Rasgos funcionales: clasificando plantas herbáceas y leñosas.

Informe Ejecutivo

Contexto

Los problemas de clasificación en Machine Learning son aquellos que requieren que un conjunto de datos sea clasificado en dos o más categorías. El ejemplo más conocido de aplicaciones de clasificación es con las bases de datos de especies de pingüinos y plantas pertenecientes al género *Iris*. Ambos ejemplos usan como predictoras las características morfológicas para predecir la especie. Sin embargo, es posible predecir características necesarias para la supervivencia de las especies en cierto hábitat cómo lo es la capacidad de los tejidos de producir tejidos lignificados, es decir las plantas leñosas.

Objetivo

En este proyecto se busca clasificar plantas de acuerdo con sus rasgos funcionales, en leñosas o herbáceas. La supervivencia está relacionada con una serie de atributos morfológicos, fisiológicos, fenológicos y comportamentales llamados rasgos funcionales (Violle et al. 2007). Contamos con una base de datos de especies vegetales por lo que se usa una aproximación de aprendizaje supervisado. En este trabajo se presenta una aproximación a la clasificación de plantas en estas dos categorías utilizando como predictores rasgos funcionales de hojas, tallo y raíz.

Método

Obtuvimos una base de datos que contiene datos de rasgos funcionales de 1,719 especies de diferentes biomas del mundo. La base de datos fue limpiada para mantener solo 301 especies con datos de rasgos funcionales empíricos completos, 10 predictores numéricos para el uso en algoritmos de clasificación, un predictor binomial como respuesta y dos predictores categóricos utilizados como identificadores de especie y bioma. Para evaluar multicolinealidad entre las variables se hallaron los índices de correlación y e calculó el factor de inflación de la varianza y se eliminaron de los análisis aquellos con un valor mayor a 5 (James et al, 2017). Para visualizar las dos clases de interés en dos dimensiones, se corrió un análisis de componentes principales. Las variables continuas se escalaron a una media de cero y una desviación de estándar uno. La base de datos fue partida en sets de entrenamiento, testeo y validación. En primer lugar, se corrió un modelo logístico con el paquete statsmodels para hallar los coeficientes del modelo, su valor P asociado, los odds ratio y efectos marginales de cada variable. En segundo lugar se corrieron los modelos de regresión logística, vecinos más cercanos y árboles de decisión del paquete sklearn luego de encontrar los mejores parámetros usando el método GridSearchCV del mismo paquete. Para evaluar el desempeño de cada modelo se halló la matriz de

confusión, el reporte de clasificación, y la curva ROC. Luego, se visualizó el área de decisión de cada modelo sobre un gráfico bidimensional de los dos primeros componentes principales. El mejor modelo fue utilizado para estimar un intervalo de confianza del 95% de la métrica de exactitud. Finalmente, para segmentar el conjunto de datos en grupos usando el algoritmo kmeans del paquete sklearn.

Resultados

De las variables predictoras, tres presentaron un VIF mayor a 5 y fueron retiradas de la base de datos antes de correr los modelos clasificadores. El análisis PCA muestra que los cuatro primeros componentes explican el 75.34% de la varianza. Desde el punto de vista estadístico el análisis de regresión logística indica que únicamente la variable altura de la planta es significativa, y el chance de ser leñosa aumenta por un factor de 29.95 cuando la planta aumenta su altura en un metro. Desde el punto de vista de aprendizaje automático, el modelo que presentó los mejores resultados fue el de regresión logística, con una exactitud de 95.83% en el set de validación con una menor tasa de falsos negativos con respecto a los otros modelos, y un área bajo la curva de 0.998 (ROC). Al estimar el intervalo de confianza de la exactitud del modelo muestran que hay un 95% de probabilidad que el intervalo (90.8% - 97.8%) capture la verdadera habilidad clasificatoria del modelo.

Conclusiones

En este proyecto evaluamos el desempeño de tres modelos de clasificación para asignar especies vegetales a las clases leñosa o herbácea de acuerdo con sus rasgos funcionales. 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 variables sobre la varianza explicada por el modelo. Por otro lado, los modelos de ML son útiles para asignar nuevos datos no vistos por el modelo a alguna de las dos clases de interés. El análisis de componentes principales permite visualizar la distribución de las clases en un plano bidimensional, pero las conclusiones extraídas de estos gráficos deben hacerse con precaución ya que solo estamos capturando el 46.70% de la varianza en las dos primeras dimensiones.

Referencias

Carmona et al (2021) Fine-root traits in the global spectrum of plant form and function. Nature. Vol 597

James G, Witten D, Hastie T, Tibshirani R. (2017) An Introduction to Statistical Learning: With Applications in R. Corr. Springer.

Violle, C., Navas, M. L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., & Garnier, E. (2007). Let the concept of trait be functional!. Oikos, 116(5), 882-892.