LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Segmentación de suelos de acuerdo con sus características fisicoquímicas a través modelos de aprendizaje automático

Ortiz, J. C.[1]; González, J. J.[2]
[1]Fundación Universitaria Los Libertadores jcortizr01@libertadores.edu.co
[2]Fundación Universitaria Los Libertadores jjgonzalezv02@libertadores.edu.co

Resumen
La segmentación de la calidad fisicoquímica del suelo permite establecer zonas que requieren manejos similares o zonas con vulnerabilidades en las que se deben enfocar estrategias para su conservación y/o recuperación; en este sentido, se tomaron 1139 muestras de suelos en predios en la zona rural de los municipios de Córdoba, Cuaspud, Iles, Ipiales y Potosí del departamento de Nariño a las que se les realizó análisis fisicoquímicos (contenidos de arenas, limos y arcillas, pH, conductividad eléctrica, contenido de materia orgánica, nitrógeno, fosforo intercambiable, azufre, calcio, magnesio, potasio, capacidad de intercambio catiónico efectiva, aluminio, hierro, manganeso, cobre, zinc, boro, saturación de aluminio, saturación de magnesio, saturación de potasio, saturación de calcio, relación calcio y magnesio, relación calcio y potasio, relación magnesio y potasio, relación calcio, magnesio y potasio); para establecer como se podrían segmentar estas muestras, inicialmente se realizó una correlación de Pearson para conocer relaciones lineales entre variables, para luego implementar un análisis de componentes principales (PCA); con esta información se aplicó varios modelos de aprendizaje no supervisado para determinar el número óptimo de clusters en los que segmentar la información; posteriormente, se decidió realizar un modelo supervisado, Random Forest (RF), teniendo en cuenta la información del PCA y de clusters, para determinar las variables originales con mayor importancia relativa en el agrupamiento de la información; finalmente se logró establecer las variables y los valores de estas que permitían el agrupamiento aplicando el modelo Decision Tree (DT); en este sentido, se logró establecer que la mejor forma de segmentar la información de las muestras de suelo es a través de tres clusters, y que las variables que mayor peso tienen en la generación de estos grupos son los contenidos de Arena y Limo, la relación Ca/Mg/K y la relación Ca/K, denotando diferencias principalmente entre suelos Franco arenosos y franco .

1. Introducción

Para lograr rendimientos óptimos en la producción de cultivos, se debe tener en cuenta la disponibilidad de nutrientes en el suelo, para así ajustar planes de fertilización adecuados. Para lograr rendimientos óptimos, se debe tener en cuenta que la producción agrícola es resultado de la interacción de factores bióticos y abióticos, siendo la fertilización uno de los más determinantes. Las plantas requieren de nutrientes esenciales para completar su ciclo biológico. En este sentido, resulta necesario asegurar su suficiencia al momento de establecer las plantaciones, ya que un inadecuado manejo de la nutrición puede generar pérdidas de rendimientos (Zoppolo & Fasiolo, 2016).

A pesar de la importancia de la práctica de fertilización, actualmente, muchos productores realizan la aplicación de fertilizantes sin tener en cuenta ningún tipo de análisis, lo cual, genera pérdidas económicas y efectos negativos en el ambiente, por el

uso innecesario o excesivo de éstos. Adicionalmente, se deben considerar las interacciones que los organismos puedan generar, ya que estas son indispensables en la disponibilidad de los nutrimentos para la planta. Debido a la necesidad de fortalecer la agricultura de precisión, se han desarrollado herramientas, que permiten el diagnóstico del estado nutricional de los cultivos por sitio específico, como lo son los análisis de suelos, permitiendo analizar características del sistema suelo-planta, para así lograr un manejo racional de la fertilidad (Castañeda, Jaramillo, & Cotes, 2014).

La evaluación de la calidad del suelo, junto con su fertilidad, es sumamente importante para tomar decisiones que permitan optimizar la producción de alimentos y así contribuir a la capacidad de adaptación al cambio climático. Los métodos de laboratorio convencionales permiten caracterizar la física, química y biología de los suelos. Sin embargo, la exploración en uso de métodos novedosos permite, eventualmente, contar con métodos más rápidos, precisos y accesibles económicamente para evaluar la calidad del suelo. Actualmente se utilizan muestras de suelo y métodos de laboratorio convencionales como la Espectroscopia de Absorción Atómica o Espectroscopia de emisión por plasma de acoplamiento inductivo y colometría para cuantificar el contenido de nutrientes en suelo. Esos métodos convencionales involucran una serie de pasos e insumos necesarios para poder obtener resultados.

El sector agrícola, tiene la necesidad de mejorar los procesos de producción para ser más eficientes con el uso de los recursos que nos provee el ecosistema como son los nutrientes del suelo; y con esto disminuir costos. Pero lograr esta eficiencia, se requiere de un sistema que permitan integrar nuevas tecnologías, para generar nuevas variables que permitan la construcción de modelos predictores para la toma de decisiones (Moreno-Carriles, 2018). Por otra parte, la gestión de un sistema agrícola involucra actividades relacionados a planificar, organizar, controlar y dirigir los recursos como los nutrientes del suelo (Fernández, Fernández, Rivera, & Calero, 2016) (Fajardo, Aguilar, Flores, Parra, & Acurio, 2017). En este aspecto, el requerimiento nutricional es diferente en cada especie vegetal, un desbalance de los nutrientes por causas naturales o por un mal manejo agrícola, tiene efectos negativos en la producción final y por ende afecta en el rendimiento financiero de la empresa (Espinoza-Freire & Tinoco-Cuenca, 2015) (Rodriguez & Fusco, 2017).

El aprendizaje automático, nace de la inteligencia artificial como apoyo al entrenamiento de modelos que respondan a la predicción de los datos en diferentes ámbitos de la ciencia (Baviera, 2017). Su estructura requiere el análisis de diferentes algoritmos que permitan evaluar a un conjunto de datos a fin de establecer si el problema es de clasificación o de regresión. Para la resolución de problemas de predicción se pretende a través de diversos algoritmos, entrenarlos con la finalidad de lograr que el modelo sea capaz de predecir el nuevo conjunto. La aplicación de aprendizaje automático emplea métodos supervisados y no supervisados, como apoyo a los procedimientos de análisis de los datos. La fase de predicción de los datos es el resultado del entrenamiento del modelo y el análisis de los hiperparametros que permiten dar mayor rigurosidad a los datos (Goya, Barquero, & Figuera, 2017). Su aporte genera elementos suficientes para dar solución estadística a los problemas que requieran la utilización de estas técnicas.

En este sentido, generar segmentaciones de suelos de acuerdo con sus características fisicoquímicas, permitirá establecer suelos homogéneos donde las estrategias de manejo de la fertilización sean similares, y adicionalmente, establecer cuales características fisicoquímicas permiten segmentar mejor los suelos, lo que podría reducir los costos al ser menos las variables a evaluar; por otro lado, permitirá establecer suelos con condiciones edáficas más vulnerables y generar planes de manejo acordes.

2. Marco Teórico

2.1. El suelo

El suelo es un bien vital para la sociedad humana y para la regulación del ambiente. Es un bien finito (es limitado, se agota), que se encuentra constituido principalmente por minerales, aire, agua, materia orgánica, macro, meso y microorganismos que desempeñan procesos fundamentales de tipo biótico y abiótico, cumpliendo funciones indispensables para la sociedad y el planeta (Ministerio de Ambiente y Desarrollo Sostenible, 2016).

El suelo es indispensable y determinante para la estructura y el funcionamiento de los ciclos del agua, del aire y de los nutrientes, así como para la biodiversidad, debido a que hace parte fundamental de los ciclos biogeoquímicos (ciclo de los nutrientes), en los cuales hay distribución, transporte, almacenamiento y transformación de materiales y energía necesarios para el desarrollo y sostenimiento de la vida en el planeta (Van Miegrot & Johnsson, 2009) (Martin, 1998). Es igualmente fundamental para la tierra, el territorio y las culturas; da soporte a la vida y a las actividades humanas permitiendo garantizar los derechos ambientales de las generaciones presentes y futuras. Sin embargo, el suelo se puede deteriorar y luego que esto ocurre, su recuperación es difícil, costosa, toma mucho tiempo y en algunos casos, no es posible recuperarlo hasta su estado inicial (Ministerio de Ambiente y Desarrollo Sostenible, 2016).

Así mismo, el suelo es un soporte para las plantas, los bosques y la biodiversidad, y está relacionado con el adecuado uso y equilibrio de los recursos agua y aire. Por ende, retomando la definición de gobernanza de los recursos naturales de la FAO, la gobernanza del suelo es un proceso dinámico y participativo que se traduce en arreglos institucionales y sociales para la toma de decisiones e implementación de estas, garantizando los derechos de las partes interesadas y su manejo, uso y conservación. De esta manera, se integran las dimensiones social, ambiental, económica, política y cultural del recurso suelo (FAO, 2015b). En ese sentido, el suelo debe ser visto de manera integral, no solo con todos sus componentes y funciones, sino en sus interrelaciones con los otros elementos o componentes del ambiente (entre ellos el subsuelo, las plantas, el agua, el aire, etc.), considerando las dimensiones social, ambiental, económica, política y cultural y ello debe traducirse, entre otras, en políticas y normas, acordes con el principio de integralidad (Ministerio de Ambiente y Desarrollo Sostenible, 2016).

La gestión adecuada del suelo constituye un factor esencial para la práctica de una agricultura sostenible y proporciona también una valiosa base para regular el clima y conservar los servicios ecosistémicos y la biodiversidad. Los suelos saludables son un

requisito previo básico para satisfacer las diversas necesidades de alimentos, biomasa (energía), fibra, forraje y otros productos, y para garantizar la prestación de los servicios ecosistémicos esenciales en todas las regiones del mundo (FAO, 2015a).

El suelo es indispensable y determinante para la estructura y el funcionamiento de los ciclos del agua, del aire y de los nutrientes. Las funciones específicas que un suelo proporciona se rigen en gran medida por el conjunto de propiedades químicas, biológicas y físicas que se hallan en dicho suelo. Así mismo, los suelos son una reserva clave de biodiversidad mundial que abarca desde los microorganismos hasta la flora y la fauna. Esta biodiversidad tiene una función fundamental en el respaldo a las funciones del suelo y, por tanto, a los bienes y servicios ecosistémicos asociados con los suelos (FAO, 2015a).

2.1.1. Servicios ecosistémicos que genera el suelo

Los servicios ecosistémicos se definen como aquellos procesos y funciones de los ecosistemas que son percibidos por los seres humanos como un beneficio (de tipo ecológico, cultural o económico) directo o indirecto. Incluyen aquellos de aprovisionamiento, como comida y agua; servicios de regulación, como la regulación de las inundaciones, sequías, degradación del terreno y enfermedades; servicios de sustento como la formación del sustrato y el reciclaje de los nutrientes; y servicios culturales, ya sean recreacionales, espirituales, religiosos u otros beneficios no materiales (Ministerio de Ambiente y Desarrollo Sostenible, Política Nacional para la Gestión Integral de la Biodiversidad y sus Servicios Ecosistémicos – PNGIBSE, 2012). La producción de alimentos, por ejemplo, depende de la disponibilidad y calidad del suelo, pues el 95% de los alimentos provienen del mismo (FAO, 2015c). De igual manera, el suelo constituye el principal reservorio de agua dulce del planeta y es determinante para la regulación de la cantidad y calidad del agua suministrada en por el ambiente. De ahí el concepto de cuenca hidrográfica como expresión de la vinculación y dependencia que existe entre estos dos recursos naturales.

Otro de los servicios ecosistémicos esenciales del suelo es la captura de carbono, que se estima en dos tercios del carbono fijado en el planeta, el cual, aunque es un factor difícil de valorar económicamente (comparado por ejemplo con el carbono fijado por los bosques) es un servicio fundamental para el mantenimiento del equilibrio ecológico en el planeta.

2.1.2. Calidad del suelo

La calidad del suelo se define como la capacidad específica que tiene un suelo para funcionar en un ecosistema natural o antrópico de acuerdo con sus funciones: (1) promover la productividad del sistema sin perder sus propiedades físicas, químicas y biológicas (productividad biológica sostenible); (2) atenuar contaminantes ambientales y patógenos (calidad ambiental); y (3) favorecer la salud de plantas, animales y humanos (Doran & Parkin, 1994). De esta manera, los servicios ecosistémicos asociados al suelo están directamente relacionados con su calidad.

2.1.3. Vocación de uso del suelo

La vocación de uso del suelo se refiere a la clase mayor de uso que una unidad de suelo está en capacidad natural de soportar con características de sostenibilidad, evaluada sobre una base biofísica. Está subdividida en cinco (5) clases: agrícola, ganadera, agroforestal, forestal y de conservación (IGAC, 2012).

Según el (IGAC, 2012), los suelos de Colombia tienen una vocación principalmente forestal, con 56,23% del territorio. En segundo lugar, aparece la vocación agrícola que representa un 19,34% y en tercer lugar se encuentran los suelos con vocación ganadera con 13,31% del país. Los suelos con vocación para la conservación y recuperación tienen una extensión equivalente al 5,52% del país y corresponden principalmente a la conservación de los recursos hídricos e hidrobiológicos y a la recuperación. Los suelos con vocación agrosilvopastoril corresponden a 3,55% del total nacional independientemente de los usos silvopastoril y agrosilvícola (IGAC, 2012).

2.1.4. Uso actual de los suelos de Colombia

Según el (IGAC, 2012) el área continental se encuentra cubierta en su mayoría por bosques, en un 53% del área total, seguido por los territorios ganaderos que corresponden a un 31%, y los territorios agrícolas a un 5%.

De acuerdo con el (IGAC, 2012) el uso adecuado del suelo en Colombia es del 68 %. El conflicto de uso por subutilización del suelo corresponde al 13% del territorio nacional, y se presenta en suelos donde la demanda es menor a la capacidad productiva de los suelos. El conflicto de uso por sobreutilización del suelo corresponde al 16% del territorio nacional e incluye los suelos donde los agroecosistemas tienen un aprovechamiento intenso, sobrepasando su capacidad productiva. Particularmente, en el uso agrícola se evidencia la subutilización del suelo, ya que la vocación equivale a 22 millones de hectáreas y el uso es de solamente 5 millones de hectáreas en Colombia. En el uso ganadero se evidencia una sobreutilización, ya que la vocación corresponde solamente a 15 millones de hectáreas y el uso de territorios ganaderos corresponde a 34 millones de hectáreas. Respecto al uso forestal es similar, la vocación es de 64 millones de hectáreas y el uso es de 65 millones de hectáreas.

2.1.5. Degradación de suelos por erosión en Colombia

La degradación de los suelos se refiere a la disminución o alteración negativa de una o varias de las ofertas de bienes, servicios y/o funciones ecosistémicas y ambientales, ocasionada por procesos naturales o antrópicos que, en casos críticos, pueden originar la pérdida o la destrucción total del componente ambiental (MAVDT & IDEAM, 2004).

La degradación de suelos puede ser física, química o biológica. En la degradación física se presenta la erosión y la compactación, en la degradación química se presenta la salinización, la acidificación/ alcalinización y la contaminación. La degradación biológica se evidencia por la pérdida de la materia orgánica, el desequilibrio de la actividad biológica y procesos de mineralización del suelo (Minambiente & IDEAM, 2015). La definición de erosión según el Protocolo de Degradación de Suelos por Erosión es “la pérdida de la capa superficial de la corteza terrestre por acción del agua y/o del viento,

que es mediada por el hombre, y trae consecuencias ambientales, sociales, económicas y culturales" (Minambiente & IDEAM, 2015).

Según el estudio de línea base de degradación de suelos por erosión (Minambiente & IDEAM, 2015) el 40% de la superficie continental de Colombia presenta algún grado de erosión. El grado de erosión ligera corresponde al 20%, el grado moderado al 17%, el grado severo al 2,7% y el grado muy severo corresponde al 0,24% del territorio nacional (Minambiente & IDEAM, 2015). Por otro lado, los departamentos más afectados por la magnitud de la erosión, es decir, por la suma de áreas de erosión ligera, moderada, severa y muy severa respecto al área del departamento son Cesar, Caldas, Córdoba, Cundinamarca, Santander, La Guajira, Atlántico, Magdalena, Sucre, Tolima, Quindío, Huila y Boyacá (Minambiente & IDEAM, 2015). Los departamentos más afectados por severidad de la erosión, es decir, por la suma de las áreas con erosión severa y muy severa en relación con el área del departamento son: La Guajira, Magdalena, Cesar, Huila, Sucre, Santander, Tolima, Boyacá, Atlántico, Norte de Santander y Valle del Cauca (Minambiente & IDEAM, 2015). En este sentido y según lo señala la Política para la Gestión Sostenible del Suelo es urgente promover la conservación del suelo, entendida como el mantenimiento de sus múltiples funciones a través de acciones de generación de conocimiento, preservación, restauración, manejo y uso sostenible del suelo (Ministerio de Ambiente y Desarrollo Sostenible, 2016).

Algunos ejemplos generales de prácticas no sostenibles de los suelos que pueden estar generando pérdida de los servicios ecosistémicos son:

- Uso inadecuado del suelo con relación a su vocación de uso.
- Excesiva o inadecuada mecanización agrícola.
- Uso excesivo de fertilizantes y plaguicidas de síntesis.
- Quemas y tala de bosques.
- Excesiva utilización de recursos hídricos y destrucción de microclimas.
- Presencia de monocultivos.
- Sobrepastoreo.
- Sobreutilización de suelos que se encuentran en áreas protegidas o bajo diferentes figuras de conservación.

2.2. Segmentación de suelos

Para hacer un buen uso y manejo de los suelos es necesario saber cuáles son, cómo son, dónde están y que superficie ocupan. Por esta razón, se han realizado esfuerzos para clasificar el suelo. La clasificación del suelo es necesaria para predecir su comportamiento e identificar limitantes que permitan tomar decisiones adecuadas de manejo en los ámbitos agrícola, pecuario, forestal, urbano, ambiental y de salud. Clasificar es una herramienta que nos permite identificar mejores usos, estimar su productividad, fomentar la investigación y es un medio de comunicación no sólo para especialistas en génesis del suelo, sino para todos los edafólogos quienes a su vez interactúan con otros científicos o gente cuya actividad se relaciona directa o indirectamente con el suelo. Las primeras clasificaciones de suelos se basaron en la textura, los suelos se clasificaban como francos, arcillosos, arenosos e incluso si eran orgánicos. Otro factor importante era el material parental, es decir a partir del cual se

habían originado, denominando a los suelos, por ejemplo: calcáreos, graníticos, arenosos, etc.

La clasificación del suelo es una tarea que demanda capacidad de cómputo, debido a la gran cantidad de datos que se tiene que procesar, por lo cual, se establece la necesidad de aplicar métodos computacionales suficientemente rápidos y efectivos. El aprendizaje computacional es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas, métodos y algoritmos que permiten que los computadores aprendan a reconocer patrones a partir de datos de un modelo de inferencia con propósitos predictivos o de segmentación (Suárez, Jiménez, Castro-Franco, & Cruz-Roa, 2017).

Entre los tipos de algoritmos de aprendizaje automático se destacan dos, el aprendizaje supervisado y el aprendizaje no supervisado. El aprendizaje supervisado ocurre cuando se proporciona al modelo de entrenamiento datos conocidos, es decir cuando se sabe a qué clase pertenece cada uno (Wang, y otros, 2014). En los algoritmos de aprendizaje no supervisado el modelo de entrenamiento está formado por entradas y no se tiene información a qué clase pertenecen los datos, por lo tanto, su objetivo es agrupar los datos por características o patrones similares en un número definido de clases.

Con el fin de disminuir el costo y el error humano, se han investigado métodos computacionales para mejorar el rendimiento y la precisión de la clasificación del suelo (Thonfeld, Feilhauer, Braun, & Menz, 2016). Previamente se han empleado entre otros métodos de aprendizaje computacional: los árboles de decisión (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012), las cadenas de Markov, la diferenciación de imágenes y cambio del vector de análisis, determinados a través de la matriz de transición de los mapas de cobertura, el análisis de componentes principales (ACP), las firmas espectrales y los índices de vegetación (Backoulou, Elliott, Giles, & Mirik, 2015).

Los árboles de decisión son una técnica de aprendizaje automático supervisado muy utilizada en muchos negocios. Como su nombre indica, esta técnica de machine learning toma una serie de decisiones en forma de árbol. Los nodos intermedios (las ramas) representan soluciones. Los nodos finales (las hojas) nos dan la predicción que vamos buscando. Los árboles de decisión pueden usarse para resolver problemas tanto de clasificación como de regresión.

3. Metodología

Se tomaron 1139 muestras de suelo en diferentes predios de municipios del departamento de Nariño (Córdoba, Cuaspud, Iles, Ipiales y Potosí), a las cuales se les realizaron análisis fisicoquímicos (contenidos de arenas, limos y arcillas, pH, conductividad eléctrica, contenido de materia orgánica, nitrógeno, fosforo intercambiable, azufre, calcio, magnesio, potasio, capacidad de intercambio catiónico efectiva, aluminio, hierro, manganeso, cobre, zinc, boro, saturación de aluminio, saturación de magnesio, saturación de potasio, saturación de calcio, relación calcio y magnesio, relación calcio y potasio, relación magnesio y potasio, relación calcio, magnesio y potasio), con el objetivo de establecer la calidad del suelo, y segmentarlos de acuerdo con esta información.

Inicialmente se realizó un análisis de correlación de Pearson, con la finalidad de observar correlaciones lineales entre variables, para luego aplicar un análisis de componentes principales (PCA), y generar un número de variables sintéticas que explique gran parte de la varianza de los datos originales. Posteriormente, con los resultados del PCA, se simularon modelos de aprendizaje no supervisado a través del método Kmeans, para realizar grupos de 1 a 10, y para determinar el número óptimo de clusters o grupos en los que segmentar la información, se utilizaron los métodos del “codo” (que tiene en cuenta cuando la inercia se reduce de forma constante, después de una determinada cantidad de clusters), el coeficiente de silueta de cada número de clusters (el número de clusters con mayor coeficiente es el mejor modelo) y se graficó la silueta para establecer visualmente el número de clusters más optimo.

Cuando se definió el número de clusters, fue necesario aplicar un modelo de aprendizaje automático con la información del PCA, para determinar las variables que más importancia relativa tenían en la diferenciación de cada cluster generado, para esto se aplico un modelo de Random Forest (RF). Y finalmente, se aplicó el modelo Decision Tree (DT), para determinar como se crearon los diferentes clusters, es decir, a partir de que variables y el valor de estas se generaron los diferentes agrupamientos.

De esta manera se logró diferenciar grupos entre las muestras de suelo obtenidas, y se determinó las variables más importantes para establecer el agrupamiento, y a partir de que valores en estas variables las muestras de suelo se diferenciaron lo suficiente como para pertenecer a un grupo distinto; información con la que se creo una segmentación de los suelos en los municipios muestreados.

4. Resultados

4.1. Correlación

Se realizó una correlación de Pearson para establecer si en la base de datos se presentaban relaciones lineales entre variables y así determinar si es posible efectuar un análisis para reducir el número de variables. De acuerdo con la Figura 1, se perciben correlaciones entre las variables Sat Al, Sat Mg, Sat K, Sat Ca, y entre las variables que relacionan contenidos de las bases del suelo (relación Ca/Mg, Ca/K, Mg/k y Ca/Mg/K); adicionalmente, se observan relaciones entre los contenidos de Ca, Mg, K, CICE con las variables que relacionan las bases de suelo (relación Ca/Mg, Ca/k, Mg/K y Ca/Mg/K). También se observan algunas relaciones entre los contenidos de arcillas, arenas y limos.

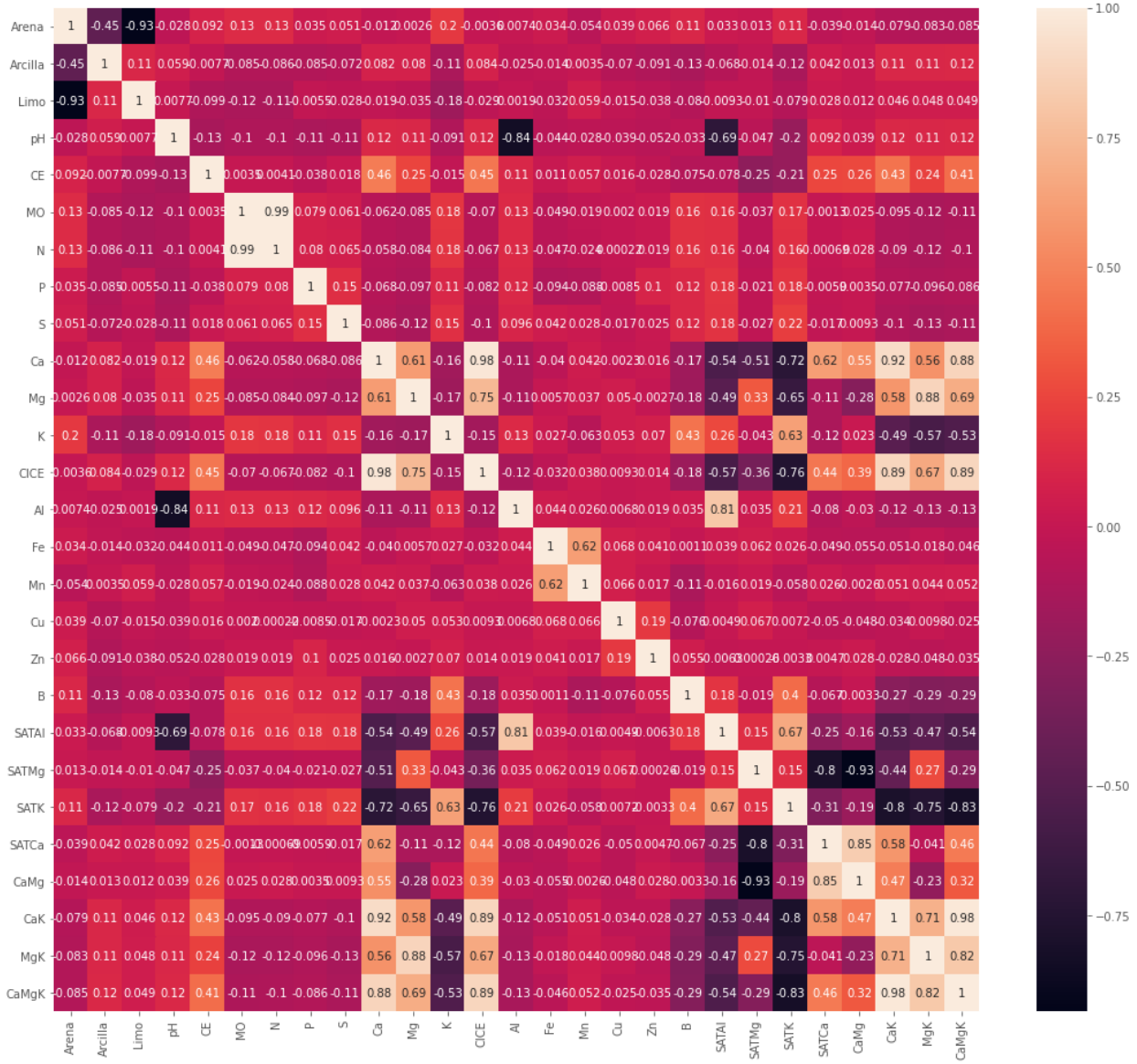

Figura 1. Correlación de Pearson entre variables.

De acuerdo con la información anterior es posible realizar un análisis de componentes principales (PCA), con la finalidad de reducir el número de variables y contener el mayor porcentaje posible de la variabilidad explicada en las nuevas variables sintéticas.

4.2. Análisis de Componentes Principales (PCA)

Se aplicó este análisis con el objetivo de crear nuevas variables sintéticas, denominadas componentes principales, no correlacionadas entre sí, que contengan la varianza de las variables originales. Como se observa en la Figura 2:

- En el primer componente se encuentra la variabilidad explicada por las variables de las relaciones entre bases del suelo (Ca/Mg, Ca/K, Mg/K y Ca/Mg/K), y las de saturación de las bases y el aluminio, los contenidos de Ca, Mg y CICE.
- En el segundo componente hay variabilidad explicada por la relación entre las bases del suelo y saturación de bases.

- En el tercer componente se encuentra la variación explicada por los contenidos de arcillas, limos y arenas, contenidos de MO y N, y las bases del suelo.

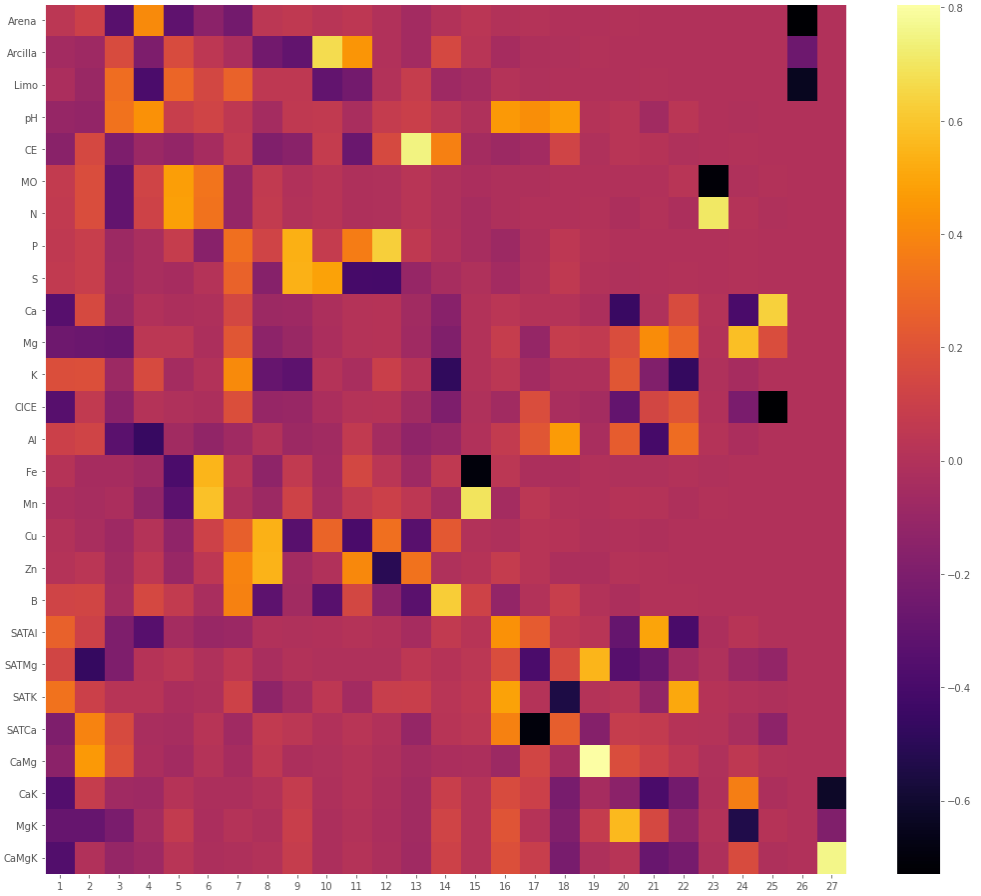

Figura 2. Mapa de calor de las variables originales y los componentes principales generados.

Por otro lado, en la Figura 3 se observa el porcentaje de varianza explicada por cada componente principal, y en la Figura 4, la varianza acumulada, de donde se extrae que hasta el séptimo componente principal se explica el 74% de la variabilidad de los datos originales.

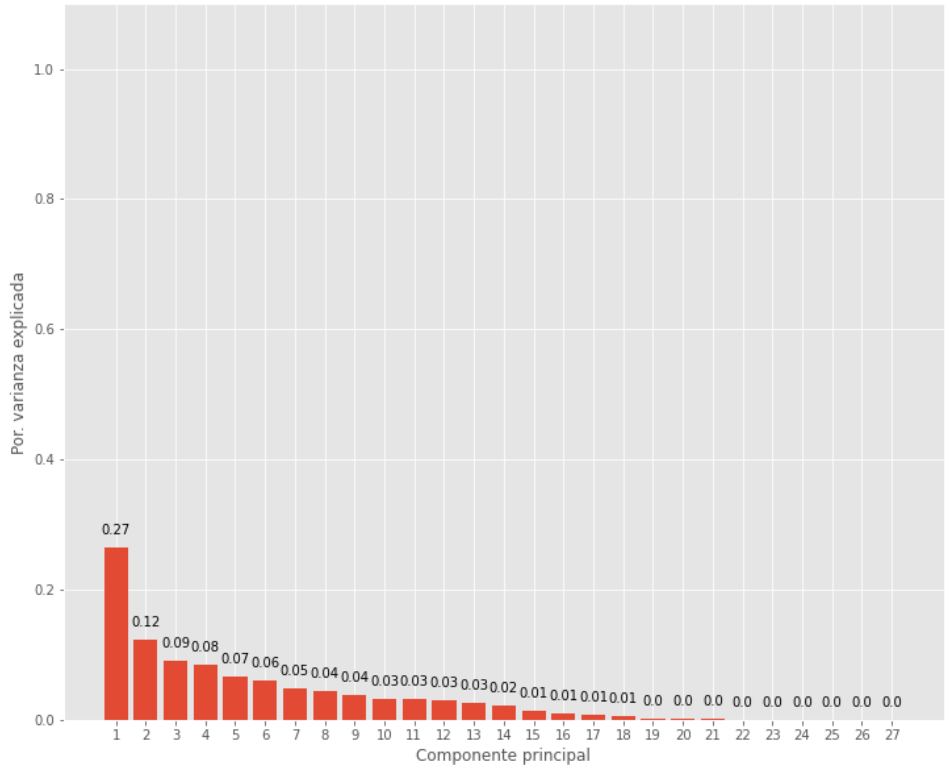

Figura 3. Porcentaje de varianza explicada por cada componente principal.

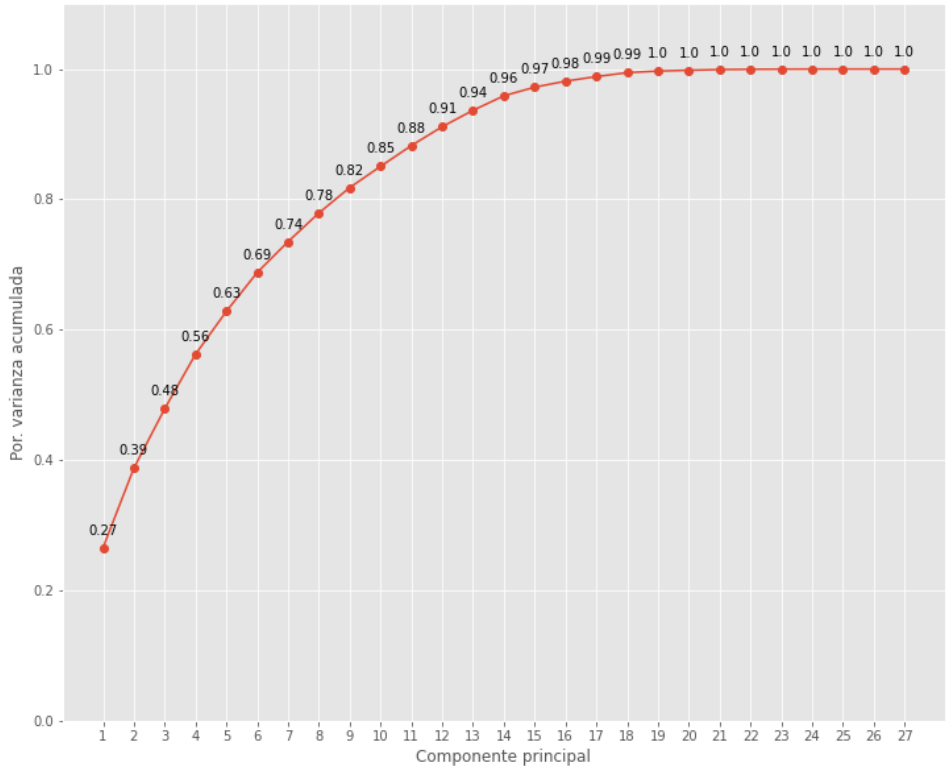

Figura 4. Porcentaje de varianza explicada acumulada.

Adicionalmente, se realizó una matriz de dispersión con las siete primeras variables sintéticas (componentes principales), en donde se visualiza la correlación entre estas (Figura 5). Se observa que la relación entre las componentes principales 1 y 2, y 1 y 3 los datos se dividen en tres grupos; la información preliminar proporcionada por la matriz de dispersión permite establecer que los datos pueden segmentarse en diferentes grupos.

Figura 5. Matriz de dispersión entre las primeras siete componentes principales.

4.3. Análisis Cluster

Para establecer el número de clusters o grupos estadísticamente más idóneo en los que se pueden dividir los datos, se realizaron previamente algunas pruebas con varios números de clusters (1 a 10), y se utilizaron diferentes métodos para seleccionar el más optimo:

- Establecer el número de grupos donde la reducción de la inercia se hace constante (método del codo).
- Establecer el número de grupos donde el coeficiente de silueta es más alto.
- Graficar la silueta y observar cuantos de los clusters de cada modelo la sobrepasan, lo que indica que es un buen modelo.

En la Figura 6 se observa el “método del codo”, que utilizando la inercia y como esta se reduce al aumentar el número de clusters, permite establecer el número “optimo” de grupos, cuando la reducción de la inercia se hace constante; en este caso se observa que tres clusters es el que mejor se ajustan a la descripción anterior, pues la inercia parece no reducirse drásticamente luego de este número de grupos.

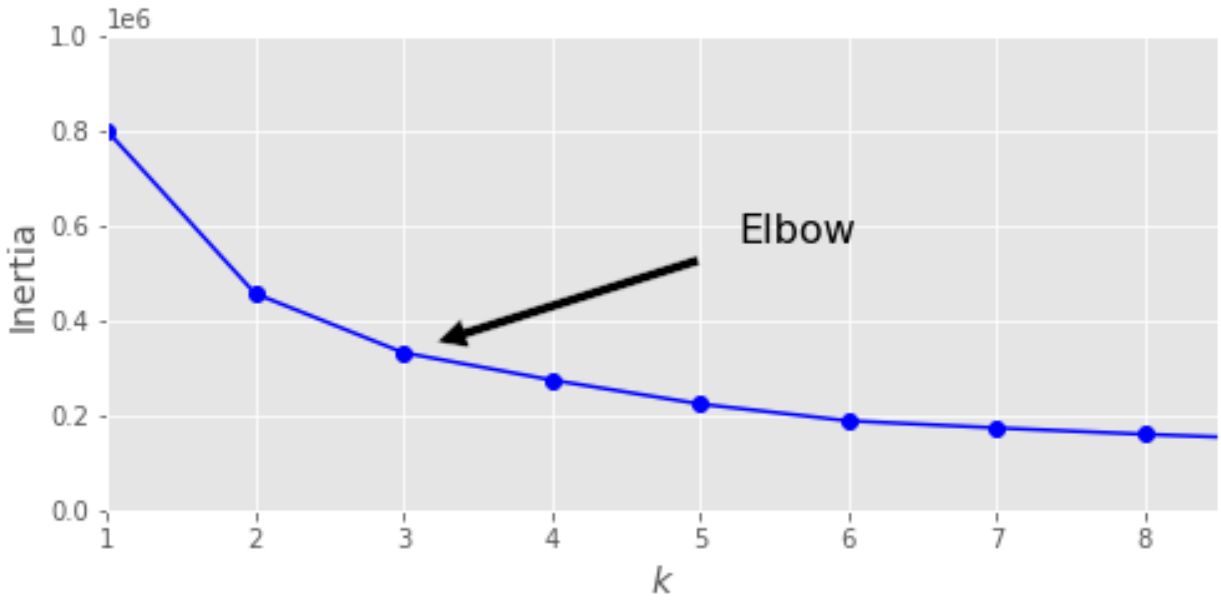

Figura 6. Gráfica del “método del codo” .

En este mismo sentido, en la Figura 7, se observa que el coeficiente de silueta más alto se presenta cuando el número de clusters es de tres, lo que sugiere nuevamente que este es el número más optimo de grupos a establecer con los datos suministrados por el PCA.

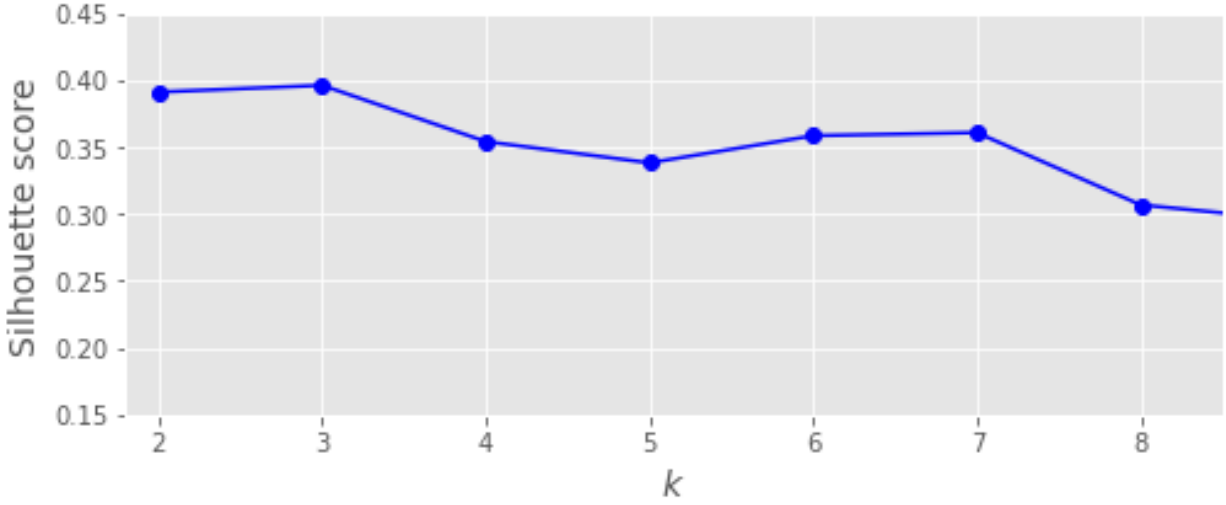

Figura 7. Gráfica del coeficiente de silueta.

En la Figura 8 se presenta la gráfica del coeficiente de silueta por número de clusters, dentro de cada modelo, el coeficiente de silueta se presenta como la línea roja, y si cada grupo dentro de cada modelo sobrepasa dicha línea roja estaría expresando que es un buen modelo de los datos suministrados por el PCA; para este caso, se observa que los modelos de 3, 4, 5 y 6 clusters, todos los grupos superan el coeficiente de silueta, por lo que serían elegibles.

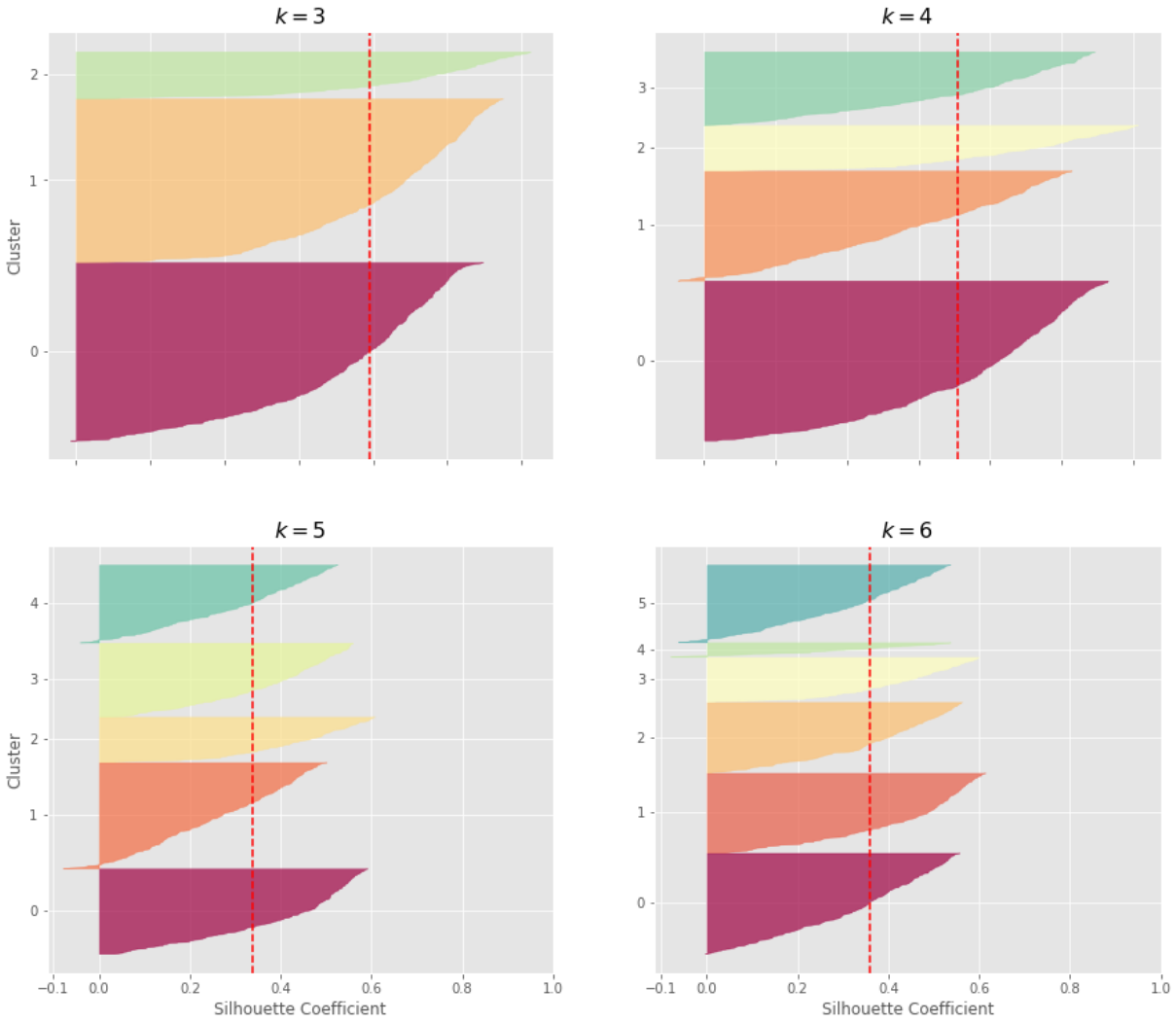

Figura 8. Grafica coeficiente de silueta para cada modelo evaluado.

Sin embargo, dado que dos de las tres pruebas realizadas sugirieron que establecer tres clusters es la forma más optima de agrupar los datos suministrados por el PCA, se estableció este como el número óptimo de grupos. Con esta información, se realiza nuevamente una matriz de dispersión (Figura 9), con las tres primeras componentes principales (PC) y los tres clusters generados.

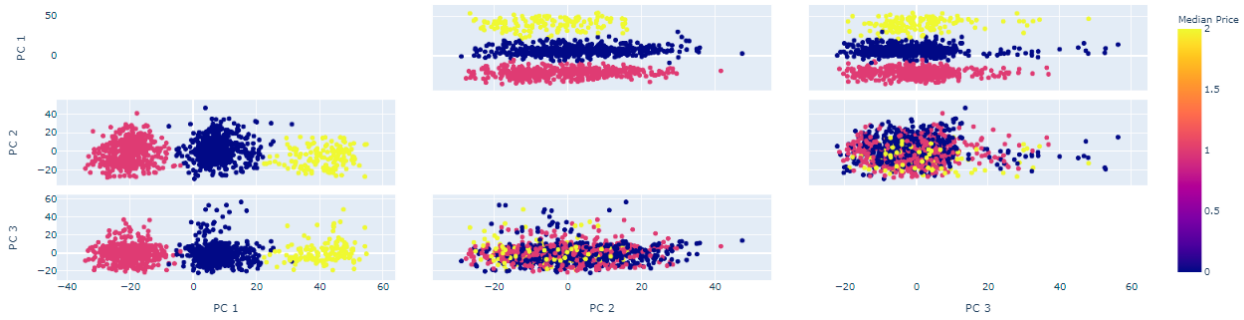

Figura 9. Matriz de dispersión para las tres primeras PC y los tres clusters generados.

Para determinar cómo estos tres clusters se diferencian, es decir que variables originales logran generar los grupos es necesario realizar otros análisis de aprendizaje automático, en este caso supervisado, utilizando como variable predictora los datos generados en el PCA.

4.4. Random Forest

Se utilizó un modelo de Random Forest (RF), con el objetivo de predecir la importancia relativa de las variables originales en los tres clusters generados. El modelo ejecutado, permitió establecer que las variables originales con mayor peso en los tres clusters, como se observan en la Figura 10, fueron Limo, Arena, Arcilla, contenido de K, Fe, relación Ca/K, entre otras.

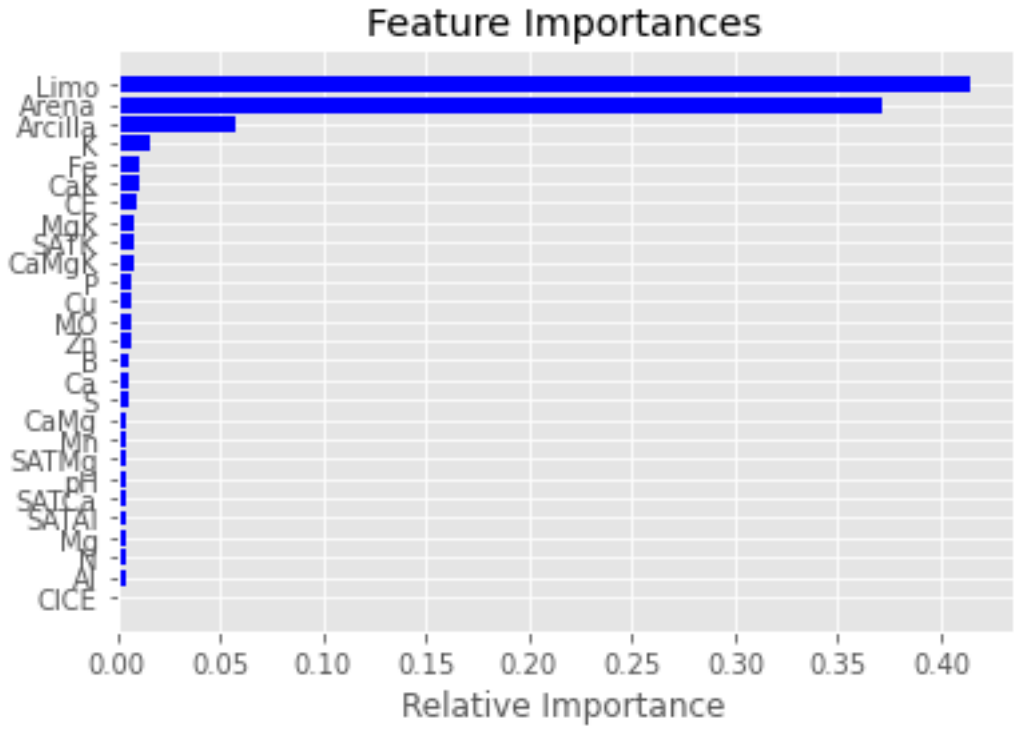

Figura 10. Importancia relativa de las variables en la generación de los clusters.

Sin embargo, esta información aún no permite establecer cuales variables tienen mayor incidencia en la generación de los clusters, para lo cual se requiere aplicar un modelo de árboles de decisión.

4.5. Decision Tree

Se aplica un modelo de Arboles de Decisión (DT), para determinar las variables originales que inciden mayormente en la segmentación de las 1139 muestras en los tres clusters generados. En la Figura 11, se puede observar que la segmentación de los clusters se generó de acuerdo con las variables: Arena, Limo, relación Ca/Mg/K y relación Ca/K. En el primer nodo (Arena<=51.5), se puede observar que un total de 522 muestras fueron agrupadas en el cluster 1, 479 en el cluster 2 y 138 en el cluster 3; por esta misma razón su coeficiente Gini es alto (menor pureza), es decir, en el nodo hay una mayor mezcla de los diferentes grupos.

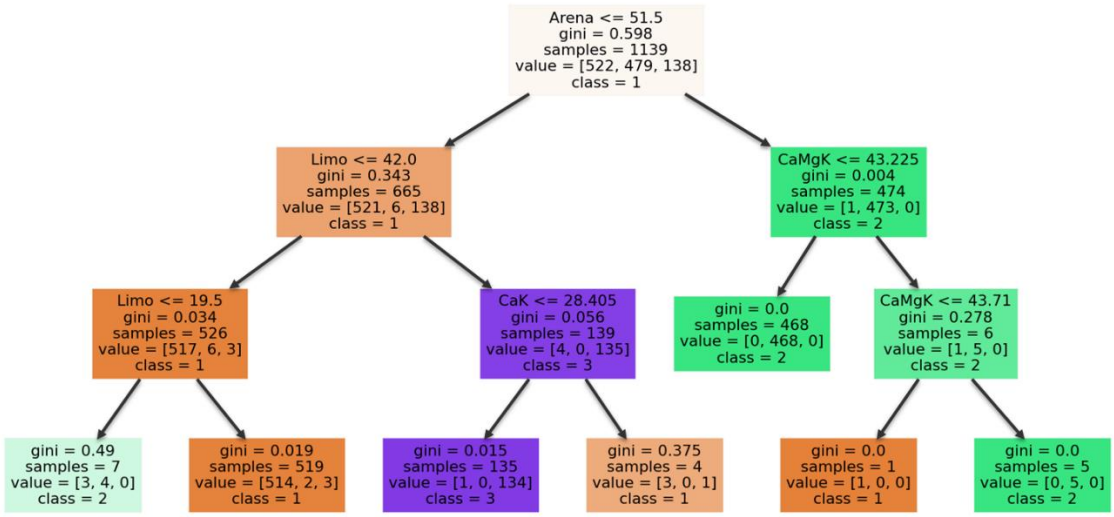

Figura 11. Visualización del modelo de árboles de decisión.

En este sentido, las muestras de suelos fueron segmentadas como sigue:

- Cluster 1: Muestras con contenidos de arena menores o iguales a 51.5, con contenidos de Limo <= 42.0 y superiores a 19.5 (dependiendo de los contenidos de arcillas podrían ser suelos franco arcillo arenosos, franco arcillosos, francos, arcillo arenosos o arcillosos).
- Cluster 2: Muestras con contenidos de arena mayores a 51.5, relación (Ca+Mg)/K <= 43.23 (lo cual indica contenidos de K adecuados).
- Cluster 3: Muestras con contenidos de Arena<=51.5, con contenidos de Limo>42.0 (dependiendo de su contenido de arcillas, podrían presentarse suelos francos, franco arcillosos, franco arcillo limosos, arcillo limosos, franco limosos y limosos), y relación Ca/K<= 28.405 (indicando contenidos de K adecuados).

5. Conclusiones

- Es posible segmentar los suelos de acuerdo con sus características fisicoquímicas por medio del uso de modelos de aprendizaje automático no supervisado.
- Específicamente el método Kmeans permitió evaluar diferentes modelos de segmentación, desde 1 a 10 clusters.
- Los diferentes métodos utilizados para definir el número óptimo de clusters permitieron inferir que el número optimo sería de tres clusters.
- Para comprender como se diferencian estos grupos fue necesario aplicar modelos de aprendizaje supervisado tomando como base el PCA y los clusters generados en la etapa del modelo no supervisado.
- Los datos obtenidos de los distintos predios muestreados en el departamento de Nariño permiten afirmar que la mayor diferencia fisicoquímica se encuentra en la textura del suelo y en la relación de las bases intercambiables.

6. Bibliografía

Backoulou, G., Elliott, N., Giles, K., & Mirik, M. (2015). Processed multispectral imagery differentiates wheat crop stress caused by greenbug from other causes. *Computers and Electronics in Agriculture (115)*, 34-39.

Baviera, T. (2017). Técnicas para el Análisis de Sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. *Revista Dígitos, 1(3)*, 33-50.

Castañeda, D., Jaramillo, D., & Cotes, D. (2014). Selección de propiedades del suelo espacialmente relacionadas con producción en el cultivo de banano. *Ciencia del Suelo*, 73-83.

Doran, J., & Parkin, B. (1994). *Defining Soil Quality for a Sustainable Environment.* Madison, Wisconsin (USA): Soil Science Society of America, Publicación Especial. Número 35.

Espinoza-Freire, E., & Tinoco-Cuenca, N. (2015). La problemática ambiental resultante de la fumigación aérea con plaguicidas a bananeras de la provincia de El Oro Ecuador. *Ciencia en su PC (4)*, 75-87.

Fajardo, J., Aguilar, L., Flores, M., Parra, D., & Acurio, C. (2017). Aspectos socio-económicos influyentes en el sector agrícola y su relación con la creación de empresas populares y solidarias de Quevedo. *Ciencias Sociales y Económicas 1(1)*, 107-116.

FAO. (2015a). *Las amenazas a nuestros suelos.* Roma (Italia): Organización de las Naciones Unidas para la Alimentación y la Agricultura.

FAO. (2015b). *Carta Mundial de los Suelos.* Roma (Italia): Organización de las Naciones Unidas para la Alimentación y la Agricultura.

FAO. (2015c). *El origen de los alimentos.* Roma (Italia): Organización de las Naciones Unidas para la Alimentación y la Agricultura.

Fernández, A., Fernández, R., Rivera, C., & Calero, S. (2016). Desafíos de la gestión de producción agropecuaria en Cuba. *Agroalimentaria, 22(42)*, 119-132.

Goya, R., Barquero, O., & Figuera, C. (2017). Enseñanza del aprendizaje automático utilizando las competiciones de Kaggle. *Nuevos enfoques en la Innovación Docente Universitaria*, 66-70.

IGAC. (2012). *Conflictos de uso del territorio Colombiano. Escala 1:100.000.* Bogotá (Colombia).

Martin, P. (1998). Soil carbon and climate perturbations: using the analytical biogeochemical cycling (ABC) scheme. *Environmental Science and Policy (1)*, 87-97.

MAVDT, & IDEAM. (2004). *Plan de Acción Nacional de Lucha Contra la Desertificación en Colombia.* Bogotá (Colombia).

Minambiente, & IDEAM. (2015). *Línea base de degradación de suelos por erosión en Colombia (2010 – 2012). Escala 1:100.000.* Bogotá (Colombia).

Ministerio de Ambiente y Desarrollo Sostenible. (2012). *Política Nacional para la Gestión Integral de la Biodiversidad y sus Servicios Ecosistémicos – PNGIBSE.* Bogotá (Colombia): Ministerio de Ambiente y Desarrollo Sostenible.

Ministerio de Ambiente y Desarrollo Sostenible. (2016). *Política para la Gestión y sostenible del suelo.* Bogotá (Colombia): Ministerio de Ambiente y Desarrollo Sostenible.

Moreno-Carriles, R. (2018). Big data, ¿pero qué es? *Angiología 70(5)*, 191-194.

Rodriguez, D., & Fusco, M. (2017). Gestión de riesgos agropecuarios en el sector del cacao en Ecuador. *Revista de Investigación en Modelos Financieros 6(1)*, 57-74.

Rodriguez-Galiano, V., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing (67)*, 93-104.

Suárez, A., Jiménez, A., Castro-Franco, M., & Cruz-Roa, A. (2017). Clasificación y mapeo automático de coberturas del suelo en imágenes satelitales utilizando Redes Neuronales Convolucionales. *Colombia Suplemento (21)*, 64-75.

Thonfeld, F., Feilhauer, H., Braun, M., & Menz, G. (2016). Robust Change Vector Analysis (RCVA) for multi-sensor very high resolution optical satellite data. *International Journal of Applied Earth Observation and Geoinformation (50)*, 131-140.

Van Miegrot, H., & Johnsson, D. (2009). Feedbacks and synergism among biochemistry, basic ecology, and forest soil science. *Forest Ecology an Management 258*, 2214-23.

Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., & Madabhushi, A. (2014). Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging (Bellingham, Wash.), 1(3)*, 34003.

Zoppolo, R., & Fasiolo, C. (2016). Análisis foliar en frutales: Herramienta de diagnóstico de alto retorno. *Revista INIA*, 27-28.