Actividad: uso de framework o biblioteca de aprendizaje máquina para la implementación de una solución.

Inteligencia Artificial para la ciencia de datos

Leonardo Gracida Muñoz A01379812

Profesor: Jorge Adolfo Ramírez Uresti

Fecha de entrega: Lunes 11 de Septiembre de 2022

- Modelo utilizado:

En este caso como la dataset que estamos utilizando son datos que se pueden separar de una manera fácil y rápida, ya que debemos separar los hongos entre hongos venenosos y hongos comestibles, pero por fortuna los hongos venenosos son parecidos entre ellos, como lo son también los hongos comestibles. El Random Forest es una buena opción para clasificar solo dos clases teniendo solo variables categóricas. Además de que no vamos a necesitar un árbol muy grande, por lo fácil que se podrían separar los datos.

- Pruebas cambio de tamaño train y test:
- Train 70, Test 30:

Train acc: 1.0

Val acc: 1.0

Test acc: 1.0

- Train 80, Test 20:

Train acc: 0.9952128568986152

Val acc: 0.9892307692307692

Test acc: 0.9969230769230769

- Train 90, Test 10:

Train acc: 0.9978720170238639

Val acc: 0.9986338797814208

Test acc: 0.998769987699877

- Train 60, Test 40:

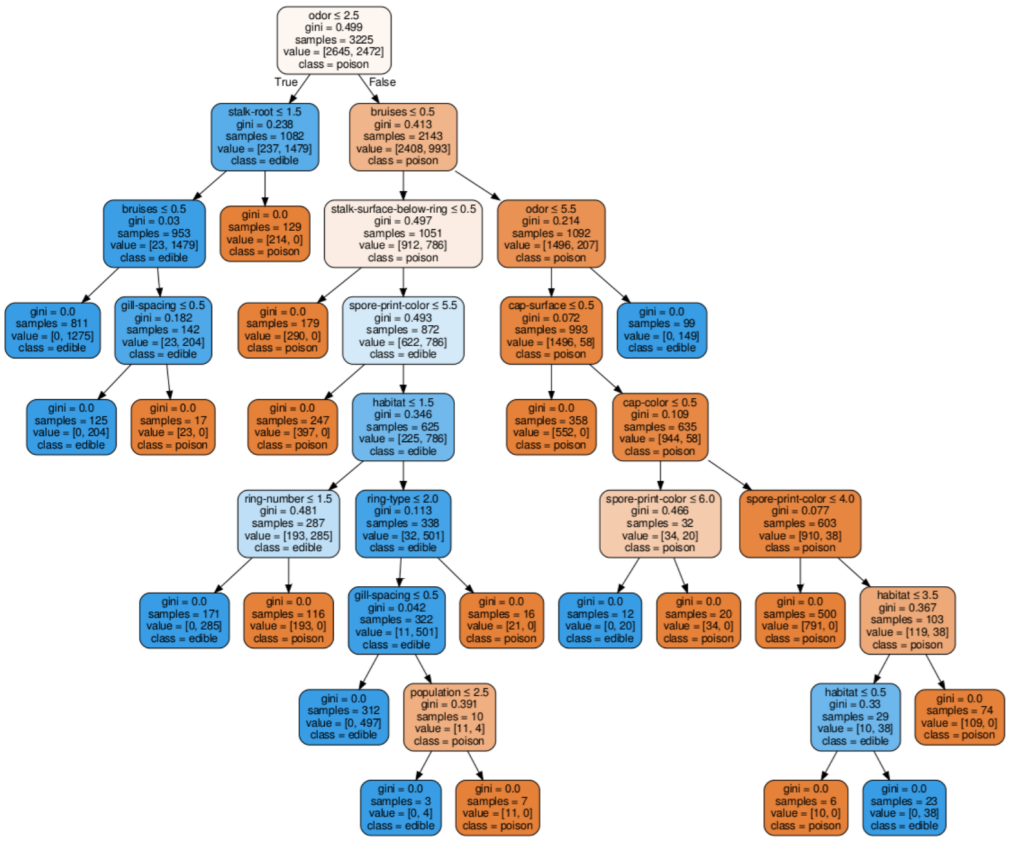
Train acc: 0.9990880072959416

Val acc: 0.9979508196721312

Test acc: 0.9990769230769231

Al ver el resultado de estas pruebas vemos que la mejor separación de datos es el de 30% de test, ya que es el que obtiene la acc cercana a 1 o que prácticamente es uno, ya que al hacer un test más pequeño entramos en un poco en overfitting y perdemos acc en la parte del test, y si la hacemos de 40% o más entramos en underfitting..

Es por eso que dejamos 30% de test como el default a usar, al escoger esa proporción y usando 15 estimadores, con un máximo de 20 hojas de profundidad obtenemos el siguiente árbol.



Como podemos ver este árbol no es muy grande, ya que es el árbol completo generado, mostrando que los hongos se pueden separar fácilmente, vemos que va obteniendo características de hongos venenosos, pero puede que una sola ya lo declare como comestible de manera automática sin tener que checar otras características, como también en el sentido contrario al ver un hongo que se va viendo que es comestible y que puede ser venenoso teniendo una o varias característica, pero no muchas.

- Pruebas de cambio de hiper parámetros:

En este caso vamos a variar el número de estimadores y el máximo de hojas.
En este caso ya vimos que obtuvimos un accuracy prácticamente perfecto o muy cercano a cero, vamos a checar el mínimo de estimadores y hojas máximas para poder obtener esa accuracy. En todas estas pruebas vamos a usar 30% de test.

- Prueba com n-estimators = 10, max leaf = 5:

Train acc: 0.9646277115497361

Val acc: 0.9543057996485061

Test acc:, 0.9569319114027892

- Prueba com n-estimators = 5, max leaf = 5:

Train acc:, 0.9777213210865742

Val acc: 0.9630931458699473

Test acc: 0.9774405250205086

- Prueba com n-estimators = 5, max leaf = 10:

Train acc: 0.9841704123509869

Val acc: 0.9806678383128296

Test acc: 0.9819524200164069

- Prueba com n-estimators = 10, max leaf = 15:

Train acc: 0.9912057846394372

Val acc: 0.9912126537785588

Test acc: 0.9909762100082035

- Prueba com n-estimators = 15, max leaf = 20:

Train acc: 1.0

Val acc: 1.0

Test acc: 1.0

Como en este caso debemos saber si un hongo es venenoso o no, debemos obtener una precisión en el test lo más alta posible, ya que decimos si un hongo se puede comer o no, al inicio vemos que la precisión en el train es de 1.0, podemos pensar que estamos haciendo

overfitting, pero al hacer las pruebas en el test y validation, vemos que obtenemos todas las pruebas correctas, mostrando que el clasificar hongos no es muy difícil, no solamente usando una técnica de aprendizaje supervisado, sino que también es fácil haciéndolo a la vista, ya que teniendo cosas como el color o tamaño podemos declarar si son venenosos o no.
Además de que al hacer las diferentes pruebas cambiando el tamaño del árbol, obtenemos una buena accuracy arriba del 0.95, pero al hacer más grande el árbol, lo permitimos aprender más las características de los hongos más neutros o más difíciles de clasificar, que al parecer no son muchos en la dataset, ya que al hacer un árbol pequeño, podemos clasificar bien la mayoría de los hongos.