Rasgos funcionales: clasificando plantas herbáceas y leñosas.

Aprendizaje Automático de Maquina I

Noviembre 2022

Paola A. Matheus Arbeláez

Contenido

01 Contexto

02 Base de Datos

03 Análisis Exploratorio de Datos

04 Modelos de Clasificación y Métricas

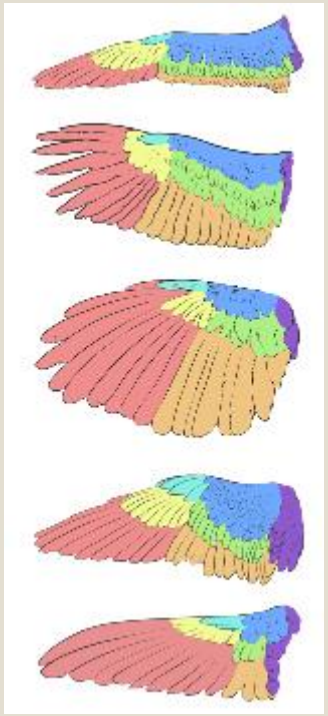
05 Intervalo de Confianza Exactitud

06 Segmentación

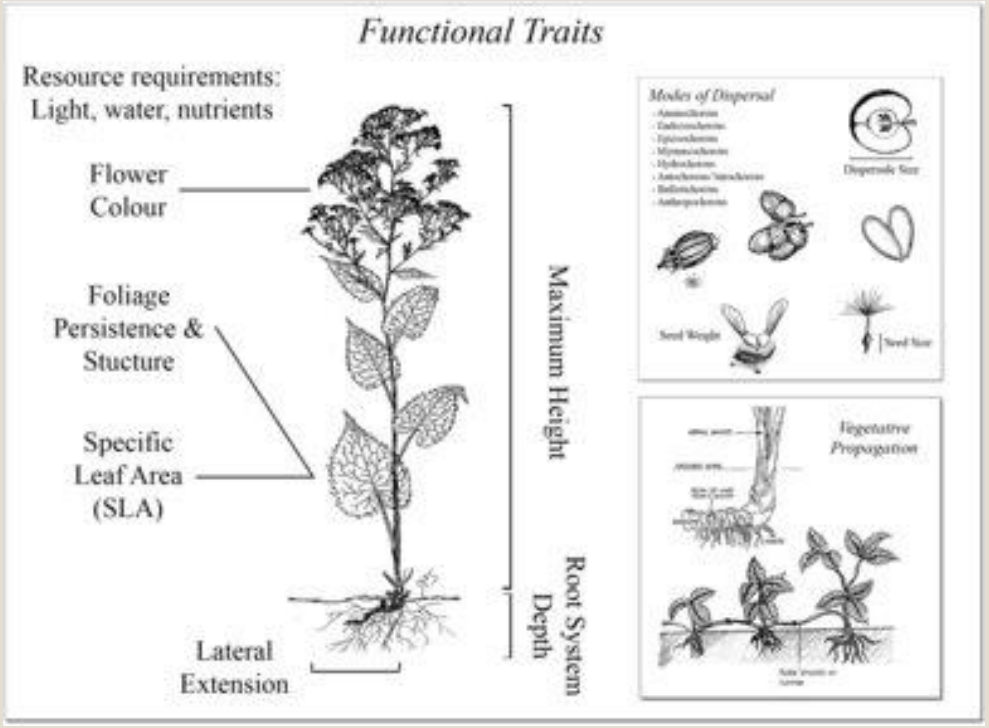
07 Conclusiones

01 Contexto

Los rasgos funcionales son características morfológicas, fisiológicas, fenológicas o comportamentales que se expresan en el fenotipo de individuos y son considerados relevantes en la respuesta de dichos organismos al ambiente. (Violle et al. 2007)



http://www.alithographica.com/



https://www.quora.com/What-are-functional-trait-based-approaches-to-plant-ecology-and-why-are-they-important

Problema de Investigación

- Aprendizaje Supervisado (Datos Etiquetados)
  - Clasificación binaria

Input

Modelo

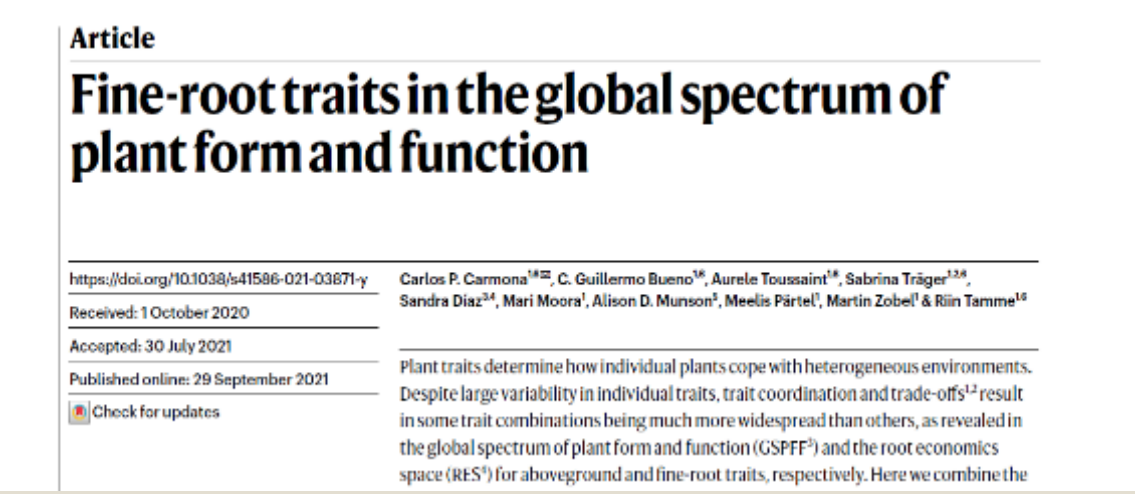
Output

Herbáceas (0)

Leñosas (1)

02 Base de Datos

Carmona et al. (2021) en Nature.



Article

Fine-root traits in the global spectrum of plant form and function

https://doi.org/10.1038/s41586-021-03871-y
Received: 1 October 2020
Accepted: 30 July 2021
Published online: 29 September 2021
Check for updates

Carlos P. Carmona[1,8✉], C. Guillermo Bueno[1,6], Aurele Toussaint[1,8], Sabrina Träger[1,2,6], Sandra Díaz[3,4], Mari Moora[1], Alison D. Munson[5], Meelis Pärtel[1], Martin Zobel[1] & Riin Tamme[1,6]

Plant traits determine how individual plants cope with heterogeneous environments. Despite large variability in individual traits, trait coordination and trade-offs[1,2] result in some trait combinations being much more widespread than others, as revealed in the global spectrum of plant form and function (GSPFF[3]) and the root economics space (RES[4]) for aboveground and fine-root traits, respectively. Here we combine the

Disponible en: **Figshare**

https://figshare.com/articles/dataset/Data_from_Fine-root_traits_in_the_global_spectrum_of_plant_form_and_function_Carmona_et_al_2021_Nature_/13140146

5 archivos .txt

- **Taxonomía:** especie, género, familia, orden.
- **Rasgos aéreos:** área foliar, nitrógeno foliar, altura planta, área especifica foliar, densidad específica de tallo y masa de semilla.
- **Rasgos raíces:** micorrizas, nitrógeno, longitud específica, diámetro, densidad de tejido.
- **Leñosidad:** especie leñosa o herbácea
- **Biomas:** 9 biomas

03 Análisis Exploratorio de Datos

Taxonomía

Leñosidad

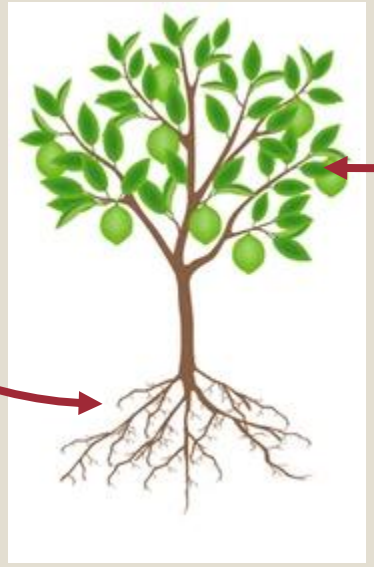
species	biomesKoeppenGroup	woodiness	SRL	D	RTD	N	la	ln	ph	sla	ssd	sm
Abies_alba	Temperate	woody	-0.144232	0.338221	-0.016036	0.111275	57.841588	15.960763	38.552800	6.127230	0.389097	61.477000
Acacia_auriculiformis	Temperate	woody	0.512350	-0.426242	0.001141	0.584477	4000.000000	24.084072	30.000000	9.500000	0.510000	20.014167
Acacia_mangium	Temperate	woody	0.116831	0.073394	0.229927	0.485716	9050.000000	23.797500	10.138776	8.768314	0.491029	14.600000

Dimensiones

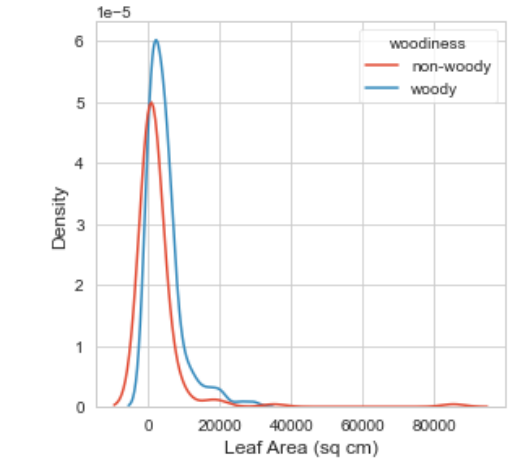
```
df.shape
(296,13)
```

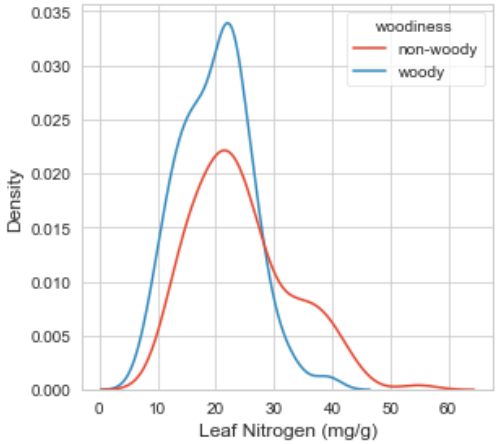
Balanceo de clases

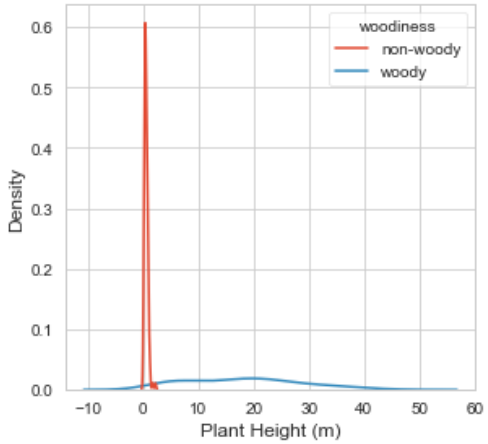
```
df['woodiness'].value_counts()
woody          162
non-woody      134
Name: woodiness, dtype: int64
```

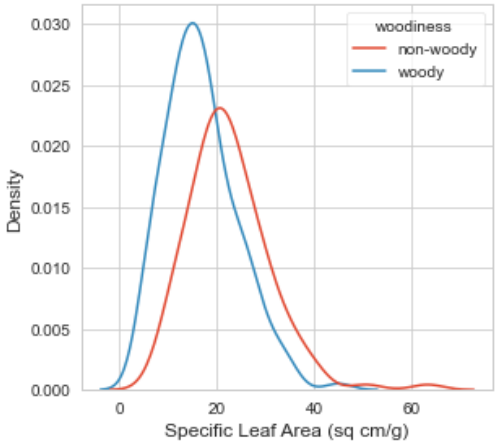


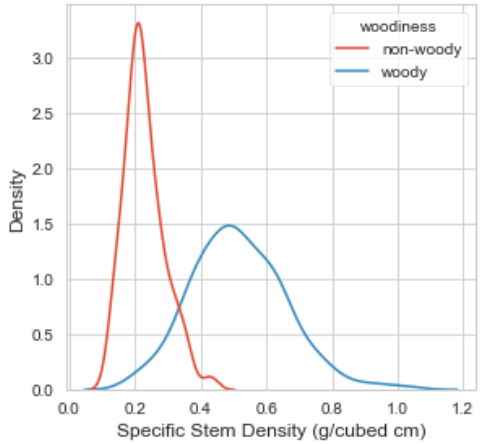
Distribuciones univariadas

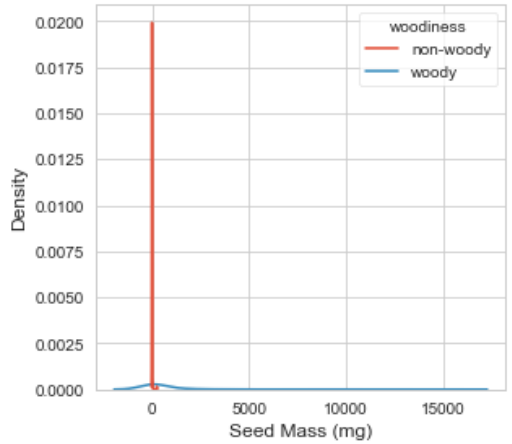


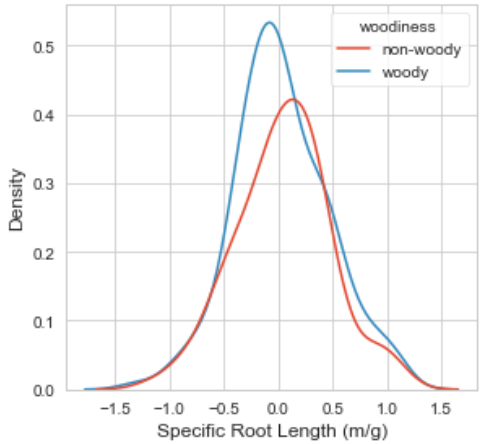


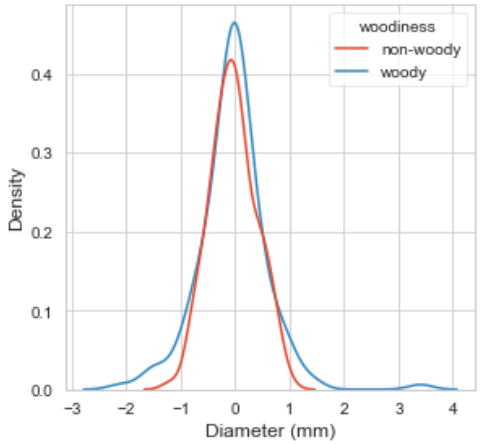


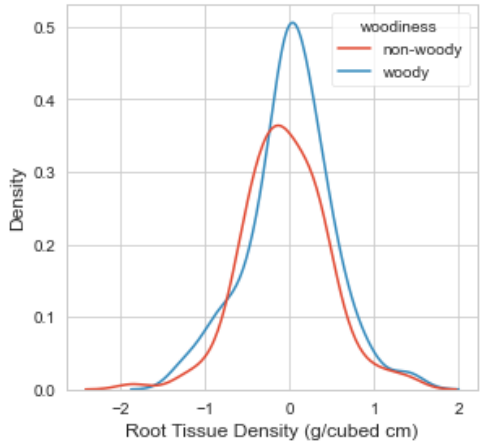


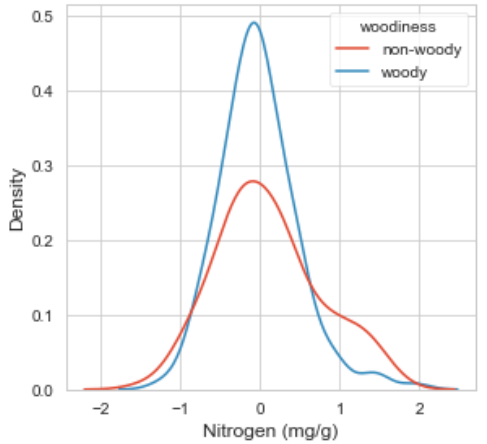




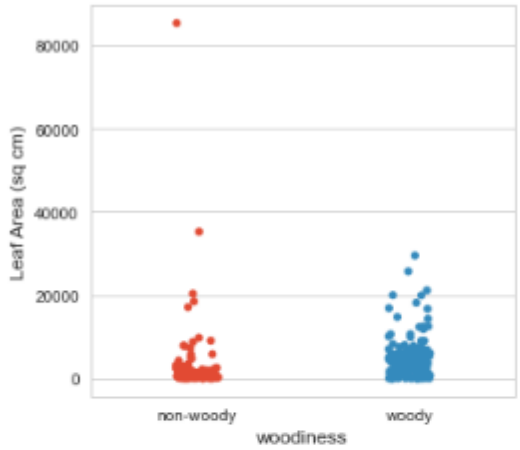


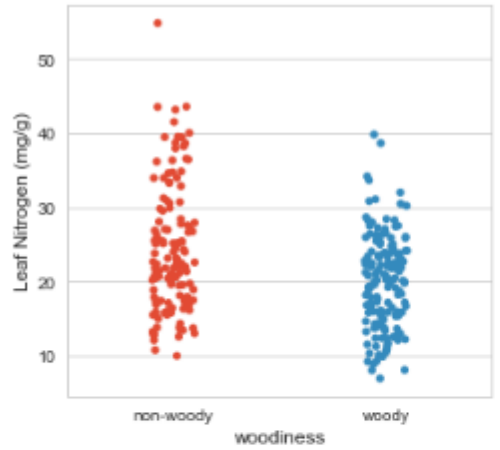


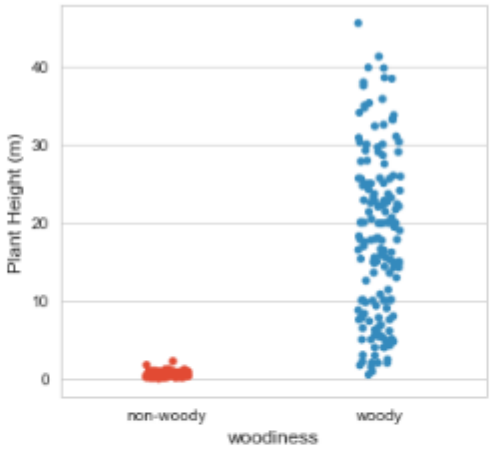


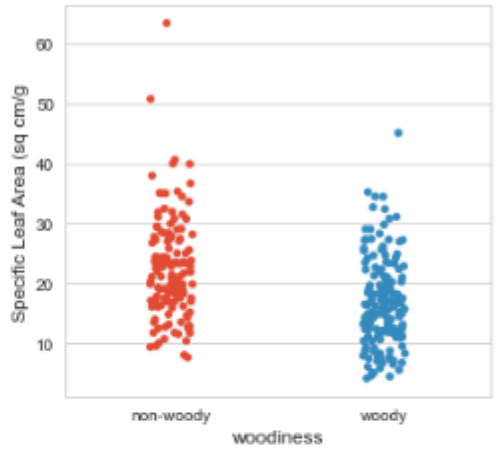


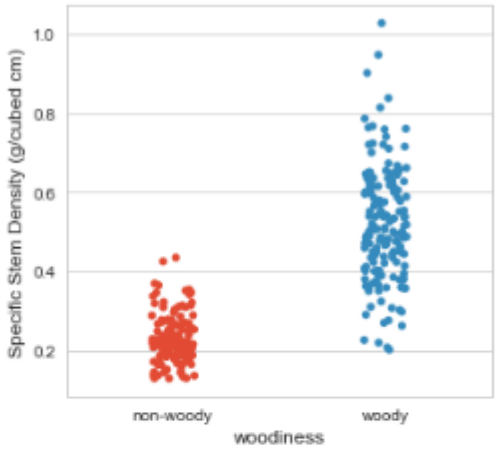
Distribuciones bivariadas

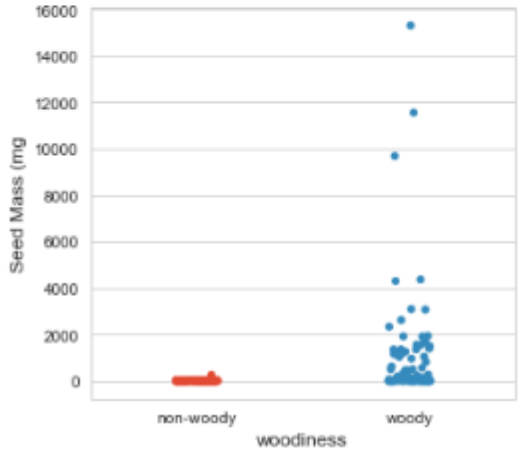


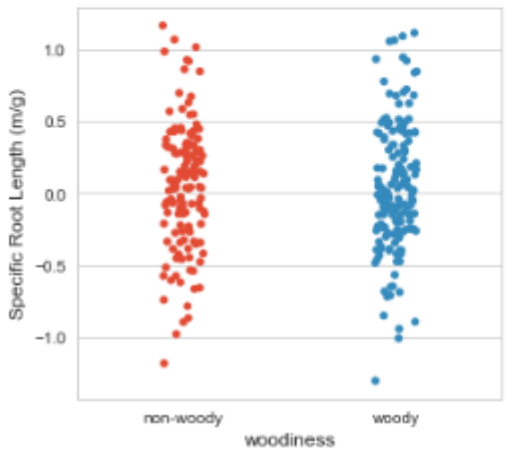


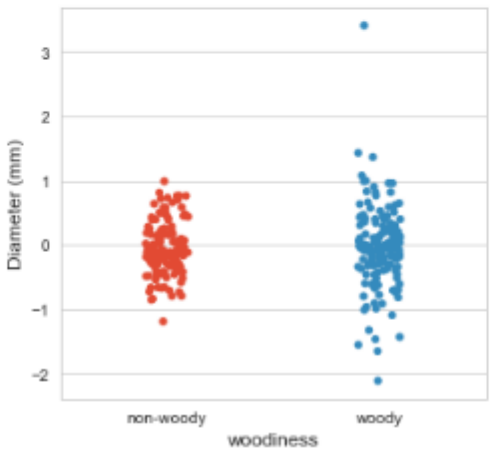


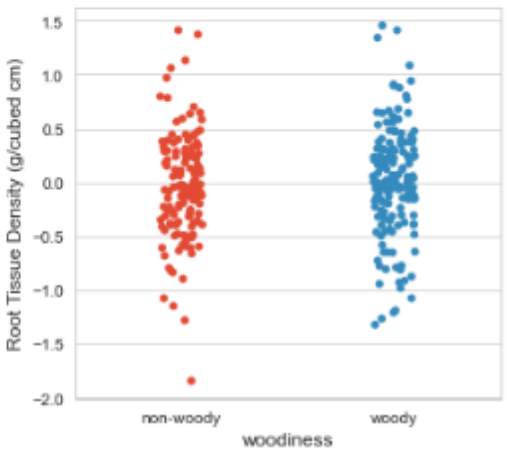


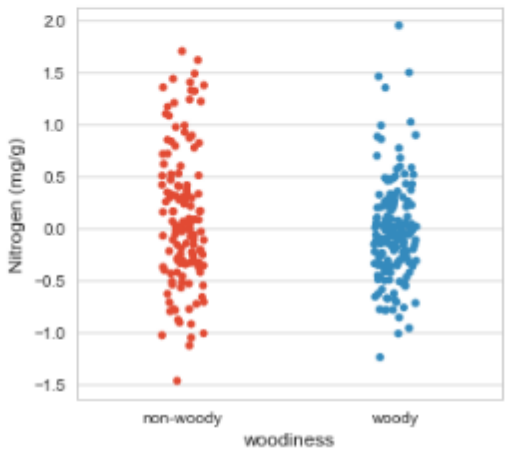




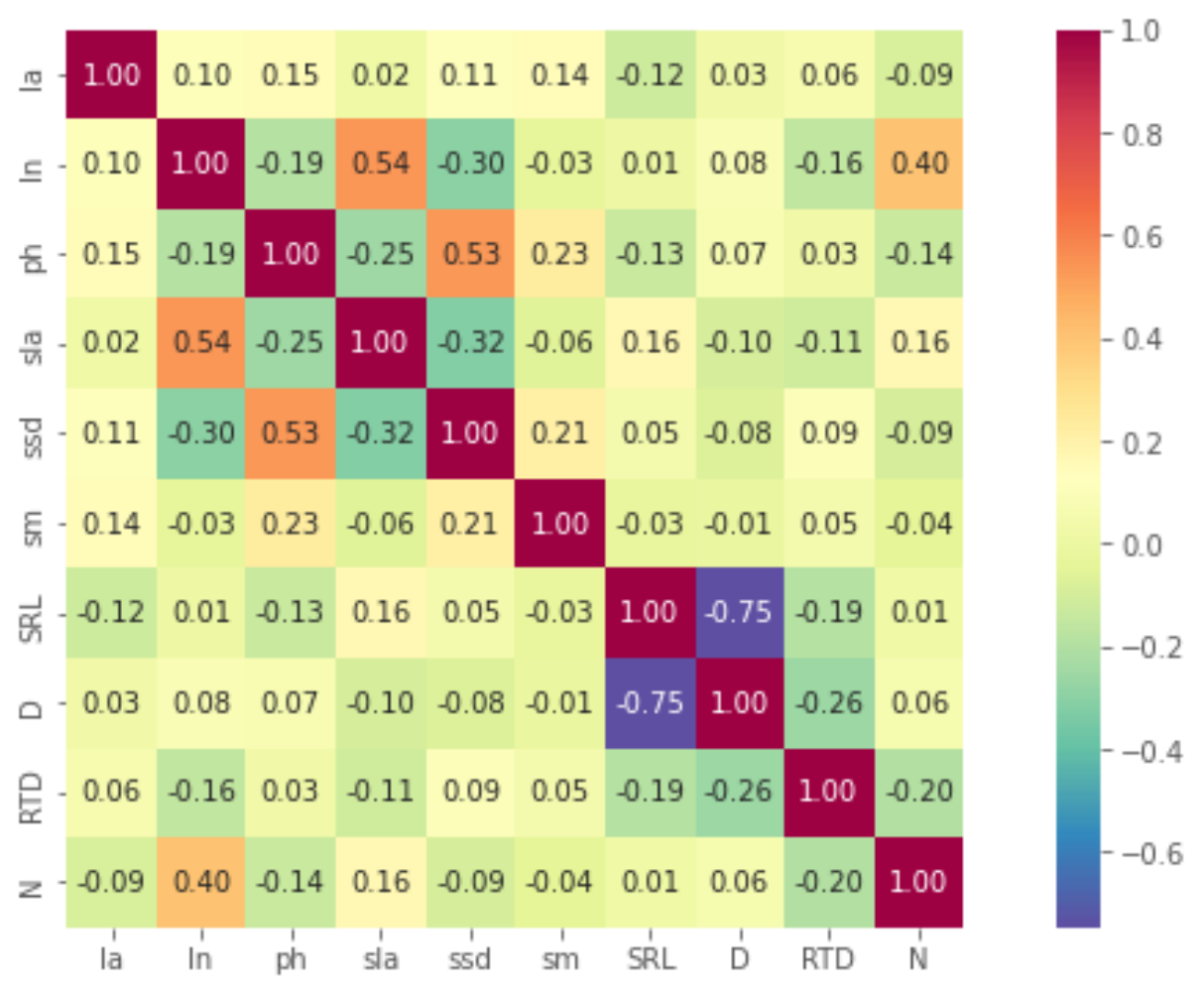






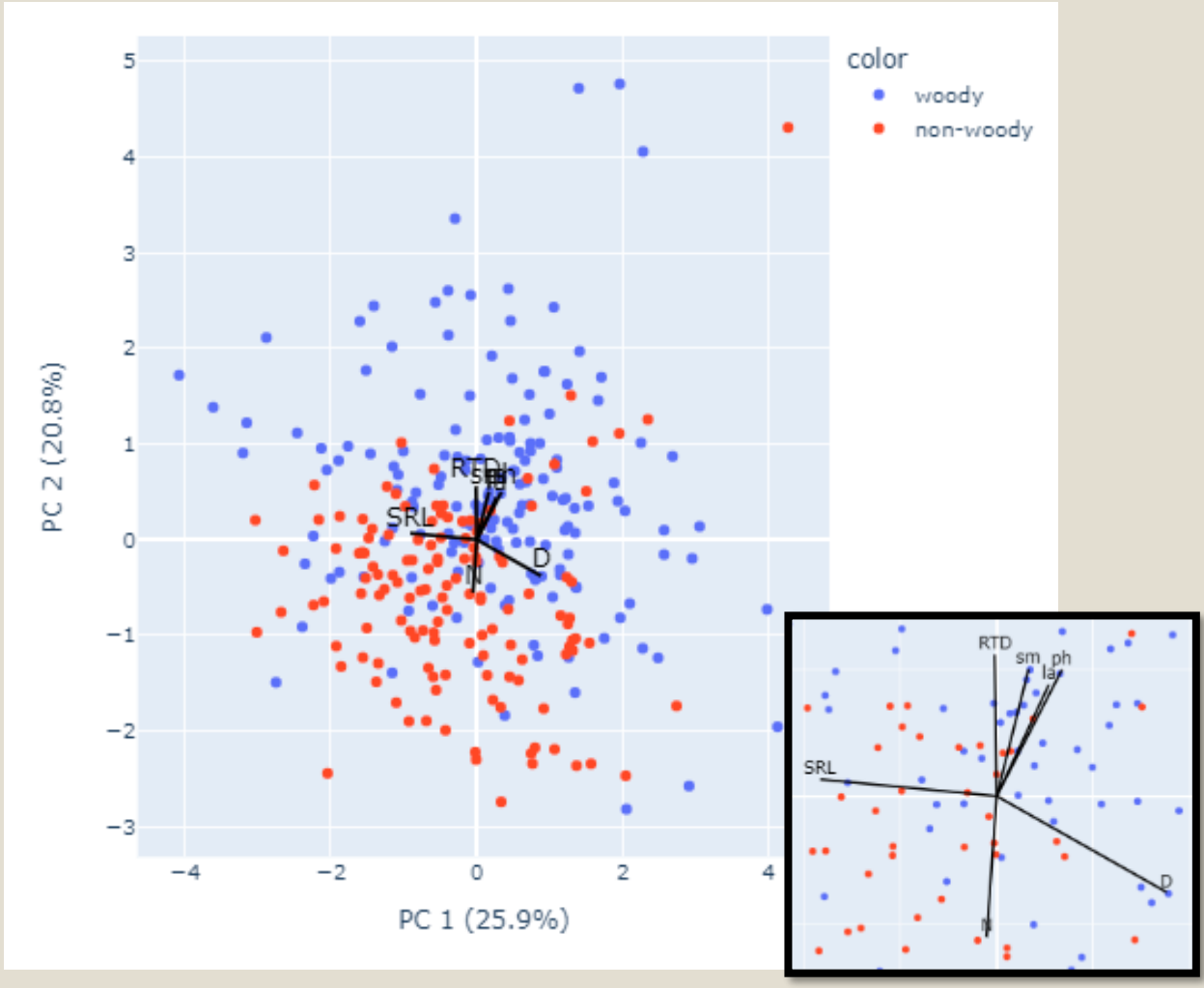


Matriz de Correlaciones



VIF >5
ln, sla, ssd

Visualización PCA (PC1 vs. PC2)



Varianza Explicada 46.70%

04 Modelos de Clasificación

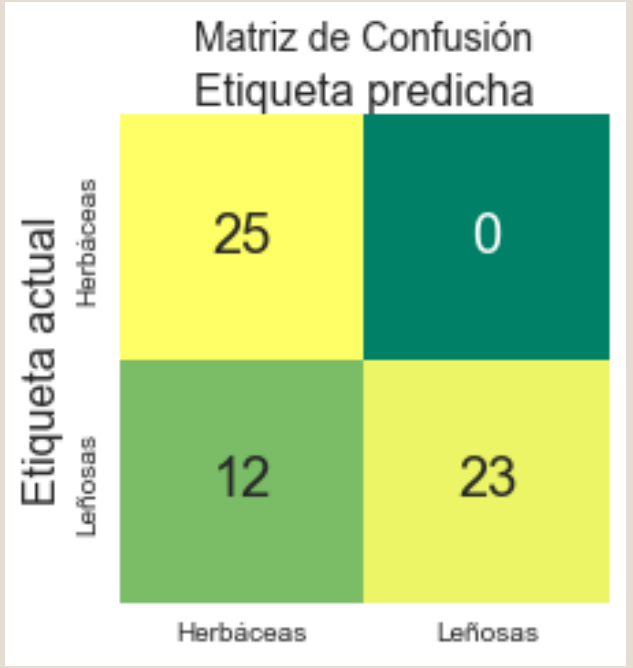
Regresión Logística usando statsmodels

Modelo	Modelo 1 – sm.logit()				
Variables Significativas	Altura de la planta				
Métricas de Evaluación		Precission	Recall	F1-score	Accuracy
	0	0.72	1.00	0.84	**Train:** 0.8563 **Test:** 0.8 **Validation:** 0.8541
	1	1.00	0.67	0.80	

Los *odds* de ser leñosa aumentan por un factor de 29.95 por cada aumento de un metro en la altura de la planta

Variable	Odds Ratio	Efectos Marginales
Longitud raíz (SRL)	1.7121	0.0455
Diámetro (D)	1.5754	0.0385
Densidad tejido (RTD)	1.3294	0.0241
Nitrógeno	1.0351	0.0029
Área Foliar (la)	0.9526	-0.0041
Altura planta (ph)	29.9585	0.2880
Masa semilla (sm)	0.3680	-0.0847

Cuando la altura de la planta aumenta en un metro, **aumenta** la probabilidad de ser leñoso en 28.80%

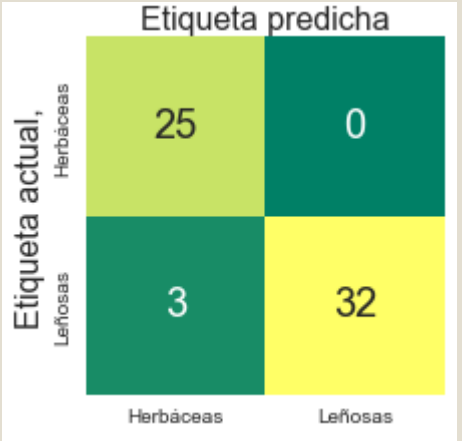


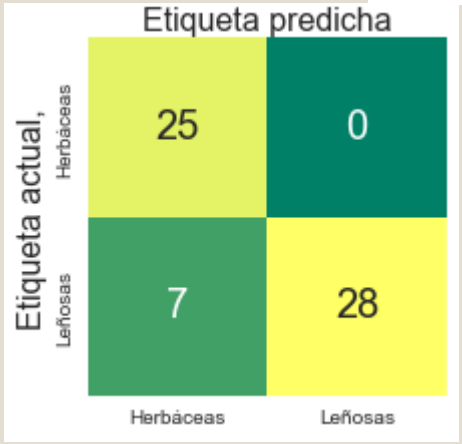
Regresión logística, KNN y Decision Tree usando sklearn

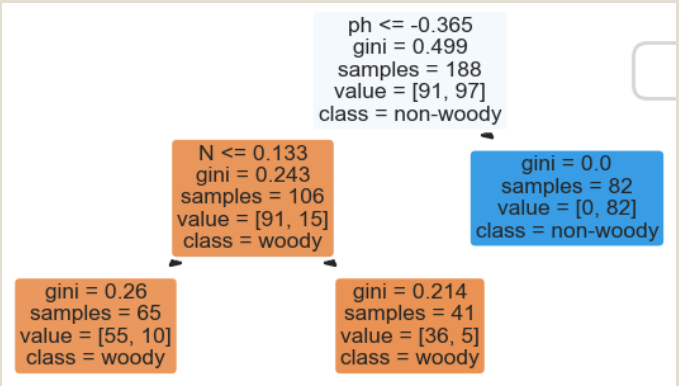
Modelo	Modelo 2 – LogisticRegression()				
Hiperparámetros	C: 1.0 Penalty: l1 Solver: liblinear				
Métricas de Evaluación		Precission	Recall	F1-score	Accuracy
	0	0.89	1.00	0.94	Train: 0.9680 Test: 0.95 Validation: 0.9583
	1	1.00	0.91	0.96	

Modelo	Modelo 3 – KNeighborsClassifier()				
Hiperparámetros	Algorithm: auto N_neighbors: 4 Weights: distance				
Métricas de Evaluación		Precission	Recall	F1-score	Accuracy
	0	0.72	0.88	0.80	Train: 1.00 Test: 0.8958 Validation: 0.8166
	1	0.90	0.77	0.83	

Modelo	Modelo 4 – DecisionTreeClassifier()				
Hiperparámetros	ccp_alpha=0 max_depth=3 min_samples_leaf=0.1				
Métricas de Evaluación		Precission	Recall	F1-score	Accuracy
	0	0.78	1.00	0.88	Train: 0.9202 Test: 0.9166 Validation: 0.8833
	1	1.00	0.80	0.89	



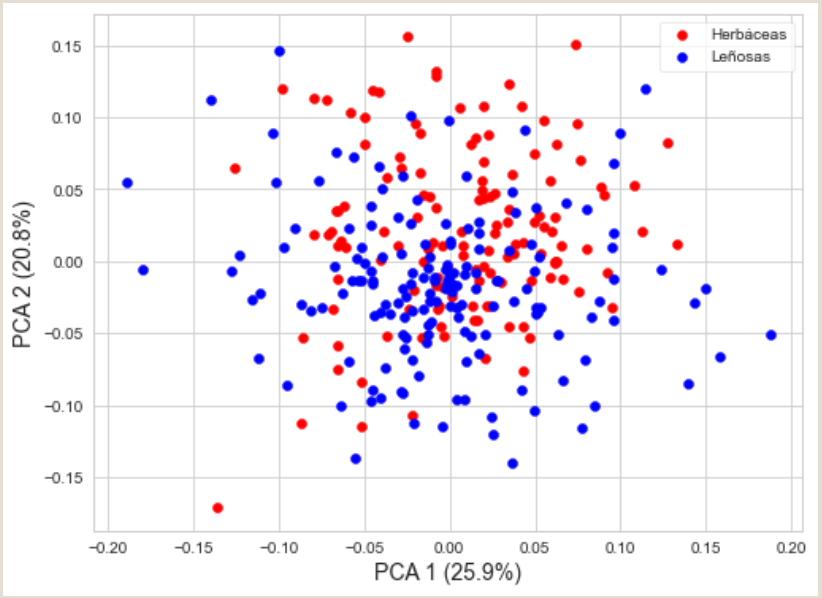


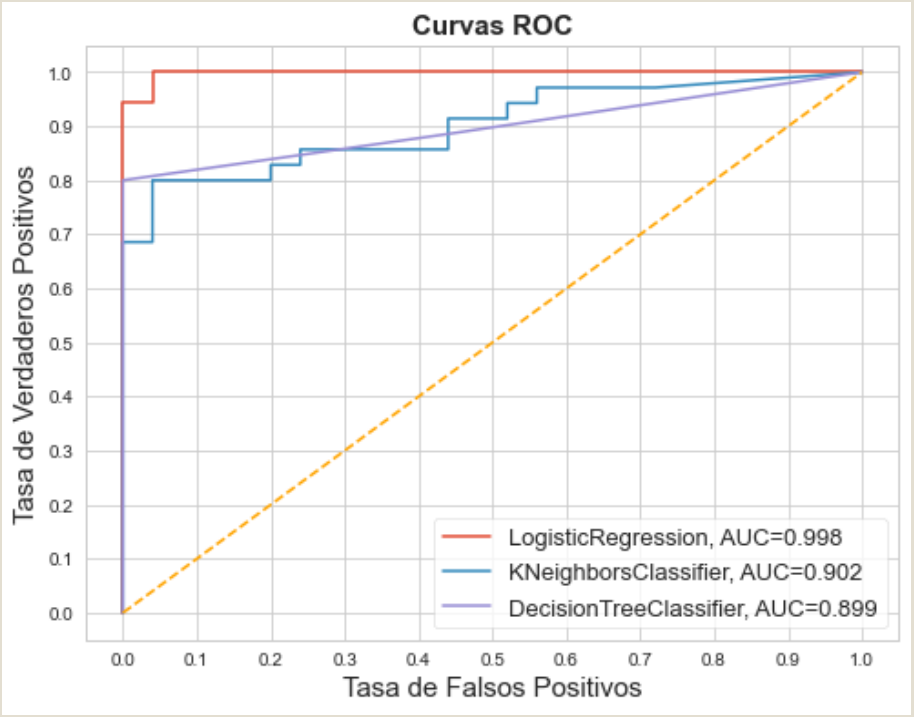


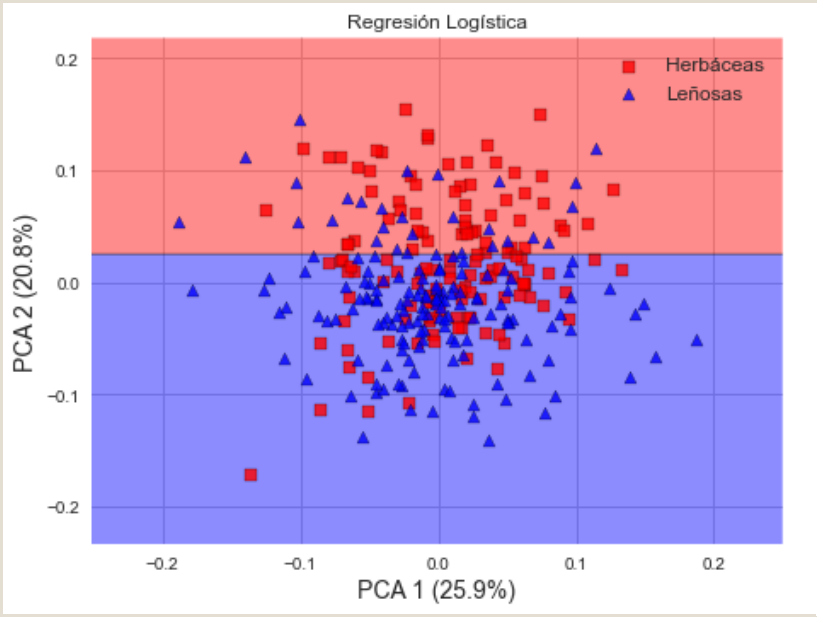
Gráficos de los modelos y curvas ROC

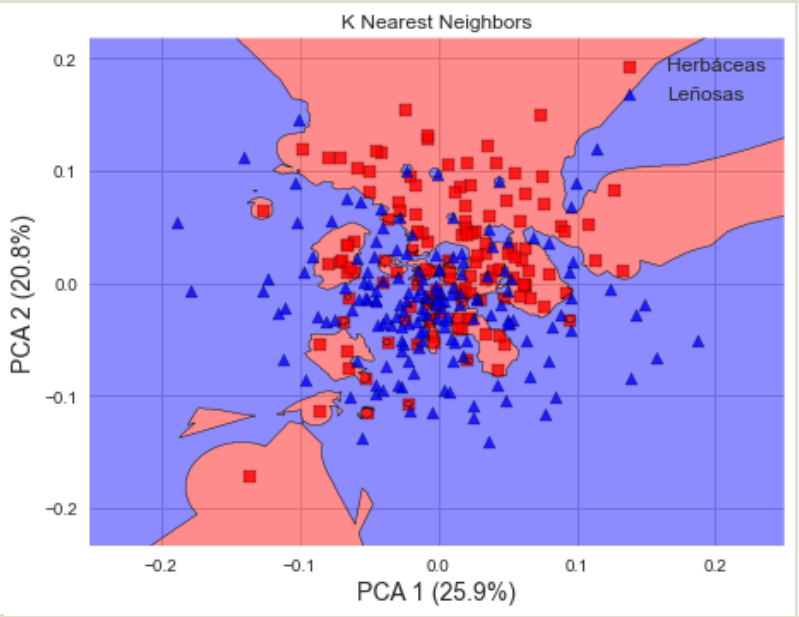
Varianza Explicada

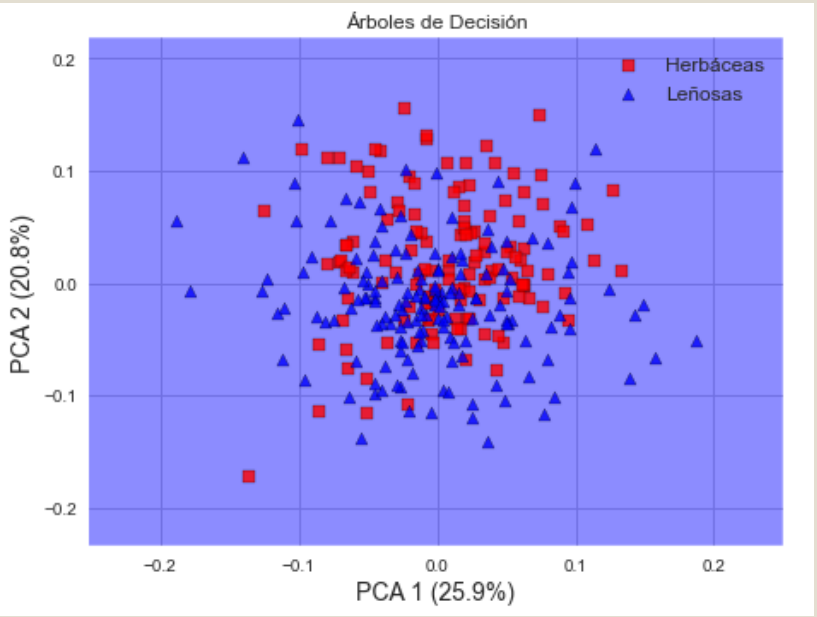
46.70%



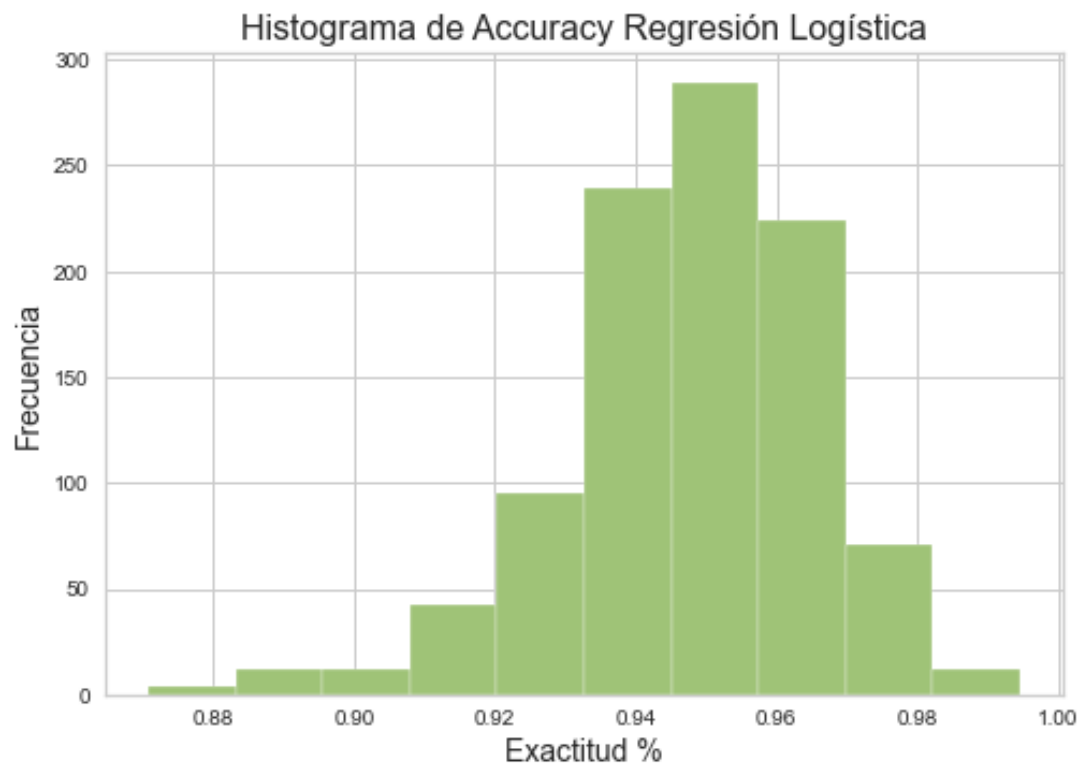








05 Bootstraping



Mejor Modelo

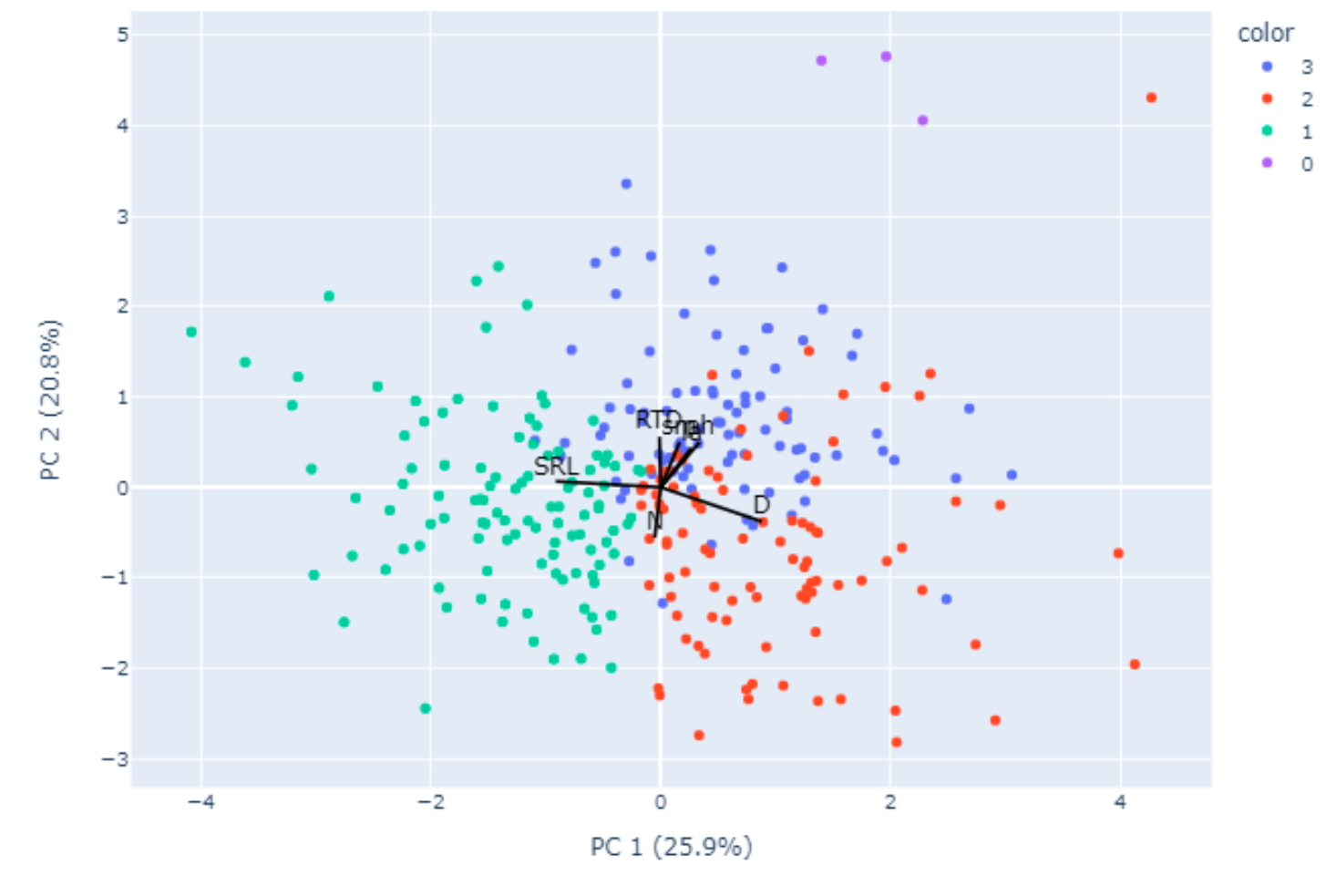
Regresión Logística

- **Hiperparámetros →** C: 1.0, Penalty: l1, Solver: liblinear
- **Accuracy → Train:** 0.9680, **Test:** 0.95, **Validation:** 0.9583
- **AUC:** 0.998

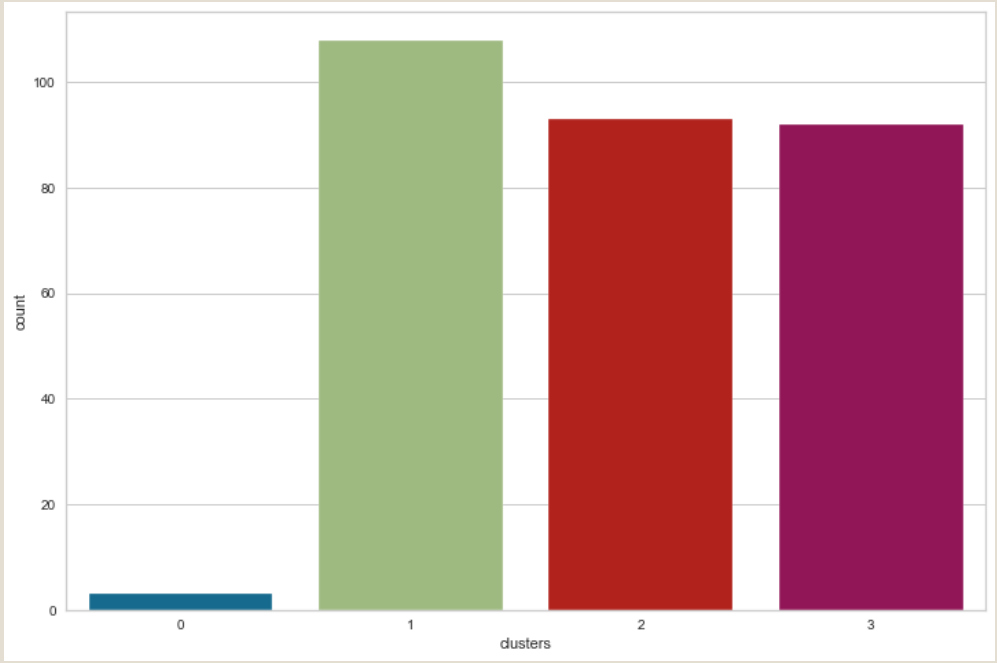
Intervalo de Confianza Regresión Logística:

Accuracy → 90.8% - 97.8%

06 Segmentación de Grupos



	SRL	D	RTD	N	la	ph	sm
0	-0.342524	0.157713	-0.031214	-0.381452	0.973922	1.050966	8.758590
1	0.889856	-0.785612	0.022788	0.099472	-0.276900	-0.591491	-0.209704
2	-0.867518	0.834902	-0.010330	0.156441	0.059796	-0.567144	-0.191367
3	-0.156497	0.073120	-0.015291	-0.262474	0.232852	1.233397	0.154016



07 Conclusiones

El modelo de regresión logística de statsmodel ofrece la posibilidad de calcular estadísticos de interés para evaluar el modelo y tomar decisiones sobre la significancia de las variables.

El análisis de componentes principales permite visualizar la distribución de las clases en un plano bidimensional, sería interesante correr los modelos sobre los datos luego del PCA.

Sería posible trabajar con la base de datos completa si se utiliza un método de interpolación para lidiar con los N/A.

08 Referencias

Carmona, C. P., Bueno, C. G., Toussaint, A., Träger, S., Díaz, S., Moora, M., ... & Tamme, R. (2021). Fine-root traits in the global spectrum of plant form and function. Nature, 597(7878), 683-687.

Código disponible en Github