Conciencia Digital
www.concienciadigital.org

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

Análisis de patrones de características de especies andinas de las reservas Chimborazo y Sangay utilizando el método k-means clustering

Analysis of patterns, Andean species, reserves, Chimborazo, Sangay, k-means method, clustering

Paúl Xavier Paguay Soxo.[1], Janneth Ximena Idrobo Cárdenas. [2], Pamela Alexandra Buñay Guisñan. [3] & Angel Patricio Flores Orozco. [4]

Recibido: 14-12-2019 / Revisado: 02-01-2020 /Aceptado: 18-01-2020/ Publicado: 07-02-2020

Abstract. **DOI:** https://doi.org/10.33262/concienciadigital.v3i1.1.1143

Over the years, researchers have recognized the importance of studying the moors and their conservation, both for their impact on the provision of water for cities, as well as their tourism potential and biodiversity. The objective of the present investigation is to conduct an analysis of patterns present in the characteristics of the different species of the Andean region of the Chimborazo and Sangay National Park reserves. For the analysis, 67 samples of different species were collected in both reserves, of which measurements of leaves, plants and flowers were made, subsequently, non-supervised machine learning algorithms called k-means clustering were applied using Python as the programming language . At the end of the species grouping process, it resulted in obtaining three categories according to the relationship between the characteristics of each species, two components being the most important for categorization, these were the height of the plant as well as The height of the leaf .

Keywords: Analysis of patterns, Andean species, Chimborazo and Sangay reserves, k-means method, clustering

Resumen.

A lo largo de los años, los investigadores han reconocido la importancia del estudio de los páramos y su conservación, tanto por su impacto en la provisión de agua para las ciudades, así como su potencialidad turístico y biodiversidad. El objetivo de la presente investigación es realizar un análisis de patrones presente en las características de las diferentes especies de la

[1] Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador, ppaguay@espoch.edu.ec
[2] Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador, janneth.idrobo@espoch.edu.ec
[3] Universidad Nacional de Chimborazo, Ecuador, pbunay@unach.edu.ec
[4] Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador, aflores@espoch.edu.ec

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

región andina de las reservas Chimborazo y Parque Nacional Sangay. Para el análisis se recolectó 67 muestras de diferentes especies en ambas reservas, de las cuales se realizaron mediciones de hojas, planta y flores, posteriormente se aplicó algoritmos de aprendizaje no supervisado de machine learning denominado k-means clustering utilizando como lenguaje de programación a Python. Al finalizar el proceso de agrupamiento de las especies, arrojó como resultado la obtención de tres categorías de acuerdo a la relación existente entre las características de cada especie, siendo dos componentes los más importantes para la categorización, estas fueron el alto de la planta así como el alto de la hoja.

Palabras Clave: Análisis de patrones, especies andinas, reservas Chimborazo y Sangay, método k-means, clustering

Introducción.

Según Vásconez (2001), los páramos "Fueron considerados tierras improductivas, hostiles, con poca oferta de diversión, con gente pobre con muchos problemas y no tan atractivos", inclusive en el campo de la botánica no llamaba la atención de los investigadores. Esta situación ha cambiado durante el transcurso de los años, luego de que entre las décadas de 1930 y 1950 Misael Acosta Solís lideró un proceso de institucionalización del conservacionismo en el Ecuador, que posteriormente los estudios fueron continuados por varios autores e instituciones gubernamentales, como por ejemplo por su impacto en la provisión de agua para las ciudades, así como su potencialidad turístico y biodiversidad.

La provincia de Chimborazo, con una extensión de 648.124 hectáreas, posee un poco más de 246.000 hectáreas de ecosistema páramo (es decir el 38% de la superficie de la provincia), y otras 49571.16 hectáreas de bosque andino y altoandinos (es decir 8%) (Bustamante et al., 2011). En la provincia de Chimborazo existen dos áreas protegidas del PANE (Patrimonio Nacional del Estado), la Reserva de Producción de Fauna Chimborazo y el Parque Nacional Sangay. Entre ellas protegen 91.667 hectáreas, que representa el 14% del total de la provincia y el 37% del total de los páramos de la provincia. En este contexto al igual que otras regiones de páramos en el Ecuador, se han visto afectadas por las actividades humanas, principalmente agrícolas y ganaderas, lo que provoca que el sustento de las comunidades cercanas se vea afectadas por la erosión o desgaste de los suelos. Por este motivo es importante el estudio de la biodiversidad en esta región para empoderar a la sociedad sobre la importancia del páramo y buscar alternativas sustentables y sostenibles de desarrollo para la población que vive a los alrededores de estas regiones.

El presente trabajo se enfoca en realizar un análisis de las especies nativas de la región andina de las reservas del Chimborazo y Parque Nacional Sangay con el fin de encontrar patrones en las características de las especies encontradas en el sitio, para este cometido la investigación en primera instancia abarca la descripción de la metodología utilizada donde este apartado se divide en dos partes, el primero que tiene que ver con el análisis y tabulación de la información de las muestras recolectadas en cuatro visitas a campo donde se ha analizado las características como dimensiones de las hojas, planta, flores, clima, altitud, color. La segunda parte de se

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

enfoca en la búsqueda de patrones aplicando algoritmos de Clasificación No Supervisada de agrupamiento por K medias (K-Means Clustering) del área de Aprendizaje Automático (Machine Learning).

Posteriormente se expone los resultados del análisis de patrones utilizando el método escogido para al final presentar las conclusiones del trabajo.

DESARROLLO

Marco Teórico Referencial

La caracterización morfológica, es una de las actividades principales en los estudios de especies florícolas de una región específica, a continuación, se cita varios trabajos referentes al tema.

Artículo Científico: "Caracterización morfológica de los algarrobos (Prosopis sp.) en las regiones fitogeográficas Chaqueña y Espinal norte de Argentina", de los autores Verga et al. (2009) en la que se encuentra relaciones de las características de las familias de especies analizadas y propone la identificación de dos formas morfológicas de algarrobos (Proposis alba).

Otra investigación en el mismo campo es la realizada por Ana María Villa (2006) en su trabajo "Caracterización diamétrica de las especies maderables en bosques primarios del Cerro Murrucucú", donde concluye que las regiones de bosques están formadas por especies con relativa diferencia de características, así también establece las tasas de crecimiento con proyección a una producción sostenible.

Cantillo H et al. (2004) en su investigación "Diversidad y caracterización florística estructural de la vegetación arbórea en la reserva forestal Cárpatos (Guasca - Cundinamarca)" encuentra dos grupos que presentan diferente desarrollo fisionómico dependiendo de altura, cobertura, estructura diamétrica y área basal.

En el trabajo "Caracterización morfológica y genética de las ectomicorrizas formadas entre Pinus montezumae y los hongos presentes en los bancos de esporas en la Faja Volcánica Transmexicana" de Garibay-Orijel et al. (2013) encuentra que existe una similitud genética de la región de los ITS y concluye que las especies analizadas podrían usarse para reforestar con plantas y hongos endémicos, lo que aumentaría la sobrevivencia, pues ambos simbiontes estarían adaptados a las condiciones ambientales locales.

Partiendo de estos precedentes se hace necesario de igual manera realizar un análisis morfológico de las especies presentes en la flora nativa andina de la provincia de Chimborazo, específicamente en las reservas Chimborazo y Parque Nacional Sangay que permita establecer una línea base para futuros trabajos con el fin de comparar diferentes variables como localización, altura, tiempo, etc.

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

Aprendizaje Automático (Machine Learning)

Para Arthur Samuel lo describió como: "El campo de estudio que otorga a las computadoras la capacidad de aprender sin ser programado explícitamente", esta es una definición más antigua, informal.

Tom Mitchell proporciona una definición más moderna:

> "***Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y la medida de rendimiento P, si su desempeño en tareas en T, medido por P, mejora con la experiencia E.***" (Carbonell & Mitchell, n.d.).

Clasificación por Agrupamiento de K-Medias (K-Means Clustering)

Es uno de los más simples y conocidos algoritmos de agrupamiento, sigue una forma fácil y simple para dividir una base de datos dada en k grupos.

La idea principal es definir k centroides (uno para cada grupo) y luego tomar cada punto de la base de datos y situarlo en la clase de su centroide más cercano. El próximo paso es recalcular el centroide de cada grupo y volver a distribuir todos los objetos según el centroide más cercano. El proceso se repite hasta que ya no hay cambio en los grupos de un paso al siguiente (Pascual et al., n.d.).

A continuación, se muestra un ejemplo de búsqueda de grupos mediante el establecimiento de 3 centroides y 6 iteraciones en un gráfico de dos dimensiones (componentes).

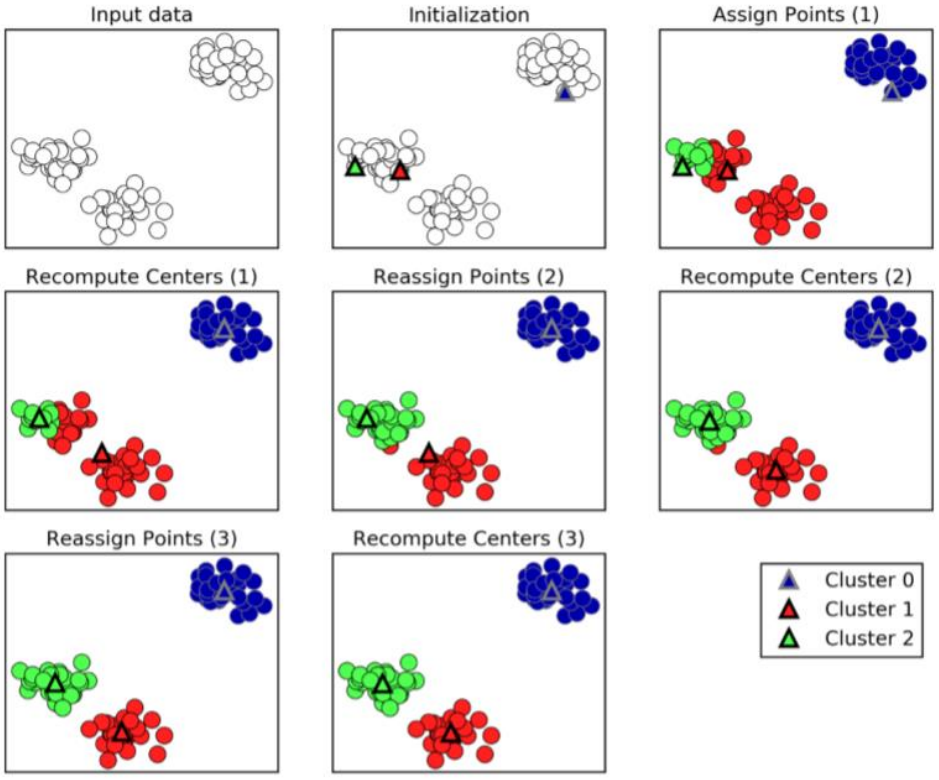


Figura 1. Iteraciones de agrupamiento con K-Means Clustering
Fuente: (Muller & Guido, 2017)

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

Análisis de Componentes Principales

Según los autores Abdi & Williams (2010), el Análisis de Componentes Principales (PCA) es una técnica multivariada que analiza una tabla de datos en la que las observaciones son descritas por varias variables dependientes cuantitativas interrelacionadas. Su objetivo es extraer la información importante de la tabla, representarla como un conjunto de nuevas variables ortogonales llamadas componentes principales y mostrar el patrón de similitud de las observaciones y de las variables como puntos en los mapas.

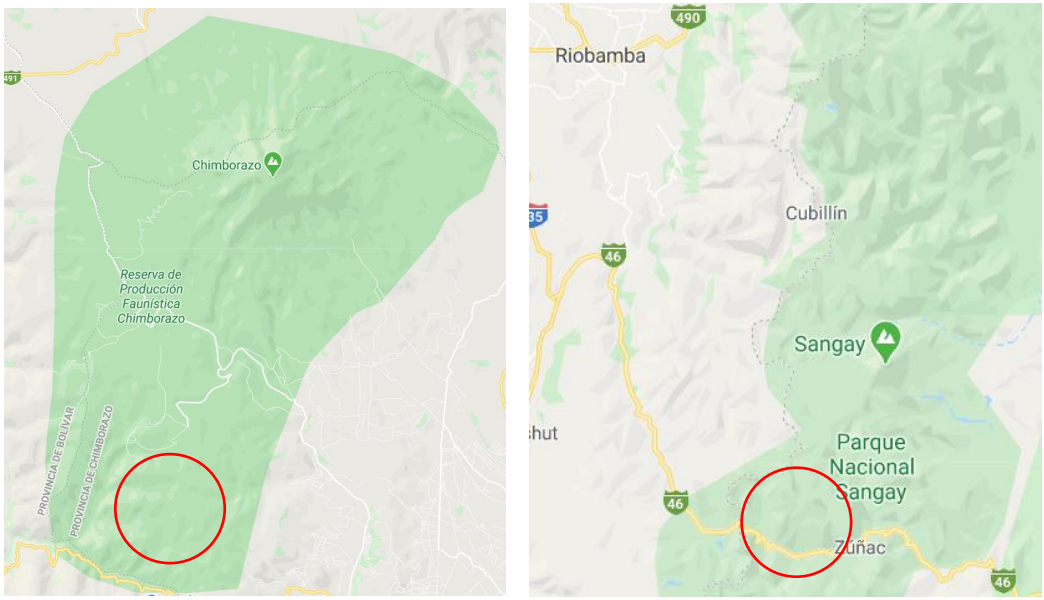
Para el proyecto esta técnica será utilizada para de las características analizadas de las especies se extraiga las características más importantes que posteriormente servirán para la búsqueda de patrones. La plataforma Python entre su vasta cantidad de librerías se tiene la de scikit-learn que contiene varias clases para la implementación de esta técnica.

Marco Metodológico

La investigación tiene un enfoque experimental donde se enfoca en establecer una muestra y aplicar algoritmos de agrupamiento de k-medias para obtener patrones de las características presentes en las muestras de las especies analizadas.

Área de estudio

El área donde se recolectó las muestras de las especies para el posterior análisis se encuentra en la Reserva de Producción Faunística Chimborazo por el sector del bosque de Polylepis a una elevación del terreno de 4200 msnm, la otra área se encuentra en el Parque Nacional Sangay por el sector de las Lagunas de Atillo con una elevación del terreno de aproximadamente de 3400 msnm en un área de 1000 metros cuadrados. En la figura 2 se observan las dos áreas de estudio para la investigación.



a. Reserva Chimborazo b. Parque Nacional Sangay

Figura 2: Áreas de estudio
Fuente: (*Google Map*, 2019)

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

Población y muestra

Para el experimento se utilizó la información de un total de 89 muestras recogidas de diferentes especies, de las cuales **67** cumplían con los requerimientos para el análisis.

Los puntos escogidos para la recolección de información fueron las dos áreas protegidas de la provincia de Chimborazo.

Herramientas

Varias herramientas fueron utilizadas en el experimento, los cuales se detallan a continuación, así como en la figura 3 se visualizan los logos de cada una.

- Lenguaje Python
- Plataforma Anaconda
- Editor Spyder
- Gestor de Base de Datos PostgreSQL

Figura 3: Herramientas para análisis de datos
Fuente: (Scikit-learn, n.d.)

Curva Elbow (Codo)

El método de "Elbow" ayuda a los científicos de datos a seleccionar el número óptimo de grupos ajustando el modelo con un rango de valores para K. Si el gráfico de líneas se asemeja a un brazo, entonces el "codo" (el punto de inflexión en la curva) es una buena indicación de que el modelo subyacente se ajusta mejor en ese punto (The scikit-yb developers., 2019) . En la siguiente figura se observa un ejemplo de Curva Elbow donde el codo demuestra que el componente 3 o incluso 4 puede ser el más ajustado para el modelo.

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

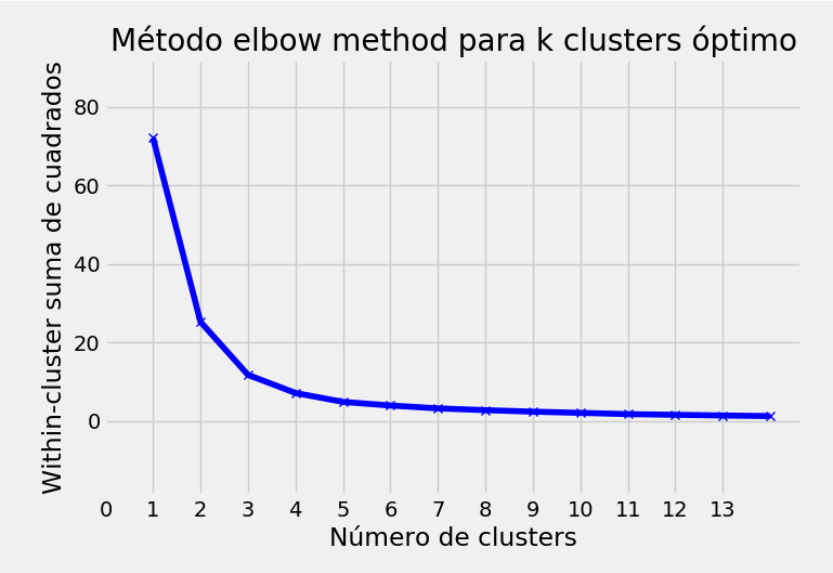


Figura 4: Curva Elbow (Codo)
Fuente: Autor

Sistema Informático

El sistema informático desarrollado se publicó en uno de los servidores de la ESPOCH con acceso externo y se encuentra en la siguiente URL **http://disenoandino.espoch.edu.ec.** En la figura 4 se observa la página principal del sistema.

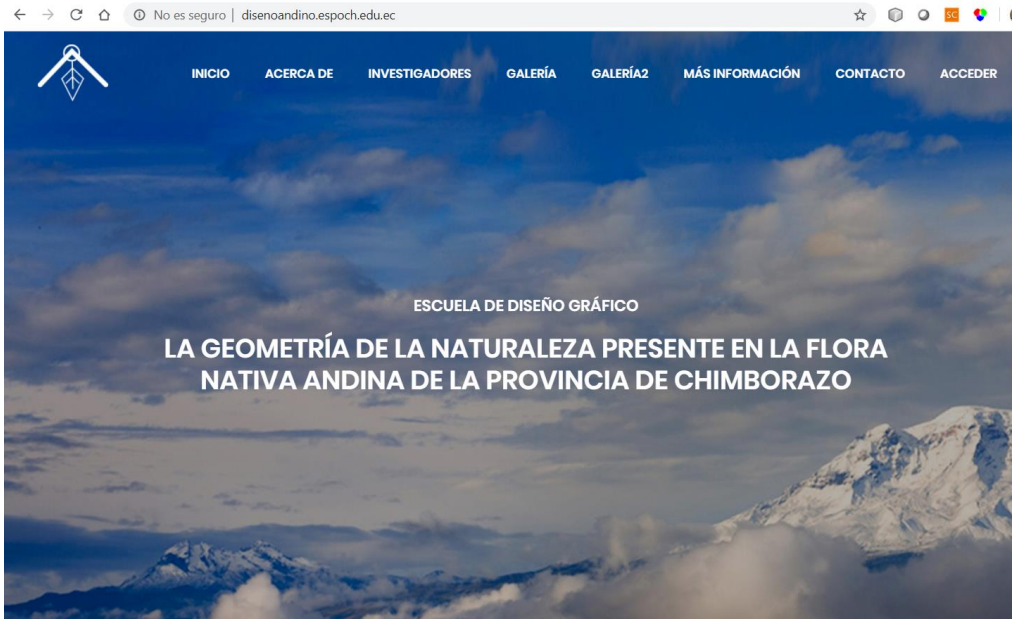


Figura 5: Página principal del Sistema Informático
Fuente: Autor

En las siguientes figuras 6 y 7 se observan las pantallas del sistema correspondientes al módulo de muestras donde se ingresa y visualiza la información almacenada.

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

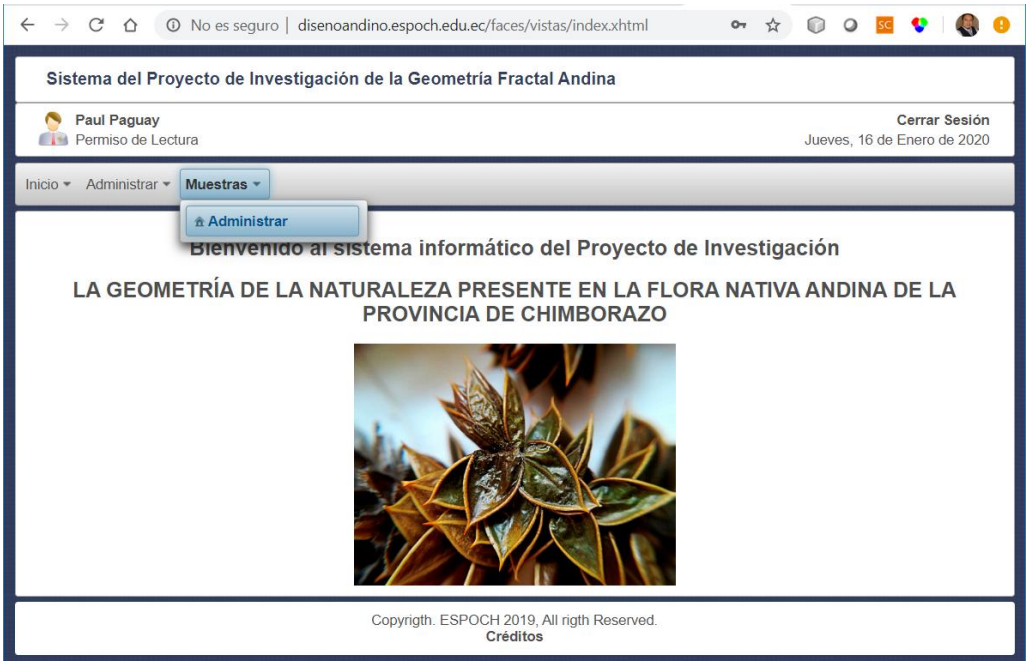


Figura 6: Página principal del módulo de administración
Fuente: Autor

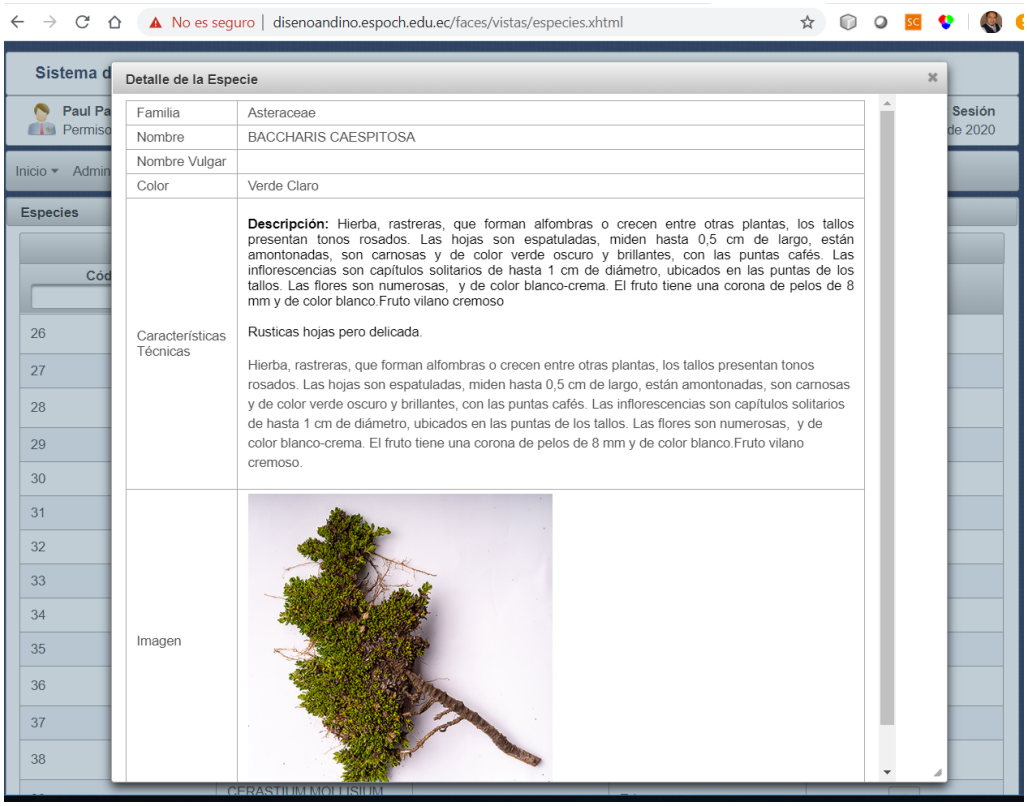


Figura 7: Módulo de Muestras de Especies
Fuente: Autor

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

Procedimiento

Para la realización de la investigación, se procedió con los siguientes pasos:

1. Recolección de especies nativas de las dos áreas protegidas
2. Análisis morfológico de las especies, almacenando sus características como alto, ancho, espesor de la planta, hoja, flor.
3. Paralelamente con la fase anterior se desarrolló el sistema informático para el almacenamiento de la información y posterior análisis
4. Desarrollo y ejecución del algoritmo no supervisado K-means Clustering para el análisis de patrones de los datos.
    - Programación de la sentencia SQL
    - Carga de información desde la base de datos
    - Selección de campos
    - Limpieza de datos
    - Transformación de datos
    - Encontrar k mediante la gráfica Elbow
    - Aplicación del algoritmo k-means
    - Representación gráfica de los resultados

Análisis de Resultados

En este apartado se muestran los resultados obtenidos del experimento.

En la figura 8 se visualiza la gráfica de Elbow de los datos analizados, donde se evidencia que un leve cambio de dirección de la función se observa entre los valores 2 y 3, por lo que el valor 3 será el utilizado para el posterior procesamiento.

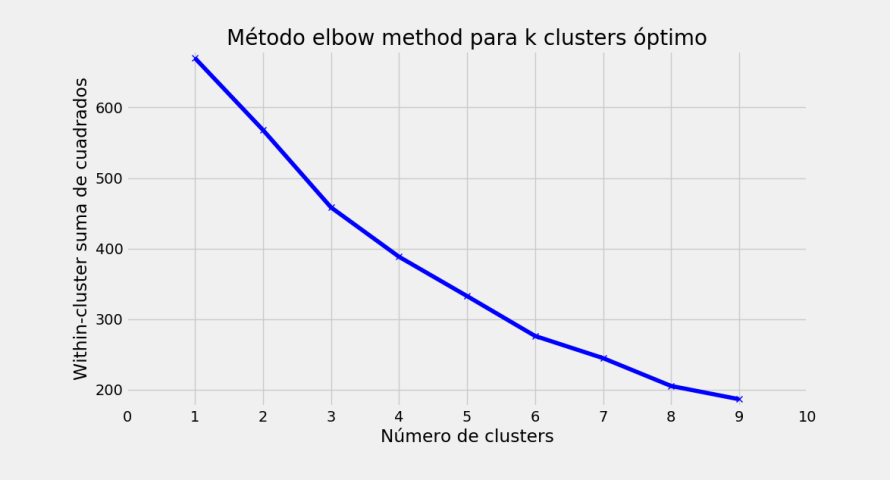


Figura 8: Gráfica de Elbow
Fuente: Autor

Utilizando la técnica del Análisis de Componentes Principales, se obtiene dos componentes importantes y estos son utilizados para la gráfica en dos dimensiones donde se visualizan los 3

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

grupos de datos, cabe recalcar que para el tercer grupo existe un solo elemento por lo que en lo posterior fue descartado para el análisis.

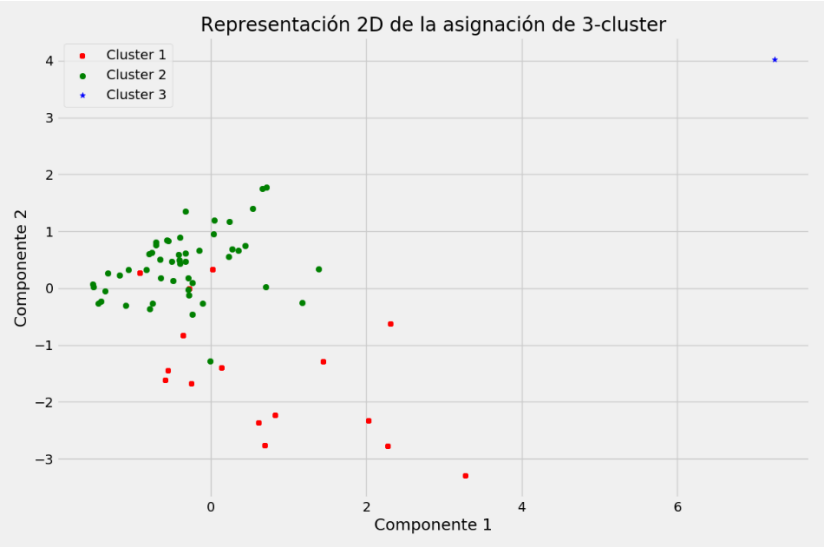


Figura 9: Representación 2D
Fuente: Autor

Nuevamente aplicando la técnica del Análisis de Componentes Principales, se buscó 3 componentes con mayor influencia en los datos, una vez encontrados se visualiza en la figura 10 la disposición de los datos en un plano tridimensional.

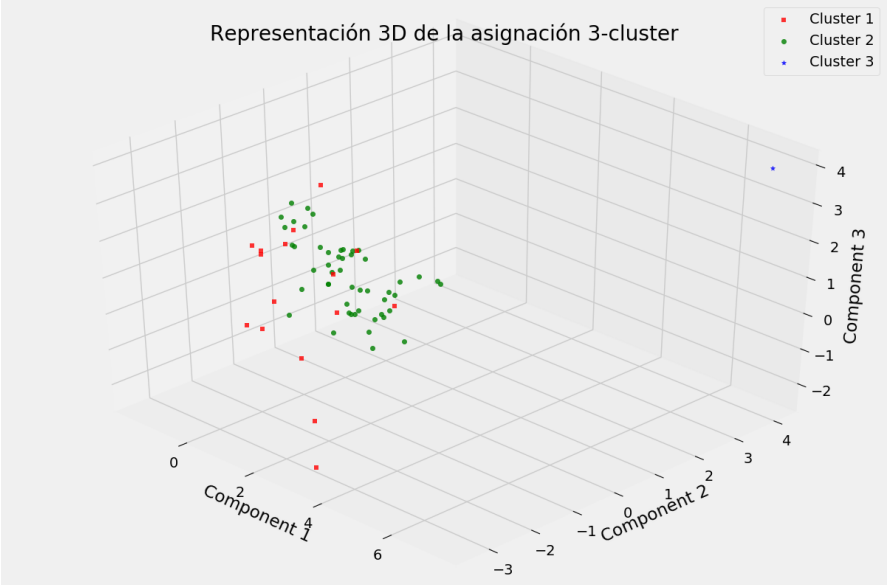


Figura 10: Representación 3D
Fuente: Autor

El total de especies por cada grupo encontrado se observa en la figura 11, como se puede evidenciar el grupo que tiene más especies es el Grupo 1 con 50 especies, seguido del grupo 2 con 16 especies y 1 de grupo 3.

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

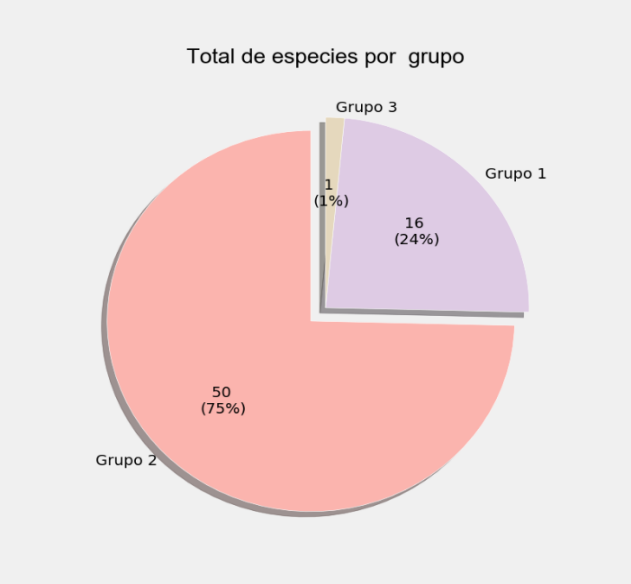


Figura 11: Totales por grupo
Fuente: Autor

Posteriormente se realizó el análisis con cada grupo, donde se encontró que para el grupo 1 las familias que más se encontró fueron las Poaceae con un 31.2% seguido de las Asteraceae con 18.8% al igual que Caprifoliaceae. Esto se puede observar en la figura 12.

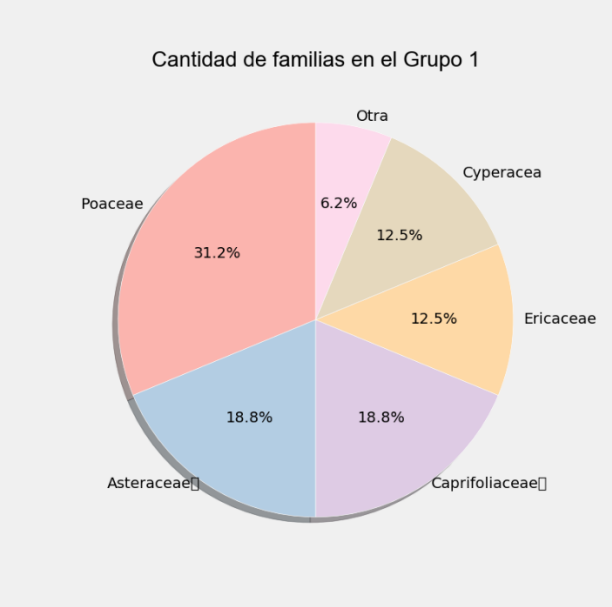


Figura 12: Familias por grupo 1
Fuente: Autor

Continuando con el análisis de familias por cada grupo encontrado se observa que para el grupo 2 las Asteraceae son las que más aparecen en este grupo con un 40%, seguido por Caprifoliaceae con un 14%, como se visualiza en la figura 13.

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

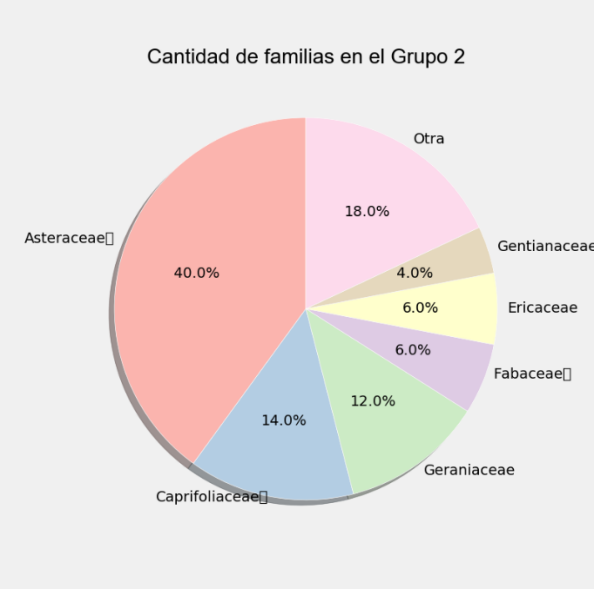


Figura 13: Familias por grupo 2
Fuente: Autor

Para el grupo tres la única especie encontrada en este grupo pertenece a la familia de las Poaceae.

En la figura 14 se observa la comparación de medias de las dimensiones de la altura, ancho y espesor de la planta y hoja. Se observa que el grupo 1 se caracteriza por tener mayor altura de la planta como de la hoja mientras que el grupo 2 se caracteriza por tener mayor ancho de la planta así como mayor espesor.

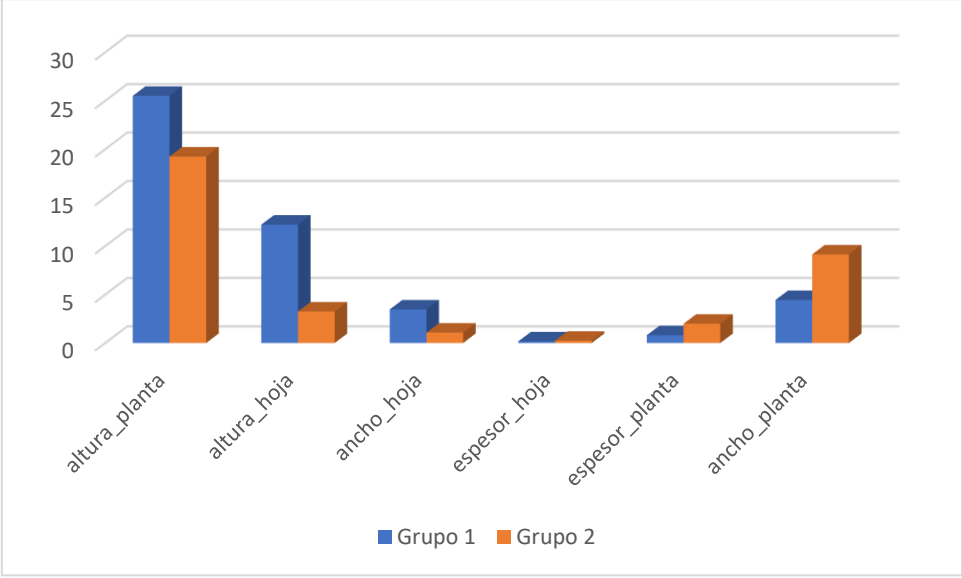


Figura 14: Comparación de Medias
Fuente: Autor

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

CONCLUSIONES

- El análisis y estudio de las características de las especies de los páramos, permite conocer la biodiversidad de estas áreas protegidas y constatar su importancia en la vida de las poblaciones aledañas.
- El resultado del presente experimento mostró la evidencia de patrones en las características de las diferentes especies, categorizadas en dos grupos que de acuerdo con su estructura muestran similitudes primordialmente en dos componentes, la altura de la planta y altura de la hoja.
- Como trabajo futuro se propone incorporar nuevas características como altitud de la recolección, color, forma, que permita mostrar nuevos patrones todavía por conocer.

Referencias Bibliográficas,

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Bustamante, M., Albán, M., & Argüello, M. (2011). *Los Paramos De Chimborazo Un Estudio Socioambiental Para La Toma De Decisiones*. 151. www.flacsoandes.edu.ec

Cantillo H, E. E., Rodríguez R, K. J., & Avella M, E. A. (2004). Diversidad y Caracterización Florística Estructural de la Vegetación Arbórea en la Reserva Forestal Carpatos (Guasca Cundinamarca). *Colombia Forestal*, *8*(17), 5. https://doi.org/10.14483/udistrital.jour.colomb.for.2004.1.a01

Carbonell, J. G., & Mitchell, T. M. (n.d.). AN OVERVIEW OF MACHINE LEARNING. In *MACHINE LEARNING: An Artificial Intelligence Approach*. Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051054-5.50005-4

Villa, A. (2006). Caracterizacion diametrica de las especies maderables en bosques primarios del Cerro Murrucucú. *Gestión y Ambiente*, *9*(2), 73–90. https://doi.org/10.15446/ga.v9n2.52064

Garibay-Orijel, R., Morales-Marañon, E., Domínguez-Gutiérrez, M., & Flores-García, A. (2013). Caracterización morfológica y genética de las ectomicorrizas formadas entre Pinus montezumae y los hongos presentes en los bancos de esporas en la Faja Volcánica Transmexicana. *Revista Mexicana de Biodiversidad*, *84*(1), 153–169. https://doi.org/10.7550/rmb.29839

Google Map. (2019). https://www.google.com/maps/

Muller, A. C., & Guido, S. (2017). Introduction to machine learning with scikit-learn. In *Kaggle's blog*. https://github.com/justmarkham/scikit-learn-videos

Pascual, D., Pla, F., & Sánchez, S. (n.d.). *Algoritmos de a grupamiento*.

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

Scikit-learn. (n.d.). Retrieved January 30, 2020, from scikit-learn.org

The scikit-yb developers. (2019). *Elbow Method*. https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

Vásconez, P., Medina, G., & Hofstede, R. (2001). Los Páramos del Ecuador. *Botánica Económica de Los Andes Centrales*, *2006*, 91–109.

Verga, A., Navall, M., Joseau, J., Royo, O., & Degano, W. (2009). en las regiones fitogeográficas Chaqueña y Espinal norte de. *Quebracho*, *17*, 31–40.

ISSN: 2600-5859
Vol. 3, N°1.1, p. 224-238, febrero, 2020

PARA CITAR EL ARTÍCULO INDEXADO.

Paguay Soxo, P. X., Idrobo Cárdenas, J. X., Buñay Guisñan, P. A., & Flores Orozco, A. P. (2020). Análisis de patrones de características de especies andinas de las reservas Chimborazo y Sangay utilizando el método k-means clustering. *ConcienciaDigital*, *3*(1.1), 224-236. https://doi.org/10.33262/concienciadigital.v3i1.1.1143