

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

PARCIAL #1

INTRODUCCIÓN AL APRENDIZAJE AUTOMÁTICO

NOMBRE DEL ALUMNO: MELISSA AIMÉ LUNA MARROQUÍN

MATRÍCULA: 1810668

GRUPO: 011

MONTERREY, N.L., MÉXICO. 31 DE OCTUBRE 2020.

**Escenario #2**

Tenemos el dataset llamado “mushroom.csv” donde tenemos valores o características de diferentes hongos y nuestro objetivo será categorizar o dividir los hongos entre los que son nocivos o no para el consumo.

Lo que podemos ver en el dataset es que tenemos muchas características como si tienen anillos, de qué color son, básicamente todas las características que puede tener un hongo.

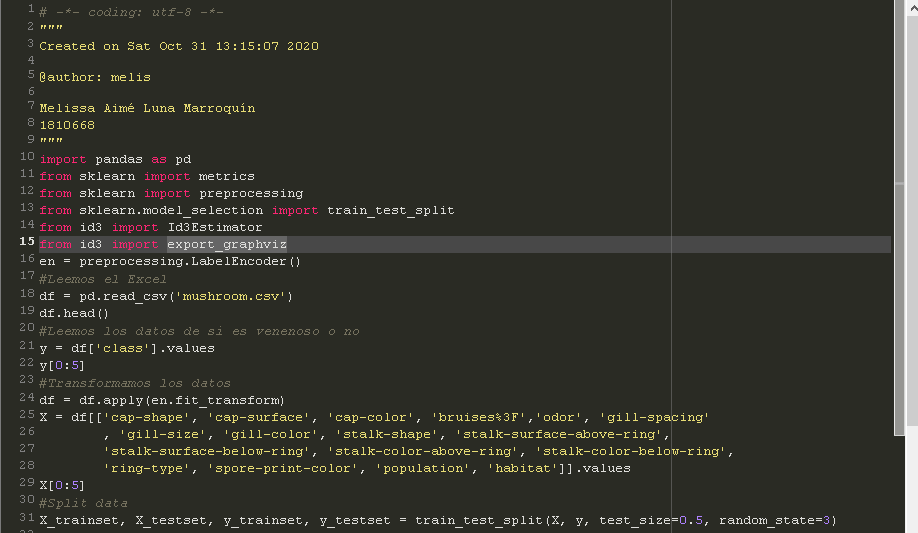
Lo primero que pude ver en el dataset es que había una columna que tenía muchos ‘?’, así que decidí eliminarla, además de que no era necesario los datos para predecir el resultado. También eliminé la columna donde mencionaban los números de anillos porque si tienen 1, podrían ser venenosos o no, entonces no nos servía de mucho, al igual que las columnas de las veil color y veil type porque todos daban el color blanco o decían que era partial entonces eliminé esas dos columnas. Y por último eliminé el gill-attachment que también cambiaba muy muy poco sus valores y es muy difícil distinguir si un hongo es venenoso o no con sólo ver ese atributo.

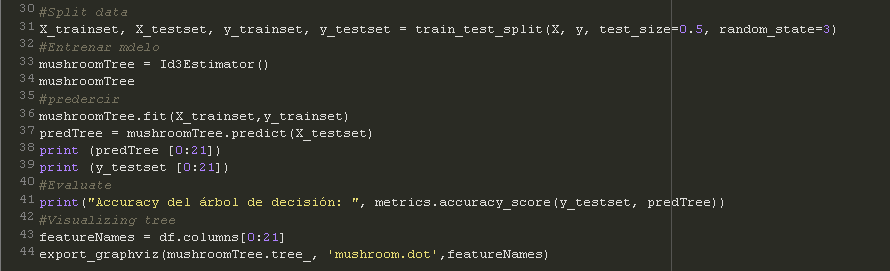
Elegí escoger el método de árbol de decisión porque como nuestro resultado es un sí o no (venenoso o no), podíamos verlo como dos ramas de decisión al final del árbol de decisión, al igual que las otras características del hongo, por ejemplo, podríamos preguntar: ¿el hongo tiene “tal” tipo de anillo? Y el resultado seguía siendo ramas de decisión de sí o no.

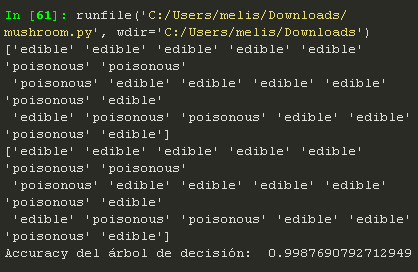
El modelo lo hice en Python, lo implementé en los dos códigos que habíamos utilizado en clase y utilicé como extra una librería de sklearn para poder cambiar los datos de categóricos a numéricos y no tener problemas al momento de ejecutar el programa.

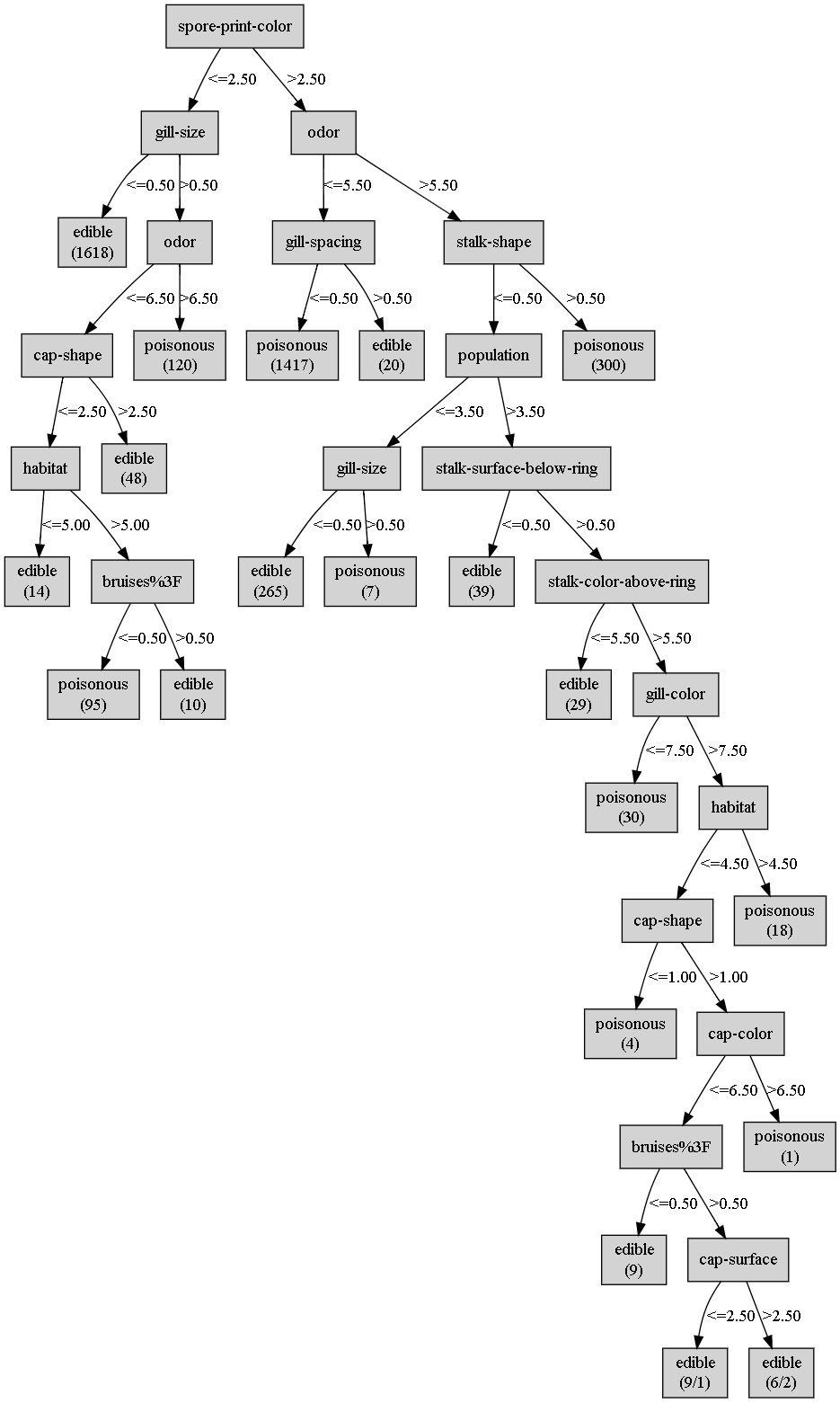
\*En ambos códigos utilicé 50% en training y 50% de test para calcular los datos

Primero pondré el código usando id3 junto con el resultado:

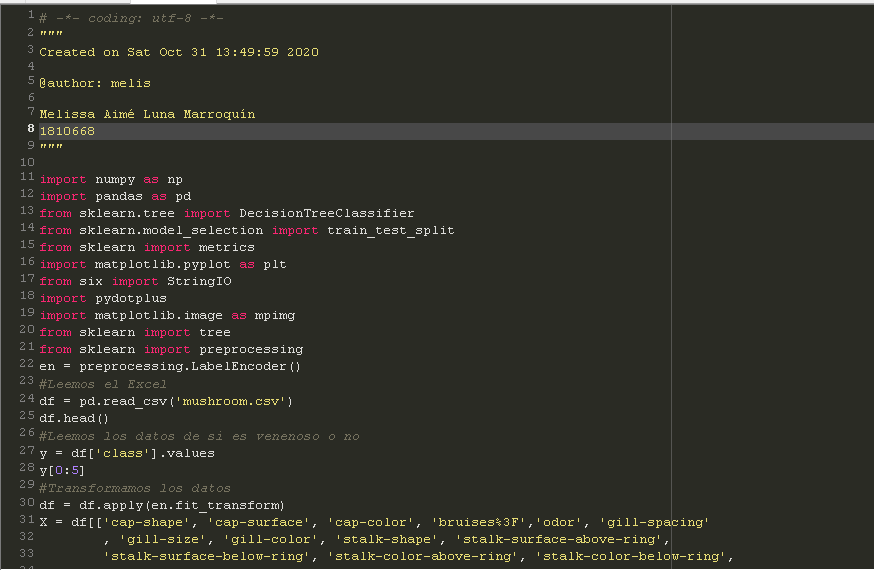


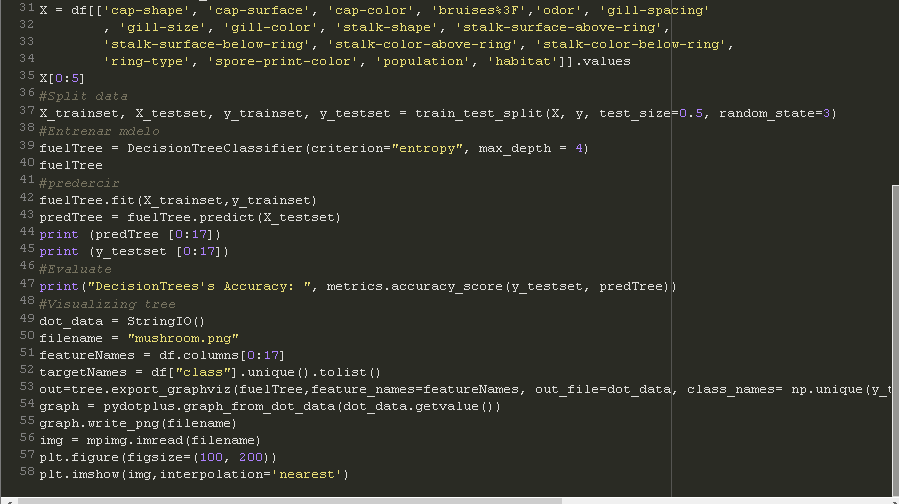


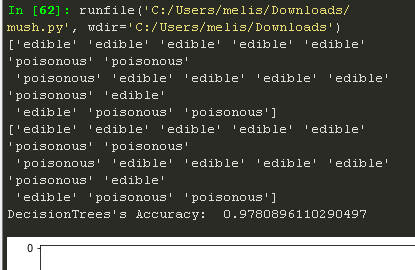


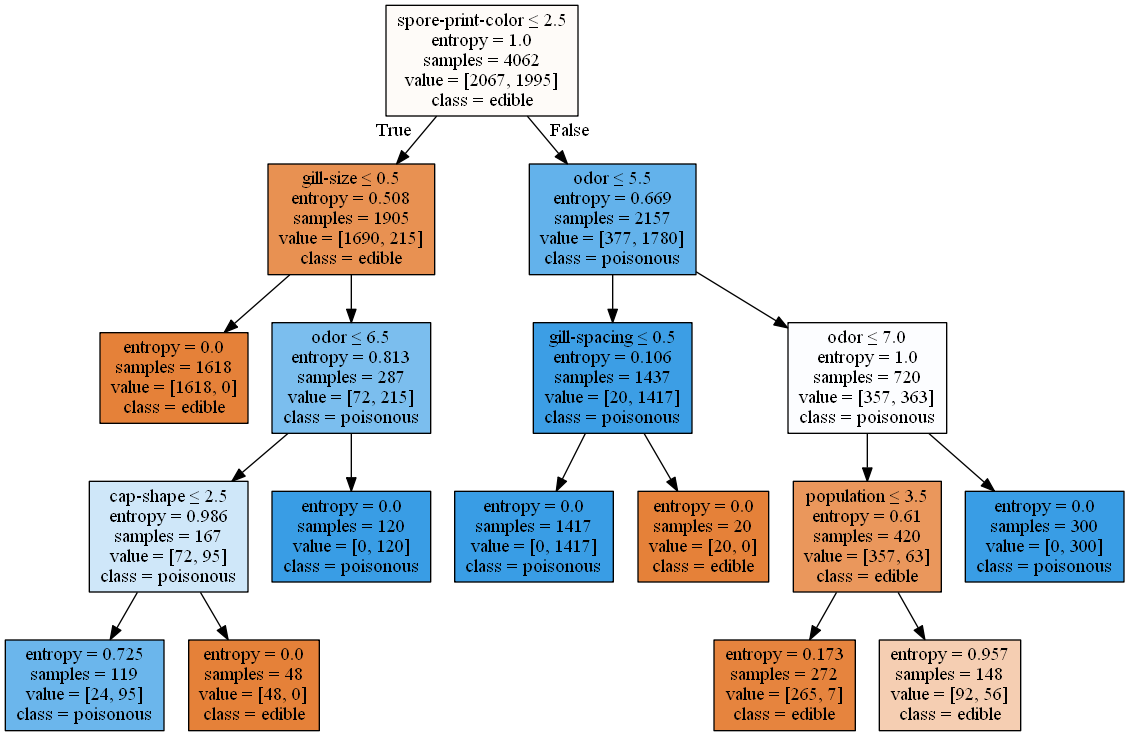


Y en las siguientes imágenes se mostrará el código utilizando Graphviz:



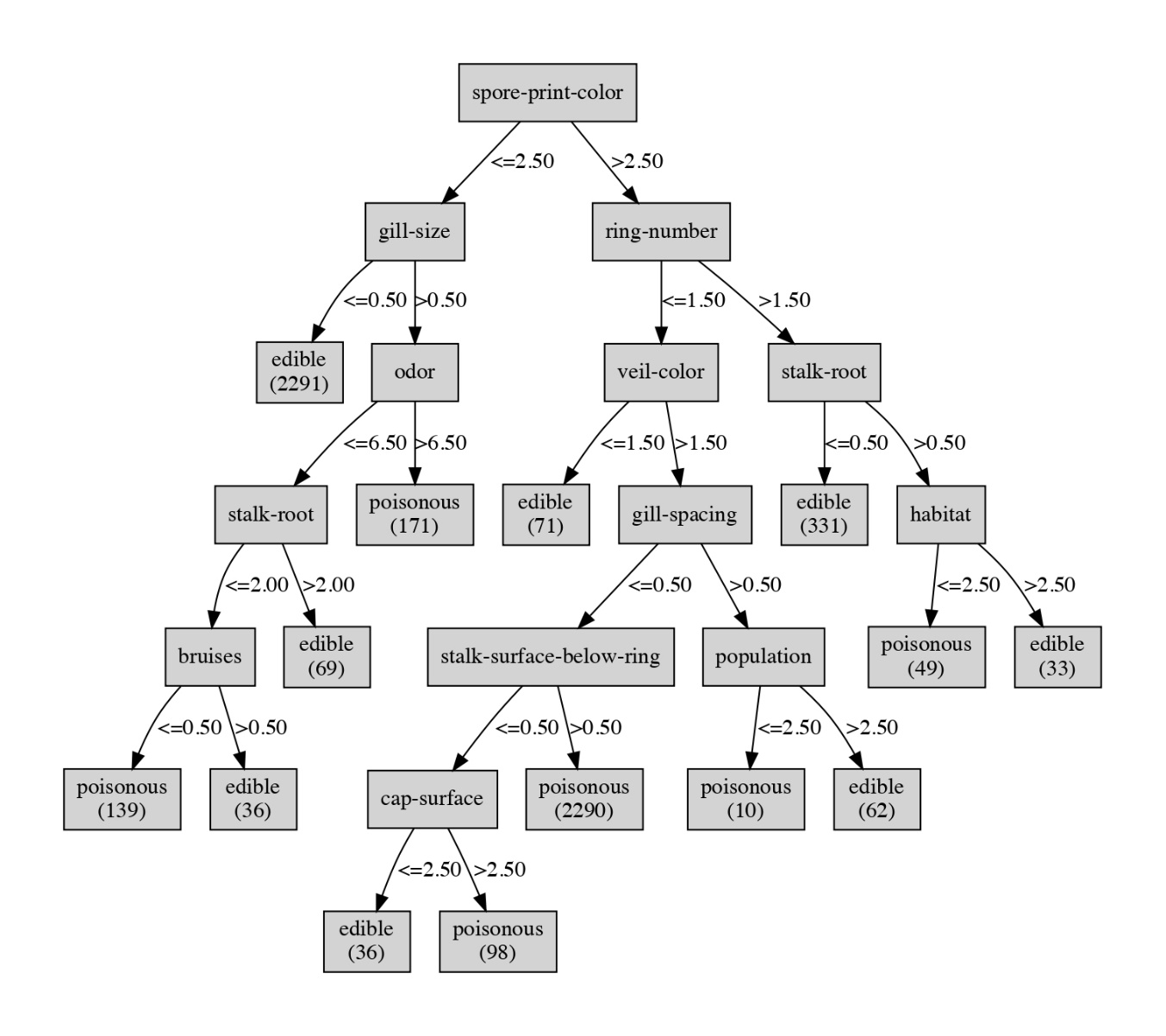






En mi opinión siento que pude eliminar o no haber eliminado columnas para poder mejorar el árbol de decisión porque pude ver que se alargó mucho y puede haber uno más compacto, en donde se llegue a la respuesta más rápido. Entonces, para poder llegar a un mejor resultado sería investigar o indagar más en sobre qué es importante para saber si un hongo es venenoso o no, e incluso podríamos haber usado más en el training y menos en el test; al igual que el max\_depth que en ambos utilicé 4 y eso también puede afectar al momento de crear el árbol de decisión.

Como pocos minutos extra, quise comprobar mi hipótesis sobre si mi árbol podría comprimirse si utilizábamos todos los datos y el resultado fue el siguiente:



Como conclusión final, se puede decir que el árbol sale muy distinto en algunas ramas y podría llegarse más rápido a si el hongo en venenoso o no, incluso con un accuracy de 1.0 en el primer código de python.