grupo de Métodos Cuantitativos, Sistemas de Información Geográfica y Teledetección, Alcalá de Henares, septiembre de 2000* [En línea]. Sevilla - España. Universidad de Alcalá. 2000. pp. 164-174. [Consulta: 7 noviembre 2019]. Disponible en: https://idus.us.es/xmlui/handle/11441/30498.

SANTANA MANSILLA, P.F.; et al. Aplicación de Algoritmos de Clasificación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos. [En línea], 2014, vol. 17, nº. 53, pp. 57-67. [Consulta: 16 marzo 2020]. ISSN 1137-3601. Disponible en: http://ri.conicet.gov.ar/handle/11336/33712.

SERNA PINEDA, S.C. *Comparacion de Árboles de Regresión y Clasificación y regresión logística* [En línea]. (Trabajo de Titulación). (Maestría). Universidad de Colombia. Medellín - Colombia. 2009. pp. 1-60. [Consulta: 16 abril 2020]. Disponible en: https://core.ac.uk/download/pdf/11051123.pdf.

SULLA TORRES, J.; et al. Aplicación de un árbol de decisión difusa con clasificación de ambigüedad para determinar el exceso de peso en escolares. *Revista mexicana de ingeniería biomédica* [En línea], 2018, vol. 39, nº. 2, pp. 128-143. [Consulta: 20 abril 2019]. ISSN 0188-9532. DOI 10.17488/rmib.39.2.1. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S0188-95322018000200128&lng=es&nrm=iso&tlng=pt.

SURVEY, U.S.G. *Landsat 8 (L8) Data Users Handbook* [En línea]. S.l.: s.n. 2019. [Consulta: 13 noviembre 2019]. Disponible en: https://www.usgs.gov/media/files/landsat-8-data-users-handbook.

TIMARÁN PEREIRA, R.; et al. Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia. *Universidad y Salud* [En línea], 2017, vol. 19, nº. 3, pp. 388-399. [Consulta: 11 abril 2019]. ISSN 0124-7107. DOI 10.22267/rus.171903.101. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0124-71072017000300388&lng=en&nrm=iso&tlng=es.

TIMOFEEV, R. *Classification and Regression Trees (CART) Theory and Applications* [En línea]. (Trabajo de Titulación). (Maestría). Humboldt University. Berlin. 2004. pp. 1-40. [Consulta: 12 mayo 2020]. Disponible en: https://d1wqtxts1xzle7.cloudfront.net/38106508/timofeev.pdf?1436188588=&response-content-disposition=inline%3B+filename%3DClassification_and_Regression_Trees_CART.pdf.

UREÑA, R.S.; et al. Estimación de probabilidades a posterior en SVMs multiclase para reconocimiento de habla continua. [En línea], pp. 1-6. [Consulta: 17 abril 2020]. Disponible en: http://www.tsc.uc3m.es/~fernando/plantilla.pdf.

VEGA CALCINES, A. Algorítmos de aprendizaje automático: Aplicación en la solución a problemas medio ambientales. [En línea], 2014, vol. 49. [Consulta: 11 enero 2020]. Disponible en: https://pdfs.semanticscholar.org/d4f5/a4e32756b90ac7d7a8b71508afd3be59e01b.pdf.

VELA CORREA, G.; et al. Niveles de carbono orgánico total en el Suelo de Conservación del Distrito Federal, centro de México. [En línea], 2012, nº. 77, pp. 18-30. [Consulta: 15 septiembre 2019]. ISSN 0188-4611. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-46112012000100003.

VÉLEZ PAREJA, I. *Decisiones empresariales bajo riesgo e incertidumbre* [En línea]. 20. Bogotá - Colombia: Editorial Norma. 2003. ISBN 978-958-04-7441-9. [Consulta: 25 mayo 2019]. Disponible en: https://books.google.es/books?hl=es&lr=&id=mGlZ7mHPsUIC&oi=fnd&pg=PA1&ots=EOhR5sq5w5&sig=qDQHiWM8eozOREzyzazH9sgGzRY#v=onepage&q&f=false.

VILLANUEVA MORALES, J.R.; et al. Aplicación de algoritmos de clasificación para el análisis de tejido mamario y detección de cáncer de mama. [En línea], 2015, vol. 36, nº. 114, pp. 260-271. [Consulta: 7 marzo 2020]. ISSN 1405-1249. Disponible en: http://itcelaya.edu.mx/ojs/index.php/pistas/article/view/302.

WANHUI CHEN; et al. Monitoring the seasonal bare soil areas in Beijing using multitemporal TM images. *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium* [En línea]. Anchorage, AK, EE. UU. Institute of Electrical and Electronics Engineers (IEEE). 2004. pp. 3379-3382. ISBN 0-7803-8742-2. DOI 10.1109/IGARSS.2004.1370429. Disponible en: https://ieeexplore.ieee.org/document/1370429.

WEISS, G.M.; et al. Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? *Department of Computer and Information Science* [En línea], 2007, vol. 7, nº. 35-41, pp. 1-7. [Consulta: 20 noviembre 2019]. Disponible en: https://www.semanticscholar.org/paper/Cost-Sensitive-Learning-vs.-Sampling%3A-Which-is-Best-Weiss-McCarthy/9908404807bf6b63e05e5345f02bcb23cc739ebd.

WILSON, N.R.; et al. Comparison of remote sensing indices for monitoring of desert cienegas. *Arid Land Research and Management* [En línea], 2016, vol. 30, nº. 4, pp. 460-478. [Consulta: 3 noviembre 2019]. ISSN 1532-4982. DOI 10.1080/15324982.2016.1170076. Disponible en: https://www.tandfonline.com/doi/abs/10.1080/15324982.2016.1170076?needAccess=true#aHR0cHM6Ly93d3cudGFuZGZvbmxpbmUuY29tL2RvaS9wZGYvMTAuMTA4MC8xNTMyNDk4Mi4yMDE2LjExNzAwNzY/bmVlZEFjY2Vzcz10cnVlQEBAMA==.

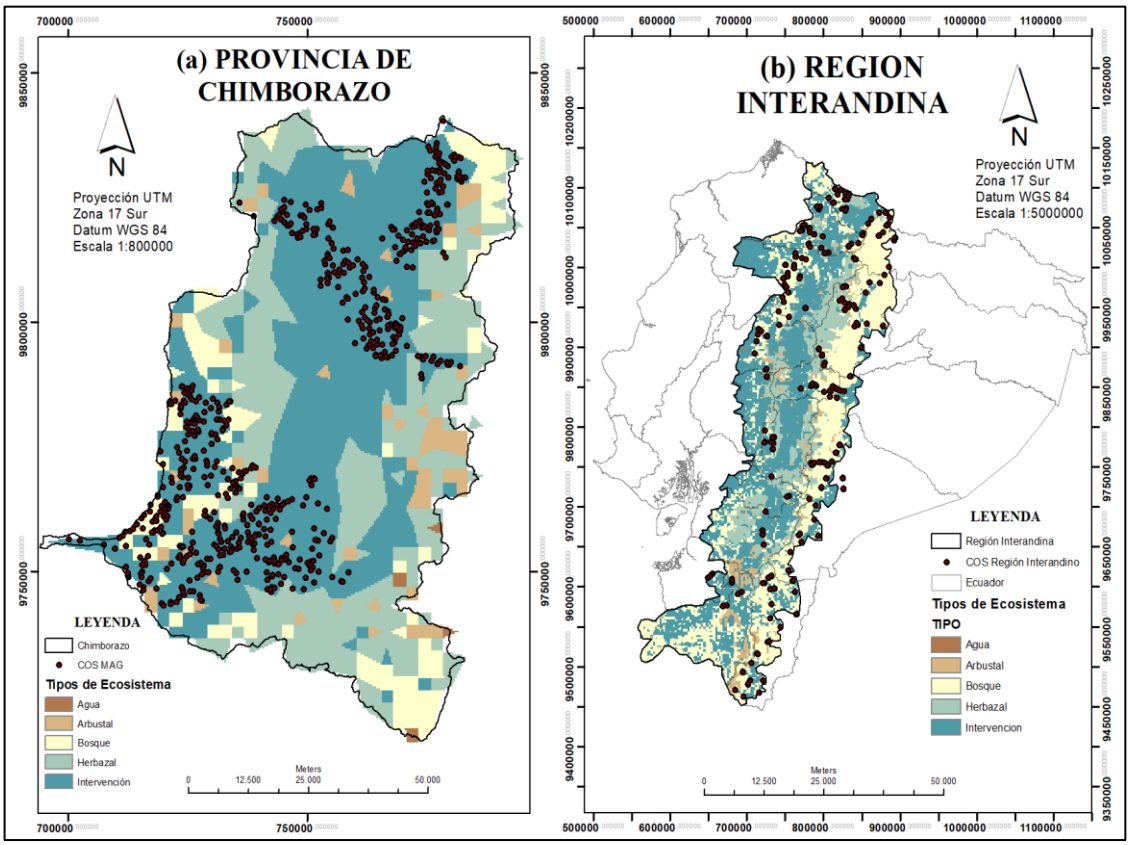
WRITTEN, I.H.; et al. *Data Mining practical Pachine Learning Tools and Techniques* [En línea]. 4. S.l.: s.n. 2016. ISBN 978-0-12-804357-8. [Consulta: 15 diciembre 2019]. Disponible en: https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5.

YATES, H.W.; et al. The role of meteorological satellites in agricultural remote sensing. [En línea], 1984, vol. 14, nº. 3, pp. 219-233. [Consulta: 18 octubre 2019]. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/0034425784900178.

ZHAO, Y. *R and Data Mining: Examples and Case Studies* [En línea]. S.l.: Academic Press. 2015. ISBN 978-0-12-397271-2. [Consulta: 14 marzo 2020]. Disponible en: https://books.google.es/books?hl=es&lr=&id=FEOh08LBD9UC&oi=fnd&pg=PR1&dq=Y.+Zhao,+%E2%80%9C+and+Data+Mining%E2%80%AF:+Examples+and+Case+Studies,%E2%80%9D+no.+December+2012,+2015.&ots=NPy5AzBGC1&sig=-y9MuTUPshKJxNndaezkjujjfo#v=onepage&q&f=false.

ANEXOS

ANEXO A: PUNTOS DE MUESTREO DE CABONO ORGÁNICO DEL SUELO



Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

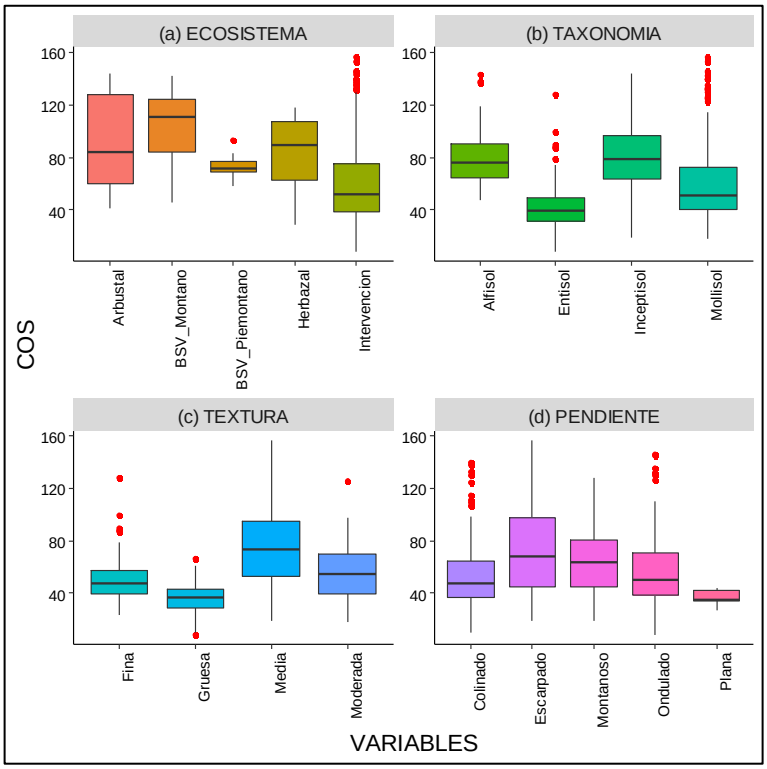
ANEXO B: MUESTRAS DE CARBONO ORGÁNICO DEL SUELO POR TIPO DE ECOSISTEMA

TIPOS DE ECOSISTEMA	PROVINCIA DE CHIMBORAZO	REGIÓN INTERANDINA
ARBUSTAL	23	8
BOSQUE	44	281
HERBAZAL	19	13
INTERVENCIÓN	505	108

Realizado por: Padilla S. Oscar R., 2020

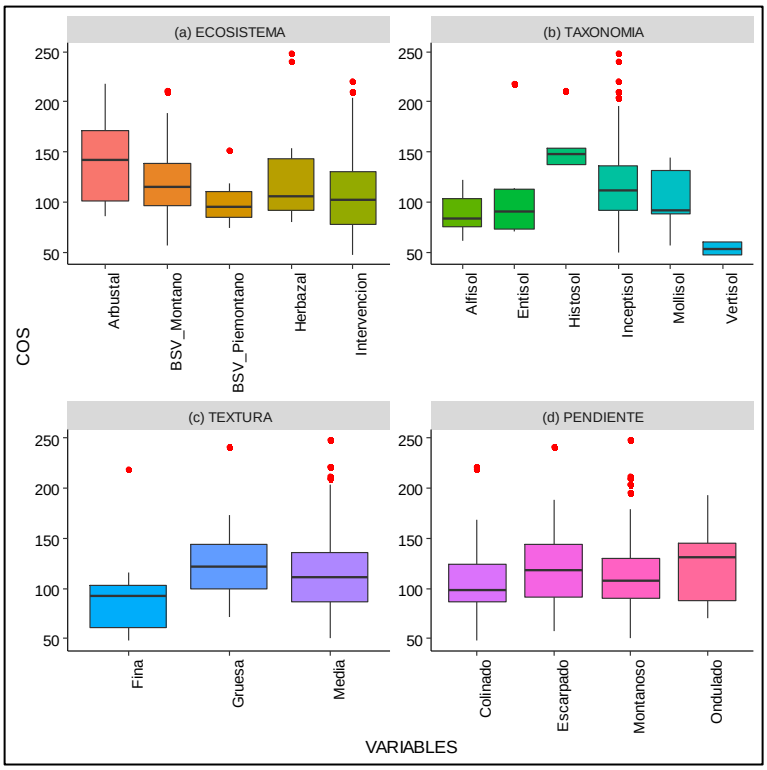
ANEXO C: ATÍPICOS POR VARIABLES CATEGÓRICOS

FAO - PROVINCIA DE CHIMBORAZO



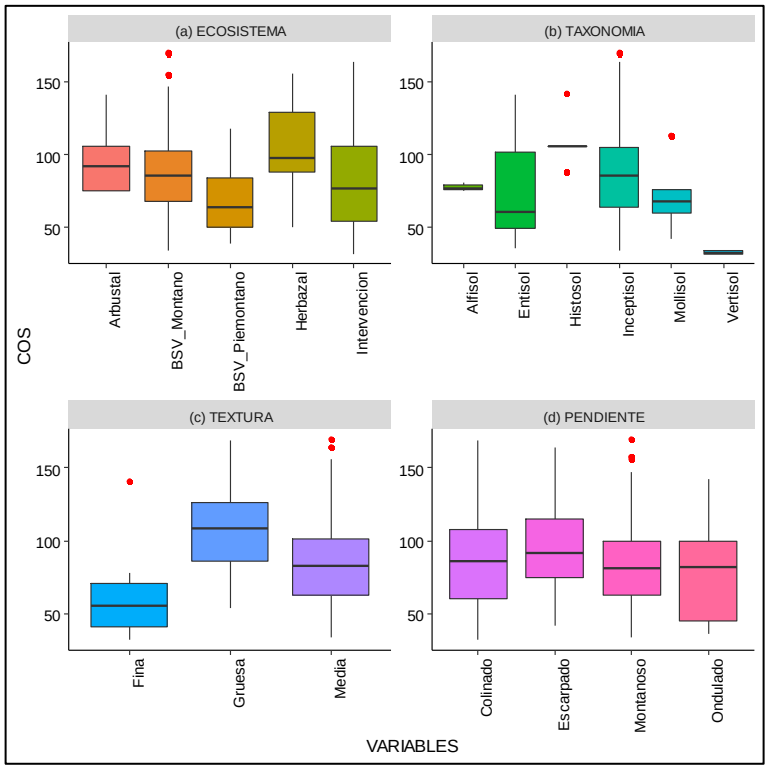
Realizado por: Padilla S. Oscar R., 2020.

FAO - REGIÓN INTERANDINA



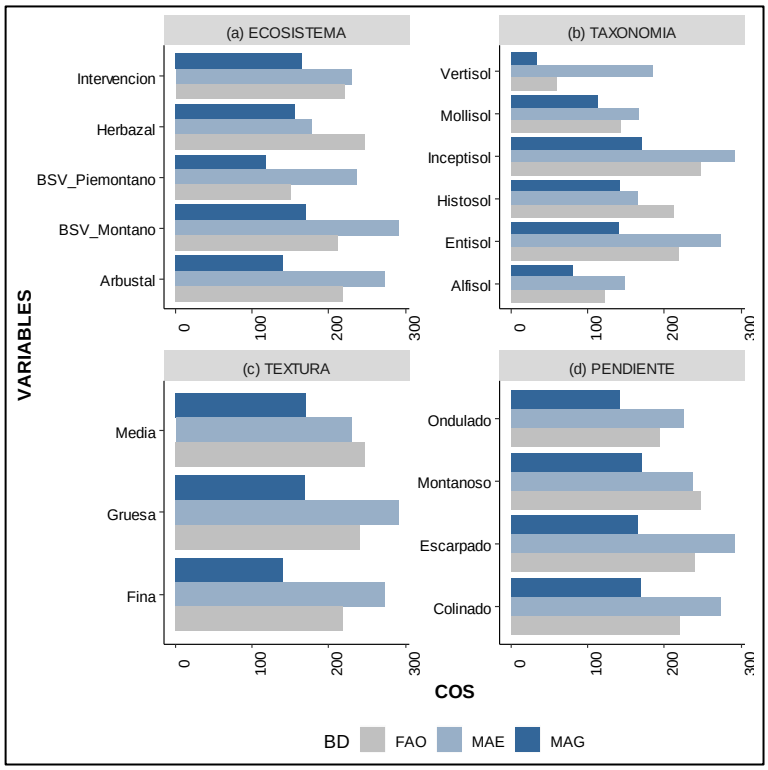
Realizado por: Padilla S. Oscar R., 2020.

MAG - REGIÓN INTERANDINA



Realizado por: Padilla S. Oscar R., 2020.

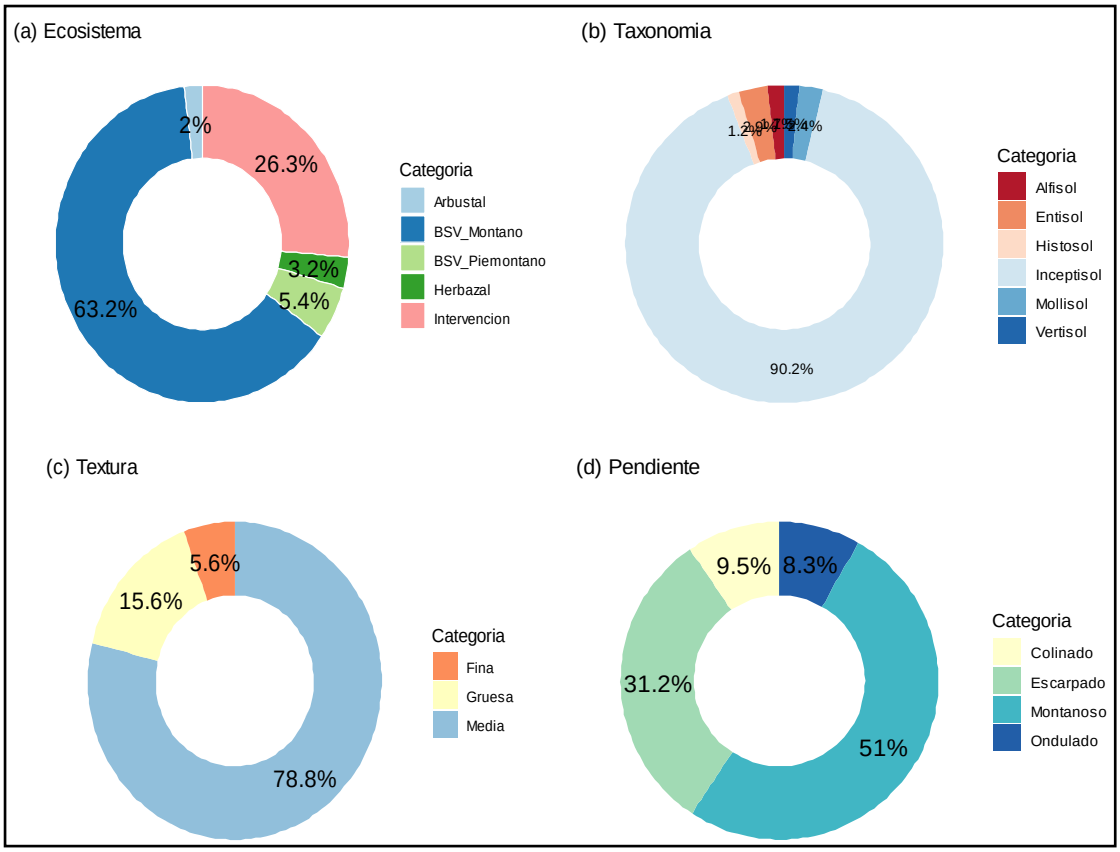
ANEXO D: CONTENIDOS DE CARBONO ORGÁNICO DEL SUELO MÁXIMOS EN LA REGIÓN INTERANDINA



Realizado por: Padilla S. Oscar R., 2020

ANEXO E: DIAGRAMA DE PASTEL DE LA DISTRIBUCIÓN ESTADÍSTICA DE FRECUENCIA POR VARIABLES DE LA REGIÓN INTERANDINA

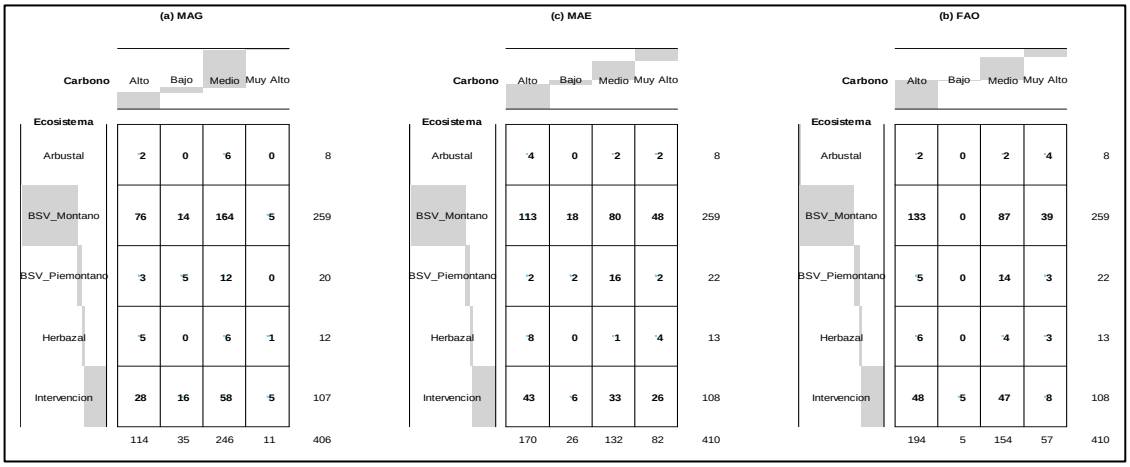
ECOSISTEMA



Realizado por: Padilla S. Oscar R., 2020

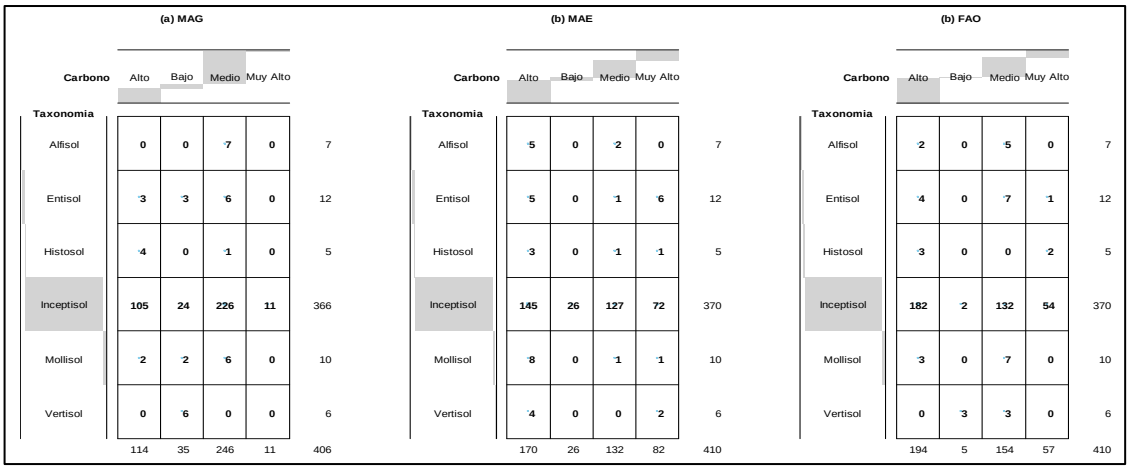
ANEXO F: ANÁLISIS DE CORRESPONDENCIA DE CARBONO ORGÁNICO EN LOS DE LA REGION INTERANDINA

ECOSISTEMA



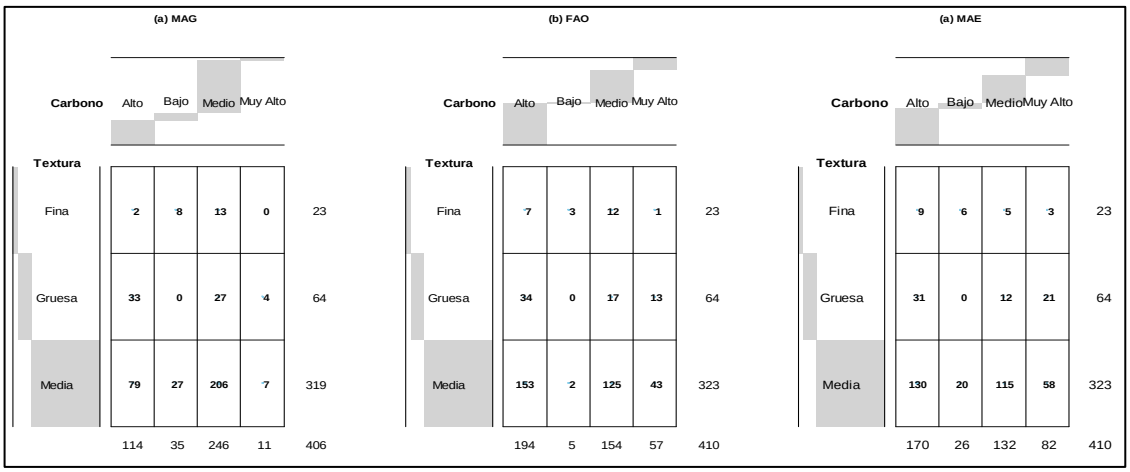
Realizado por: Padilla S. Oscar R., 2020

TAXONOMÍA



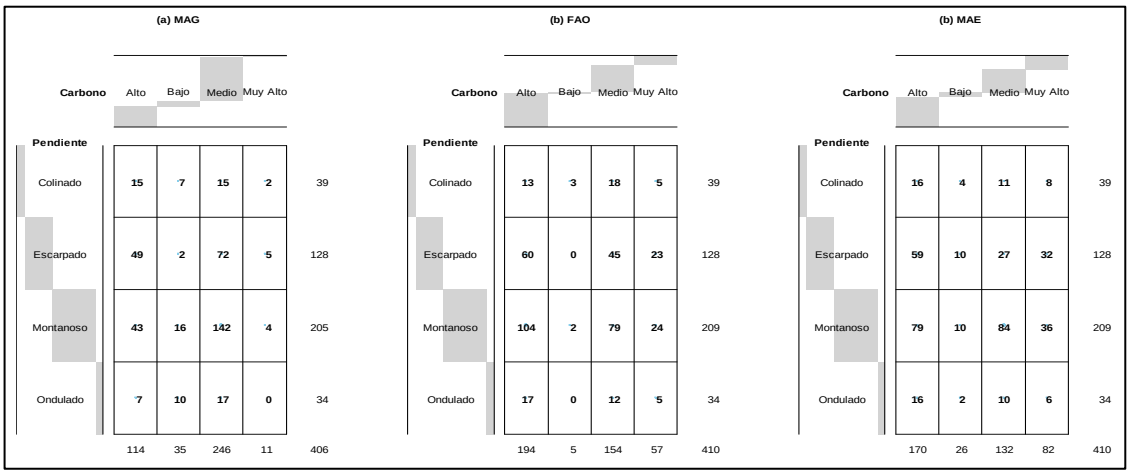
Realizado por: Padilla S. Oscar R., 2020

TEXTURA



Realizado por: Padilla S. Oscar R., 2020

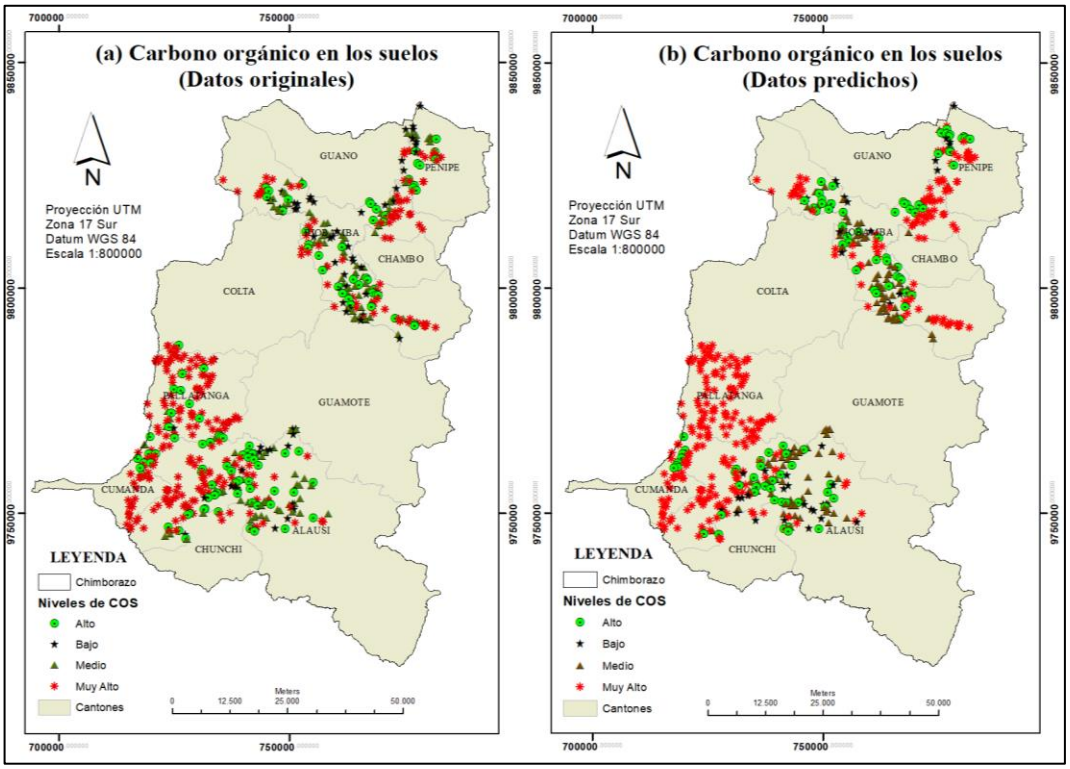
PENDIENTE



Realizado por: Padilla S. Oscar R., 2020

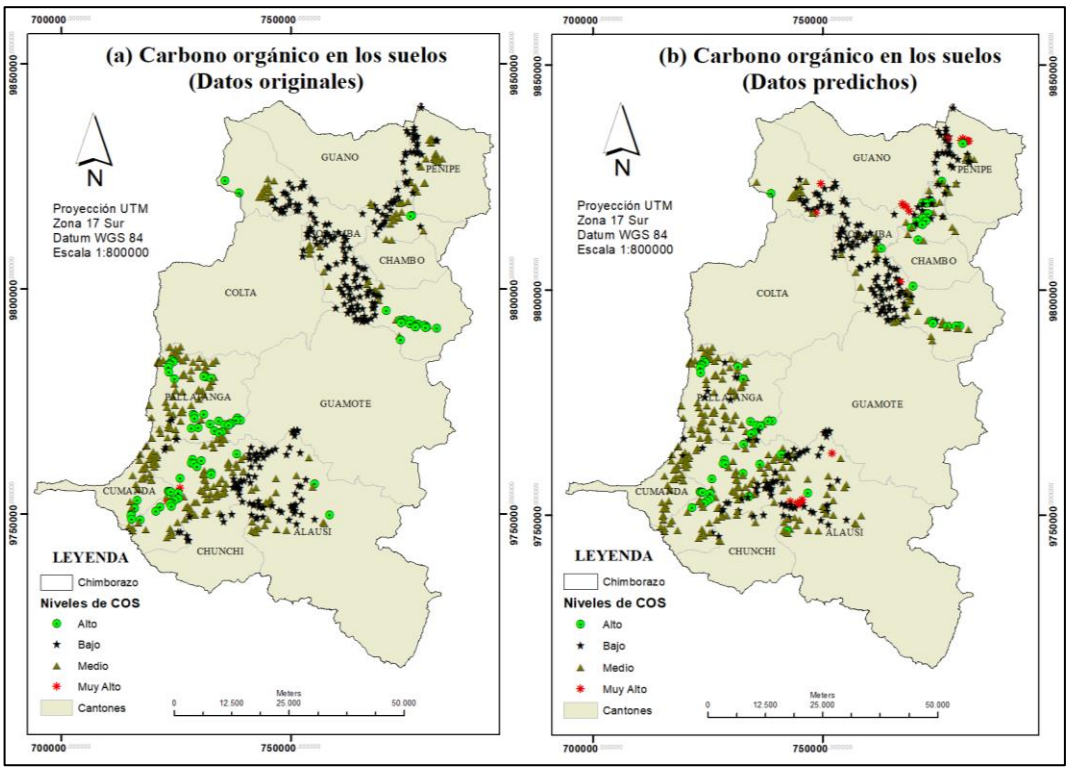
ANEXO G: CARBONO EDÁFICO EN LA PROVINCIA DE CHIMBORAZO

MAG



Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

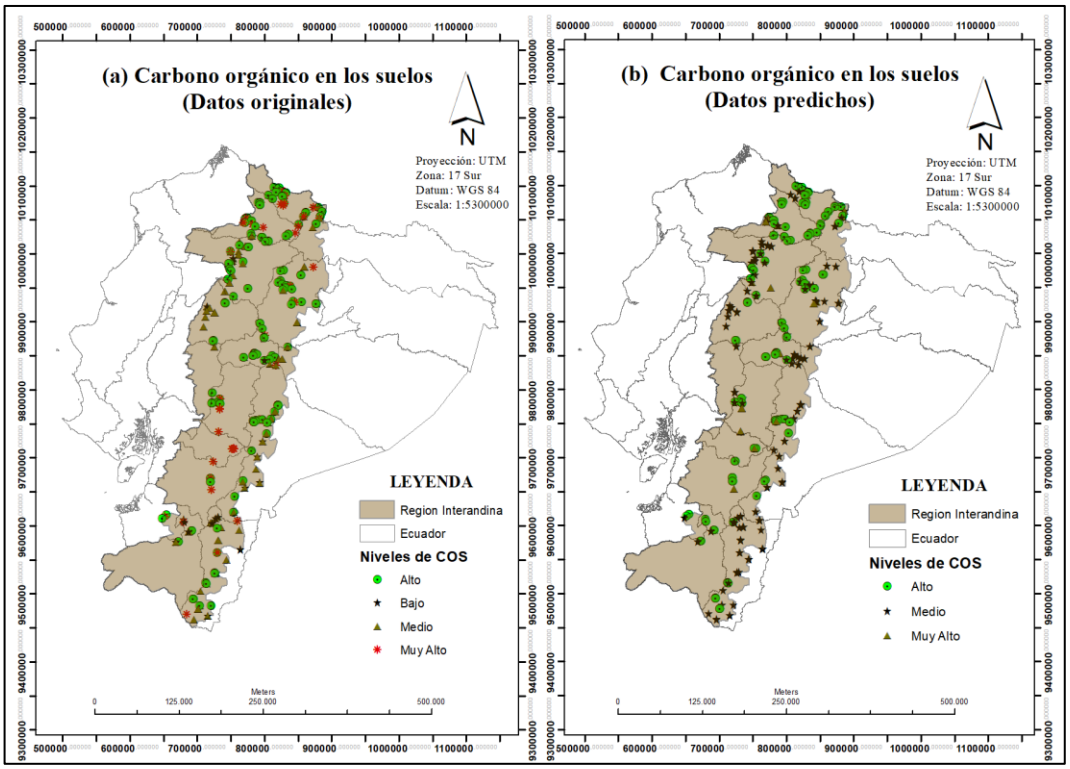
FAO



Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

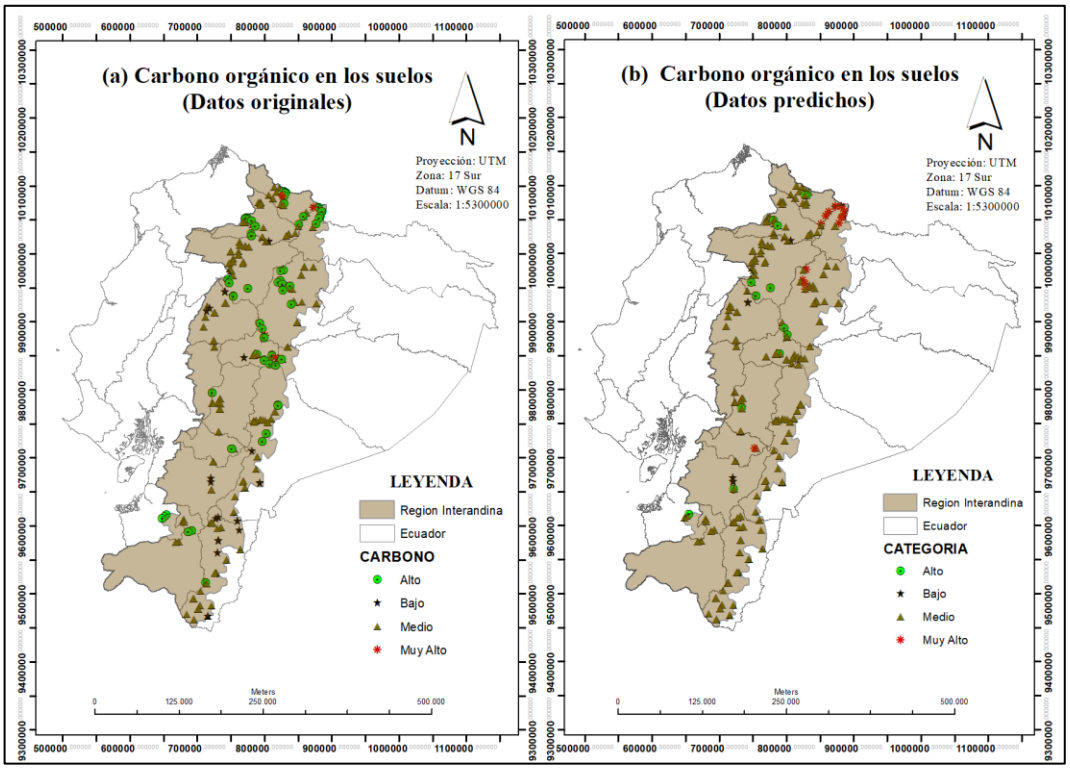
ANEXO H: CARBONO EDÁFICO EN LA REGION INTERANDINA

MAE



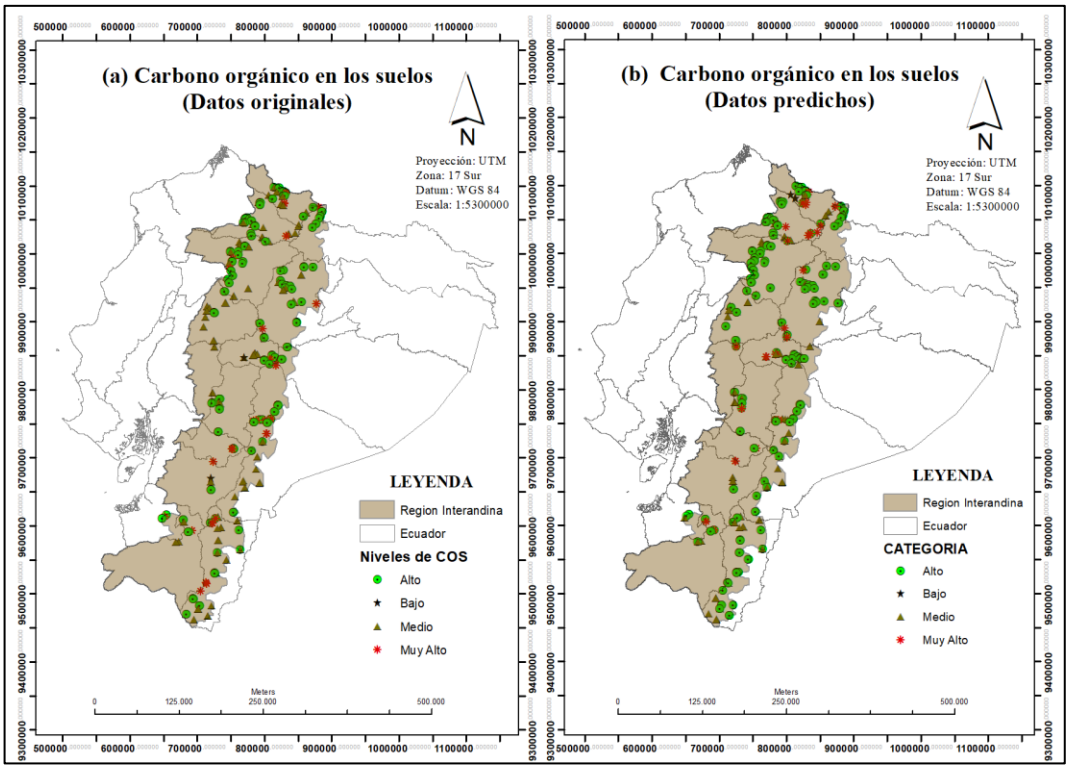
Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

MAG



Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

FAO



Realizado por: Padilla S. Oscar R., 2020 - GIDAC.

ANEXO I: AVAL DE LA INVESTIGACIÓN

MINISTERIO DEL AMBIENTE

CERTIFICADO

Yo: Abogada Jessica Estefanía Coronel Carvajal en calidad de **DIRECTORA NACIONAL FORESTAL** certifico:

Que el señor **OSCAR ROBERTO PADILLA SEFLA** con CI: 060471015-2, estudiante de la carrera de Ingeniería en Estadística Informática de la Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Escuela de Física y Matemática va a desarrollar el trabajo de titulación denominado: **"ANALISIS DE ARBOLES DE DECISION PARA LA VALORACION DEL CARBONO EDAFICO DE LA PROVINCIA DE CHIMBORAZO MEDIANTE EL USO DE VARIABLES DE EVALUACION NACIONAL FORESTAL MAE-FAO"**, con los datos de la provincia de Chimborazo proporcionados por esta dependencia.

El interesado señor Oscar Padilla puede utilizar el presente certificado con la finalidad de proceder con la matrícula de su proyecto de titulación para lo cual la Dirección Nacional Forestal compromete su apoyo para la ejecución del trabajo en mención.

Quito, 9 de julio del 2019

Abg. Jessica Estefanía Coronel Carvajal
DIRECTORA NACIONAL FORESTAL

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

DIRECCIÓN DE BIBLIOTECAS Y RECURSOS PARA EL APRENDIZAJE Y LA INVESTIGACIÓN

UNIDAD DE PROCESOS TÉCNICOS

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 17/09/2020

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: Oscar Roberto Padilla Sefla
INFORMACIÓN INSTITUCIONAL
Facultad: Ciencias
Carrera: Ingeniería en Estadística Informática
Título a optar: Ingeniero en Estadística Informática
f. Analista de bibliotecas responsable: Lic. Luis Caminos Vargas Mgs.

LUIS ALBERTO CAMINOS VARGAS

Firmado digitalmente por LUIS ALBERTO CAMINOS VARGAS
Nombre de reconocimiento (DN): c=EC, l=RIOBAMBA, serialNumber=0602766974, cn=LUIS ALBERTO CAMINOS VARGAS
Fecha: 2020.09.17 11:43:15 -05'00'

0266-DBRAI-UPT-2020