

Clasificación de especies de Frailejón del sector del páramo del Sumapaz mediante la morfología foliar

Juan David Leal Campuzano

Universidad Nacional de Colombia



1 INTRODUCCIÓN

Las especies a menudo se consideran unidades básicas de diversidad biológica en ecología, evolución, biogeografía y biología de la conservación [3]; la intervención de la mano del hombre y el cambio climático global está afectando el ecosistema de los páramos en Colombia y en general del territorio suramericano, esto está generando que un género de plantas que tiene gran predominio en estos terrenos y que además sirven como medio de tratamiento del agua y retención de minerales en el suelo, presenten cambios debido a una alteración de su ecosistema. Existe una necesidad en distinguir el ecosistema en donde se trabaja, esto genera la necesidad de identificar las diferentes especies que componen el mismo, en el presente proyecto y dado el predominio del frailejón en los páramos, se creará un algoritmo que permita clasificar tres distintas especies de frailejón endémicas del páramo del Sumapaz a través de la morfología de sus hojas: *E. Grandinflora*, *E. Aregentea* y *E. Summapacis*, especies predominantes en este lugar.

En un proceso de estos de clasificación, hay una alta demanda de tiempo y costos, la aplicación de este algoritmo permitirá a los especialistas contar con ayuda al momento de clasificar entre especies de frailejón y tener una mayor certeza, por otro lado, en la literatura no existe un algoritmo que se especialice en plantas de esta especie, por ende será algo nuevo y de gran ayuda para los investigadores; el desempeño a medir en este algoritmo y lo que concierne a los investigadores tiene que ver con la precisión del modelo, es la finalidad del proyecto a nivel cuantitativo.

Los datos en bruto consisten en 180 fotografías tomadas en el páramo del Sumapaz a las cuales se les debió realizar un preprocesamiento específico y además se debieron obtener ciertos tipos de características de las imágenes preprocesadas para poder entrenar el algoritmo [4].

Los resultados que se obtuvieron parten de 4 diferentes algoritmos, tres de ellos presentaron una precisión que osciló entre el 80% y el 84% al momento de clasificar cada especie, la otra tuvo una precisión que osciló entre el 60% y el 70%, los primeros tres arrojaron una precisión prometedora teniendo en cuenta la calidad de los datos y el parecido existente entre la especie *Grandinflora* y la especie *Summapacis* lo cual dificultó la identificación entre ellas, los algoritmos tuvieron buena precisión al identificar la especie *Argentea* que se diferenciaba bastante bien de las otras por su forma.

El proyecto tuvo buenos resultados a pesar de las complicaciones iniciales con las fotografías, se espera que el algoritmo pueda ser generalizado a otras especies dado que parte del algoritmo es la extracción de características, algo que no depende enteramente de los datos iniciales que son las hojitas preprocesadas, esto quiere decir que dado el caso en que obtengan nuevas fotografías de otras especies, quizás el modelo permita de igual manera distinguirlas y clasificarlas, lo único faltante sería tener esas nuevas hojas previamente clasificadas por el experto para así poder entrenar de nuevo el modelo.

2 DATOS, MÉTODOS Y RESULTADOS

2.1 DATOS

Para tal estudio, en principio se han obtenido un conjunto de 180 fotografías tomadas en el páramo del Sumapaz [4] en las que se pudieron encontrar alrededor de 4 o 5 hojitas por fotografía, para el preprocesamiento de estas fotografías se decidió eliminar el fondo y separar la hojas en imágenes por unidad, este trabajo tuvo un alto grado de complejidad ya que las fotografías originales no tuvieron unos estándares al ser tomadas, por ende, tamaños y colores se perdieron o se transformaron en preprocesamiento de las fotografías originales.

Después del preprocesamiento se continuó con la extracción de características de las imágenes; al ser figuras cerradas se pensó en hallar características representativas de una figura cerrada, se tomó como base la elipse y a partir de esta figura se obtuvieron las siguientes características básicas: área, perímetro, longitud del eje mayor y del eje menor y la medida del eje focal según [2]. En principio se pensó en usar estas características pero, como las imágenes perdieron proporción, tamaño y color en el preprocesamiento entonces se decidió crear nuevas características a partir de las básicas halladas, estas características muestran en el apéndice en el anexo 1:

Como se puede observar, todas las características son razones y por ende pierden la desventaja del tamaño y la proporción y además como un plus son valores pequeños, esto permite al algoritmo tener un control sobre los cálculos

2.2 MÉTODOS

Después de la obtención de las características mencionadas anteriormente, se realizó un análisis de los datos a través de la creación de diferentes espacios de características, se obtiene la matriz de correlación entre los datos y se representan dos a dos observando una separación evidente entre cada especie con algún error e identificando una correlación alta entre algunas características.

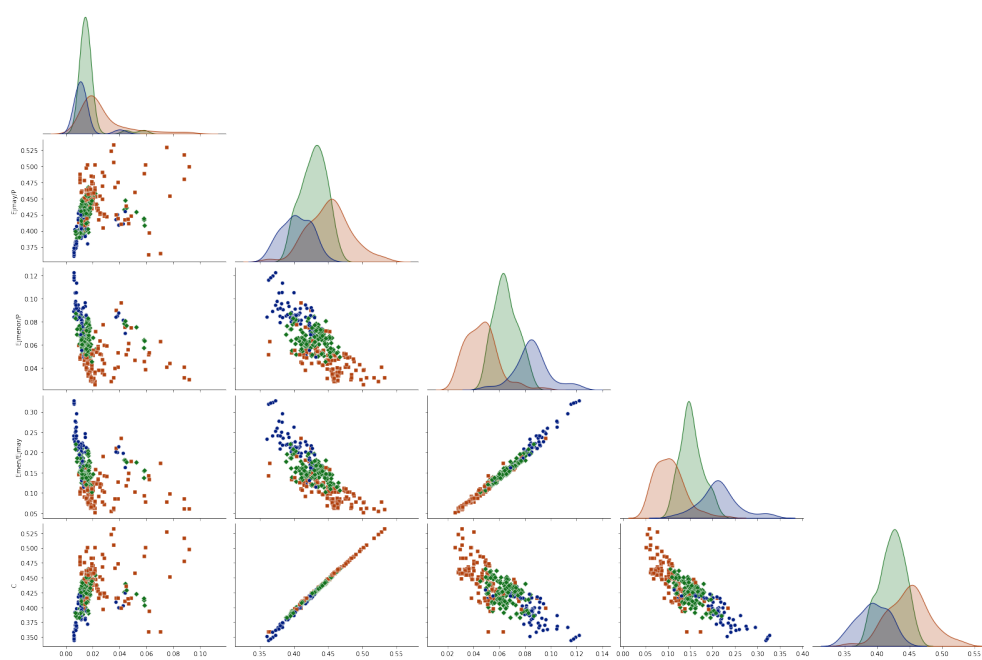


FIGURE 1: REPRESENTACIÓN DE LAS TRANSFORMACIONES DE LAS CARACTERÍSTICAS BÁSICAS EN DOS DIMENSIONES. (GRÁFICA HECHA EN PYTHON, SNS)

En la mayoría de gráficas de la Figura 1 se logra observar una buena separación sin embargo en la figura 2 se

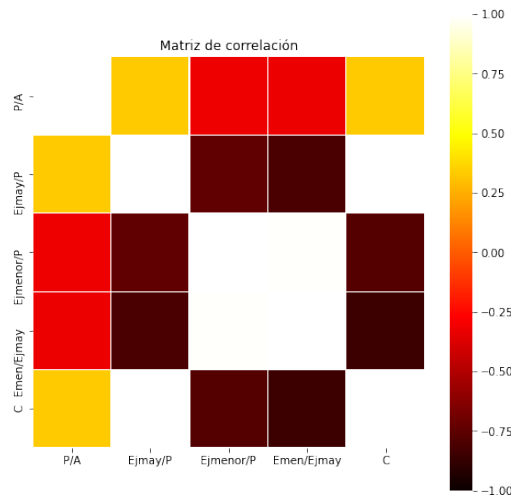


FIGURE 2: MATRIZ DE CORRELACIÓN PARA LOS DATOS BÁSICOS

ve una alta correlación algunas variables.

Se decide crear a partir de las 5 características obtenidas grupos de parejas, triplas, cuartetos y por último las 5 componentes, esto a razón de que si se aplica el algoritmo teninedo una dimesionalidad relativamente alta, por la poca cantidad de datos, quizás el algoritmo no clasifique bien, además se esperaba que usando grupos de características con una baja dimesionalidad se pudiera separar el espacio para hacer más óptimo los modelo.

Para tener una mejor observación den baja dimensionalidad se aplicaron os algoritmo de PCA y TSNE para vi-sualuzar los datos del espacio completo, se observó un buena separación usando PCA, algo que no se obtuvo con TSNE

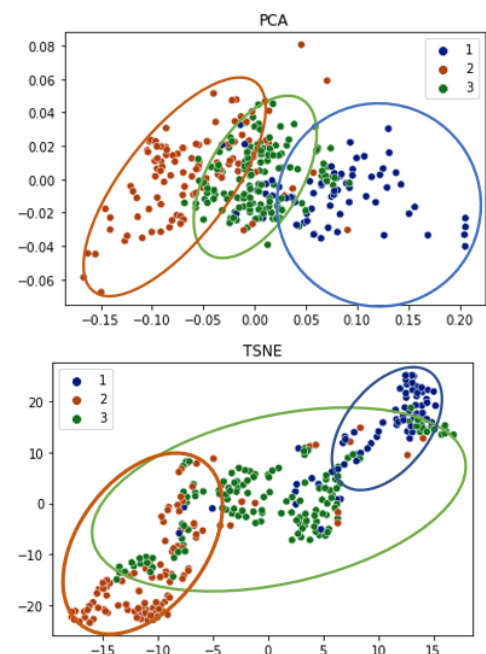


FIGURE 3: PCA Y TSNE SEPARADOS POR GRUPOS DE ESPCIE EN LAS CARACTERÍSTICAS BÁSICAS

Los algoritmos usados para la clasificación fueron: RFC(RandomforestClassifier), ETC(ExtraTreesClassifiers), MLP(Multi-layer Perceptron), KNC(KNeighborsClassifier), modelos escogidos por tener una alto éxito en la clasi-ficación multiclase como se menciona en [1], [7], [5], además que hizo una busqueda hiperparámetros en cada algoritmo, se muestran en el Apéndice en el anexo 2.

2.3 RESULTADOS

En la aplicación de los cuatro modelos al conjunto de características completo se obtuvieron las siguientes matrices de confusión

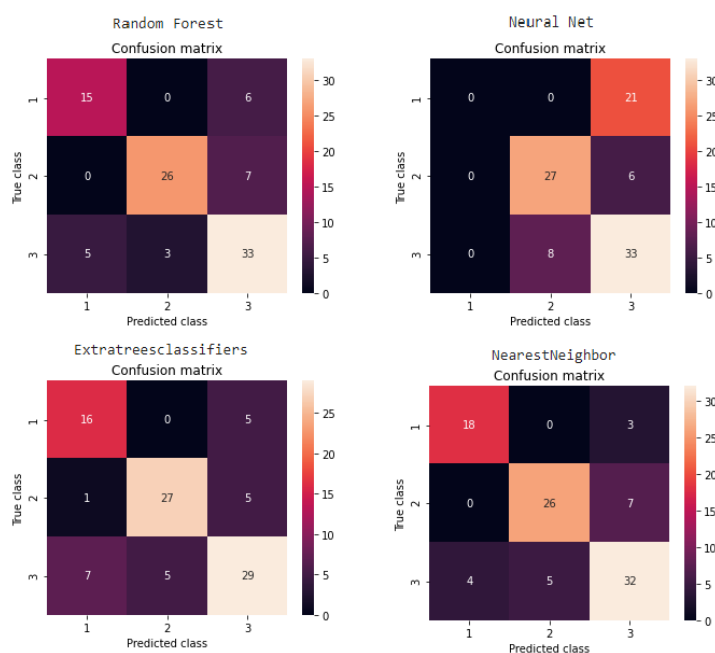
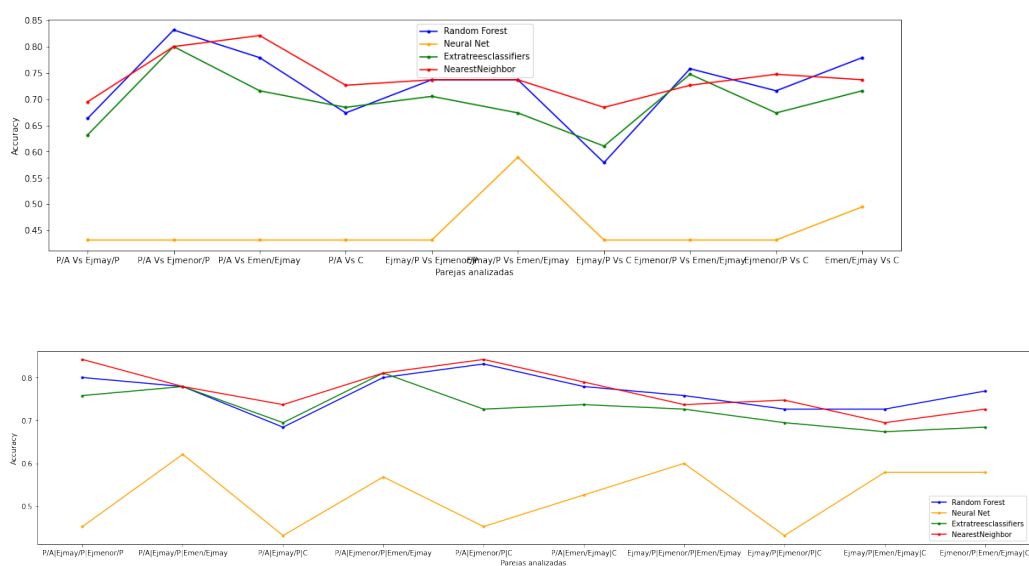


FIGURE 4: MATRIZ DE CONFUSIÓN PARA CADA MODELO ENTRENADO CON EL ESPACIO 1 COMPLETO

Se puede observar que existe una confusión de los modelos entre la especie 1 y la especie 3 (E. Grandinflora y E. Summpacis), la especie 2 (E. Argentea) tiene mejores resultados a pesar de confundirse un poco con la especie 3, también se puede observar el desbalance de los datos y a pesar de eso tener resultados medianamente buenos.

Al entrenar cada modelo con cada sugrupo hecho del espacio completo se obtuvieron las siguientes gráficas que muestran los resultados obtenido



La tabla 1 se muestran los mejores desempeños con los mejores grupos respecto a cada modelo y sus respectivos parámetros

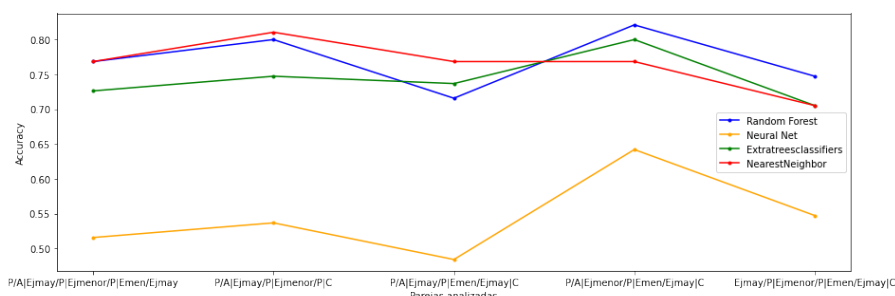


FIGURE 5: ANÁLISIS POR SUBGRUPOS, ARRIBA POR PAREJAS, EN EL MEDIO POR TRIPLETAS Y ABAJO POR CUARTETOS

Grupo	Modelo	Desempeño	Parámetros	Subgrupo
Parejas	RFC	0.83	criterion: entropy, n_estimators: 100	P/A Vs Ejmenor/P
	MPL	0.58	act: tanh, alpha: 0.001, hidden_layer: (128,)	Ejmay/P Vs Emen/Ejmay
	ETC	0.8	criterion: entropy, n_estimators: 20	P/A Vs Ejmenor/P
	KNC	0.82	algorithm: auto, p: 1	P/A Vs Emen/Ejmay
Triplas	RFC	0.83	criterion: gini, n_estimators: 10	P/A, Ejmenor/P, C
	MPL	0.62	act: tanh, alpha: 0.01, hidden_layer: (128,)	P/A, Ejmay/P, Emen/Ejmay
	ETC	0.81	criterion: gini, n_estimators: 20	P/A, Ejmenor/P, Emen/Ejmay
	KNC	0.84	algorithm: auto, p: 1	P/A, Ejmenor/P, C
Cuartetos	RFC	0.82	criterion: gini, n_estimators: 10	P/A, Ejmenor/P, Emen/Ejmay, C
	MPL	0.64	act: tanh, alpha: 0.01, hidden_layer: (128,)	P/A, Ejmenor/P, Emen/Ejmay, C
	ETC	0.8	criterion: gini, n_estimators: 20	P/A, Ejmenor/P, Emen/Ejmay, C
	KNC	0.81	algorithm: auto, p: 1	P/A, Ejmay/P, Ejmenor/P, C

TABLE 1: RESULTADOS DE LOS SUBGRUPOS DEL ESPACIO FORMADO A PARTIR DE LAS CARACTERÍSTICAS BÁSICAS.

3 CONCLUSIONES

El objetivo principal de este proyecto es crear un algoritmo que logre brindar una ayuda al especialista en identificar entre tres especies de Espeletia: E. Grandiflora, E. Argentea y E. Summapacis y ahorre tiempo, materiales y costos al momento de realizar esta clasificación.

Se planteó tener un desempeño de al menos 80%, los modelos de RFC(ExtraTreesClassifiers), ETC(ExtraTreesClassifiers) y KNC(KNeighborsClassifier) han logrado el propósito creado, a pesar de que no se obtuvo una precisión de más del 90%, haciendo una observación sobre la calidad de las imágenes se logró una buena precisión con diferentes combinaciones de cada grupo.

Como se mencionó anteriormente, los algoritmos tiene en cuenta la extracción de características de las imágenes preprocesadas, esto quiere decir que es posible generar un modelo con especies diferentes de Espeletia y quizás se logre obtener de la misma manera tan buen desempeño como el obtenido en este documento, ese será un siguiente objetivo a trabajar.

Se trabajará en aumentar el desempeño mediante la búsqueda de otras características que permitan identificar con un mayor desempeño cada especie, sobretodo las especies en las que el modelo presenta confusión.

4 APÉNDICE

ANEXO 1: CARACTERÍSTICAS OBTENIDAS A PARTIR DE LAS BÁSICAS

$$P/A = \frac{Perimetro}{Area}$$

$$Ejmenor/Ejmay = \frac{Eje_{menor}}{Eje_{mayor}}$$

$$Ejmay/P = \frac{Eje_{mayor}}{Perimetro}$$

$$Ejmenor/P = \frac{Eje_{menor}}{Perimetro}$$

$$C = \frac{\sqrt{Eje_{mayor}^2 - Eje_{menor}^2}}{Perimetro}$$

ANEXO 2: HIPERPARÁMETROS PARA CADA ALGORITMO

- RandomForestClassifier y ExtratreesClasiffier:
 "n_estimators": [5, 10, 20, 50, 100]
 "criterion": ["gini", "entropy"]
- Knearestneighbor:
 "algorithm": ['auto', 'ball_tree', 'kd_tree', 'brute']
 "p": [1, 2, 3, 4]
- NeuralNetwork: "activation": ['logistic', 'tanh', 'relu']
 "hidden_layer_sizes": [(16,), (32,), (64,), (128,)],
 "alpha": [0.1, 0.01, 0.001]

References

- [1] Bobadilla J. *Machine Learning y Deep Learning: Usando Python, Scikit y Keras*, Ra-ma, 2020.
- [2] Burger W., Burger M. *Principles of digital image processing*, Springer 2009.
- [3] M. F. Cárdenas Agudelo *Ecohydrology of páramos in Colombia: Vulnerability to climate change and land use*, Universidad Nacional de Colombia 2016.
- [4] Pineda Yam, Cortés Andrés, Madriñán Santiago, Jiménez Iván *The Nature of Espeletia Species*, 2020.
- [5] Prinzie A., Van den Poel D. *Random Forests for multiclass classification: Random MultiNomial Logit*, Department of Marketing at Ghent University, Ghent, Belgium, 2008.
- [6] Wäldchen J, Mäder P *Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review*, Springer, 2017.
- [7] Shichao Zhang, Debo Cheng, Ming Zong, Lianli Gao *Self-representation nearest neighbor search for classification*, University of Electronic Science and Technology of China, Chengdu, China, 2016.