UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería
CC3074 – Minería de Datos
Sección 30
Ing. Leonel Guillén

Hoja de Trabajo 3

Alejandro Martinez - 21430
Samuel Argueta - 211024

GUATEMALA, 10 marzo de 2024

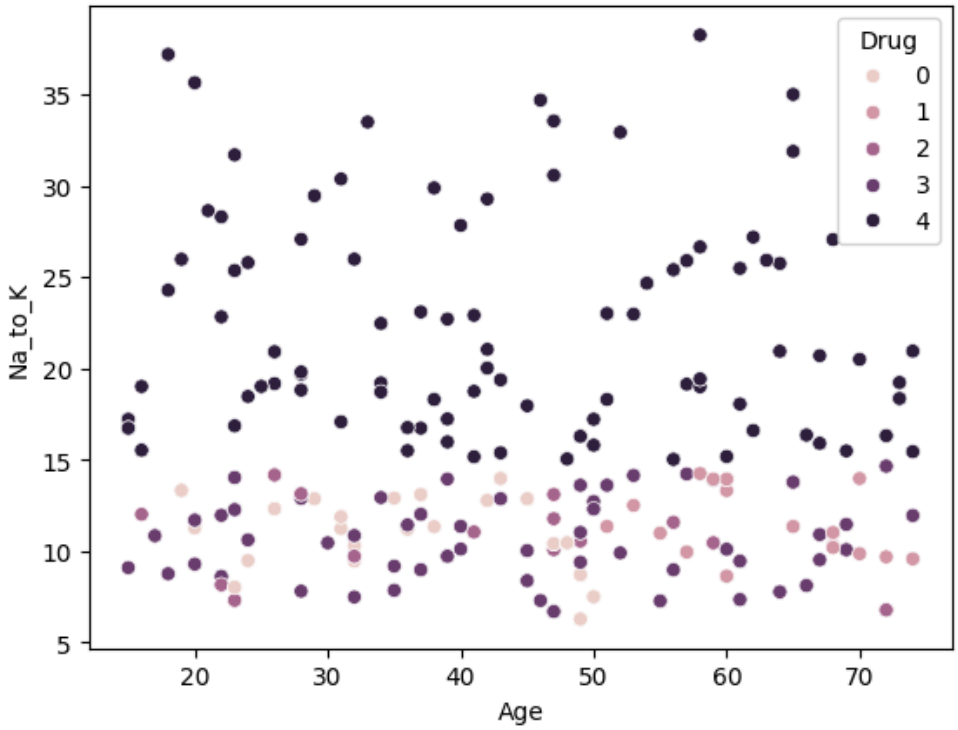
DATOS

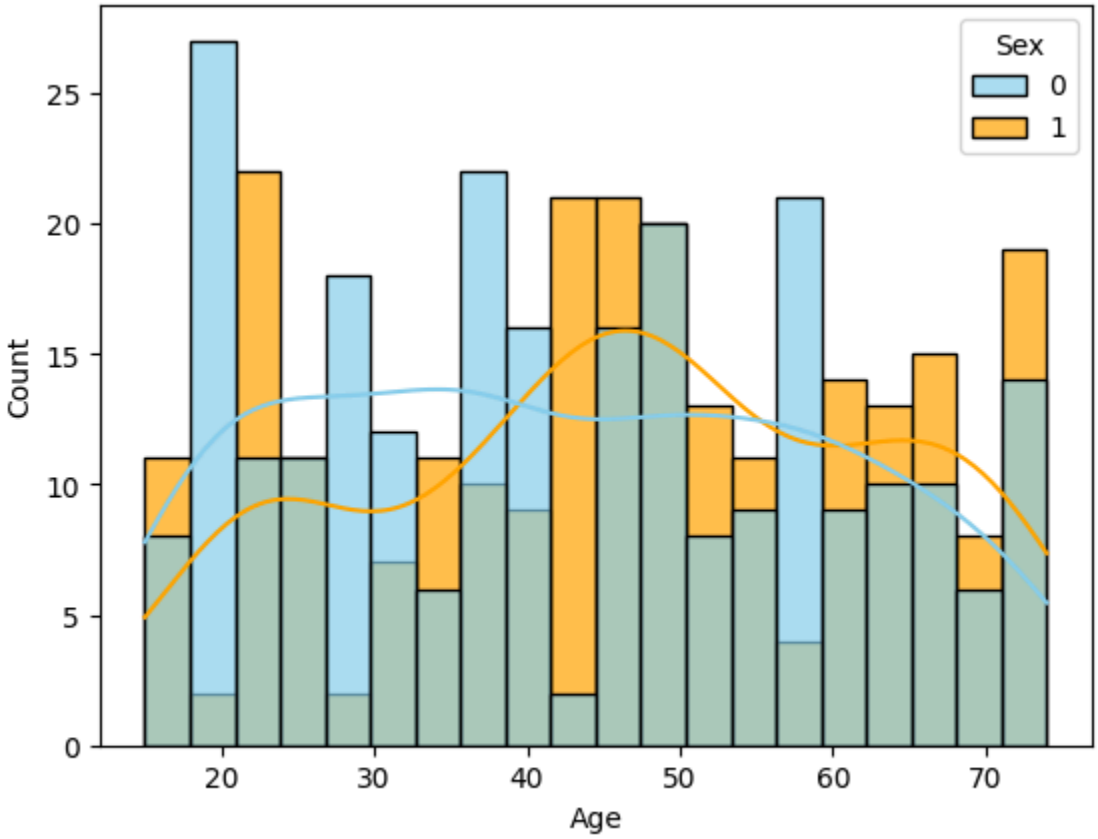
Para este conjunto de datos, se desea analizar el tipo de medicamento a recetar según las necesidades del paciente o bien, según su estado físico, pues ciertos medicamentos tiene afecciones severas según ciertas complicaciones de salud, por ello es que se receta un medicamento B en lugar de un C.

Sobre los datos, se decidió hacer una conversión numérica para aquellos que no eran números (palabras, caracteres, booleanos) para poder trabajar con todos los datos, se hizo la conversión de para género se uso la convención M=1, F=2. Para el colesterol HIGH=2, NORMAL=1, LOW=0. Para las medicinas, se hizo A=0, B=1, C=2, X=3, Y=4 y para el BP se hizo los mismo que para Colesterol

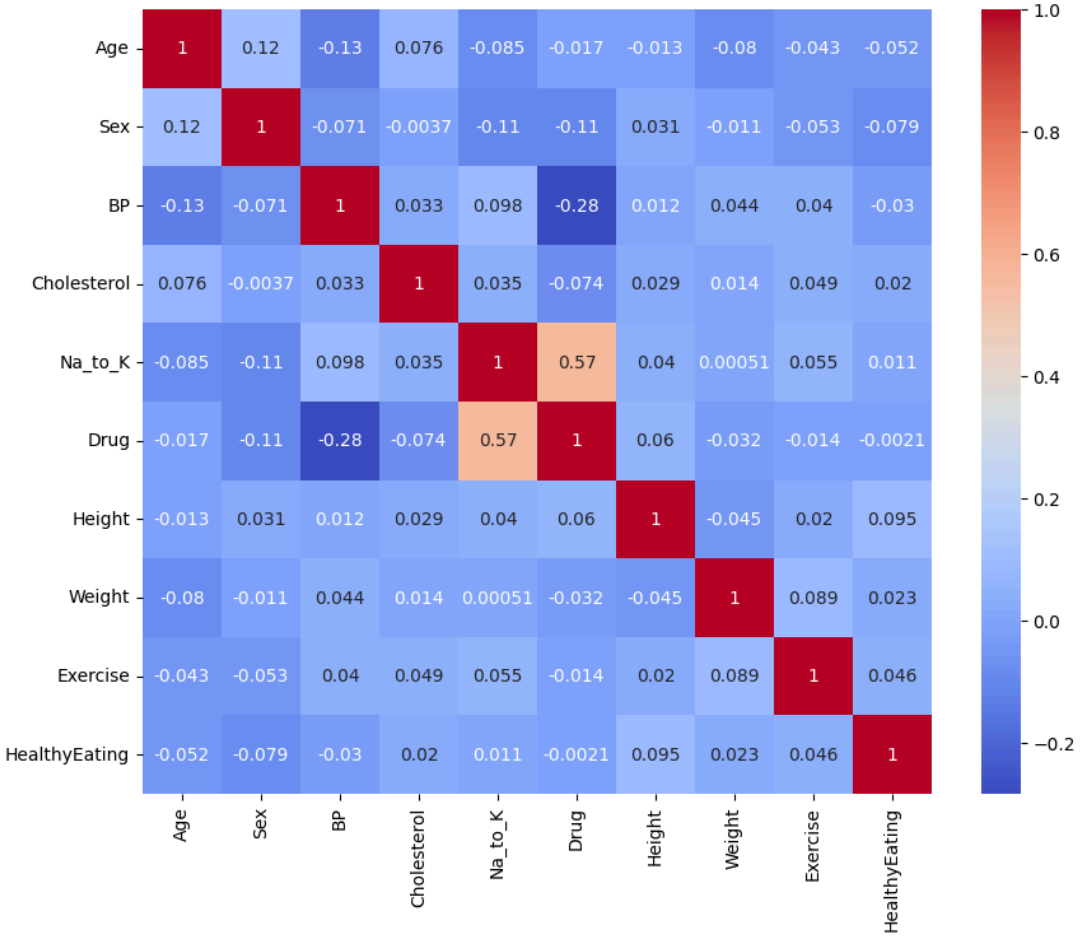
Relación entre la edad,el medicamento que se utiliza y su presión arterial. Validando si es que afecta en algo o se tienen ciertas consideraciones con los pacientes que poseen ciertos rasgos compartidos, como la edad y su presión arterial. El cómo es que ciertos medicamentos son mejor para ciertas edades con menor presión arterial, así como casos en donde la presión arterial juega un papel importante pues hay medicamentos más recetados a mayor presión arterial.

En este caso, se apreció que a mayor presión arterial, se receta el medicamento Y y X, aunque el X se receta en varios rangos, desde 15 hasta 35, en donde se tiene más consideraciones, es para los medicamentos A, B, C, donde estos se medican a los pacientes con presión arterial menor a 15, sin importar la edad


Drug
0
1
2
3
4
Na_to_K
Age
5
10
15
20
25
30
35
20
30
40
50
60
70


Sex
0
1
Count
Age
0
5
10
15
20
25
20
30
40
50
60
70

En este caso, la gráfica anterior se toma las endades por rangos de 10 en 10, en el eje vertical muestra la frecuencia o conteo de individuos en cada rango de edades. La altura de la barra indica cuántos individuos hay en ese rango de edades específico. Las curvas suaves sobre las barras son curvas de densidad de kernel, que proporcionan una representación suavizada de la distribución de edades para hombres y mujeres. Estas curvas ayudan a visualizar la forma general de la distribución.
Se observa cómo se distribuyen las edades en el conjunto de datos para hombres y mujeres. Las barras y las curvas suaves indican en qué rangos de edades hay concentración de individuos.
Donde las barras y las curvas se superponen, se puede inferir que hay similitudes en la distribución de edades entre hombres y mujeres. Donde hay separación, indica diferencias en la distribución de edades.
La transparencia ajustada con alpha puede ayudar a distinguir áreas de superposición y visualizar la contribución relativa de cada género en diferentes rangos de edades.



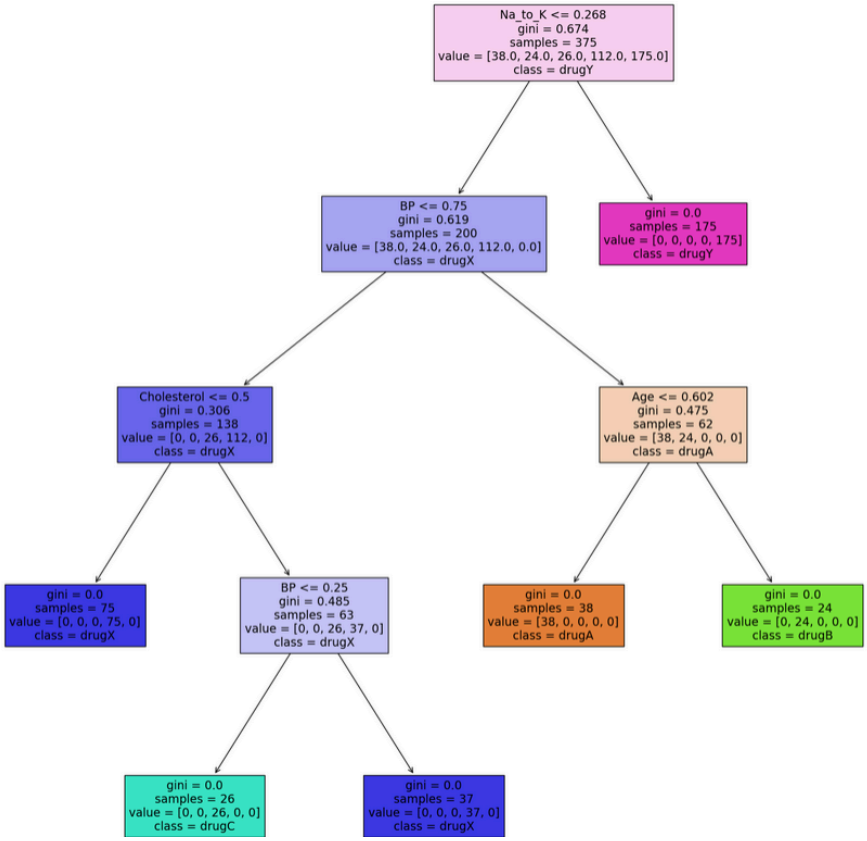
La correlación que se presenta es entre la presión arterial y la medicina, donde esté es proporcional, entre más presión arterial, se categoriza el tipo de medicamento que el paciente debe ingerir o se le debe recetar.

EXPLICACION ARBOL

Para el tema de los conjuntos de entrenamiento, para los de test y prueba, en este caso, para los X e Y se utilizó la columna Drug como variables objetivo, siendo el valor de la variable X, y el resto de valores para la variables y.
El tamaño predeterminado para el conjunto de prueba es del 25% de los datos. La proporción de clases en el conjunto de entrenamiento y prueba sea similar a la proporción original en el conjunto de datos completo. Dado que no se ha utilizado la estratificación, las clases se mantendrán en proporciones similares en ambos conjuntos. La estratificación asegura que la distribución de clases en el conjunto de entrenamiento y prueba sea similar a la distribución original en todo el conjunto de datos.

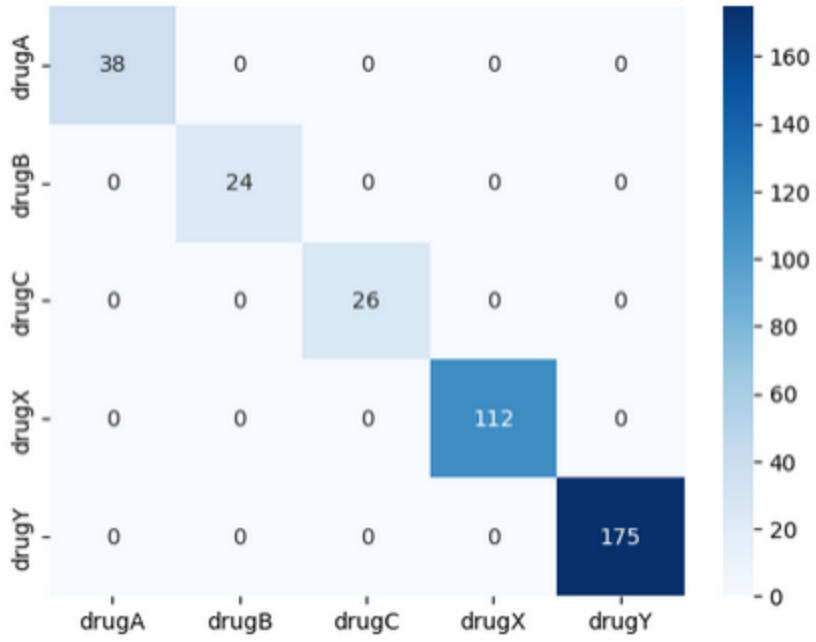
Pre Punning

Post punning

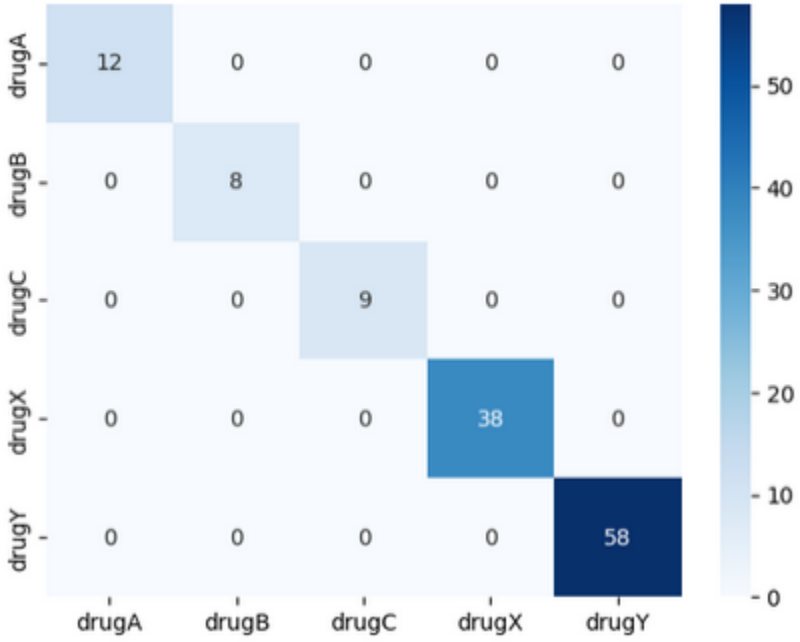


En esta caso particular, tanto para pre como post se obtuvo el mismo resultado, donde se aprecio como es que llega de la medicina Y, que esa la raíz, hasta la medicina X o C como rama más lejana. Claro, se valida el tema de la presion arterial, donde al hacer tal validacion se obtiene que se utiliza la medicina B o A

```
Train score  1.0
Test score  1.0
Train Matriz de Confusión
```



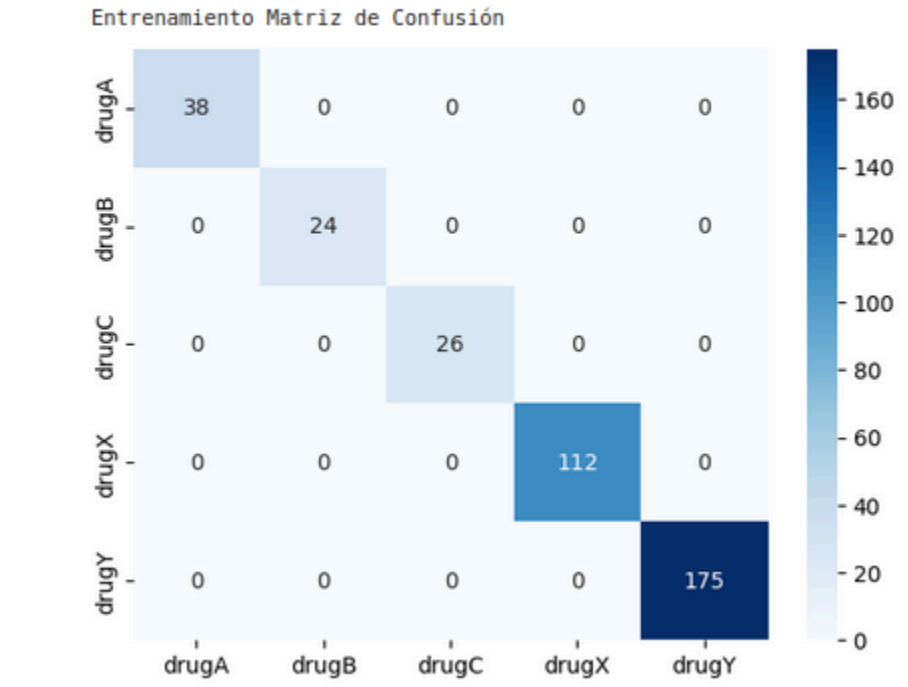

```
Test Matriz de Confusión
```

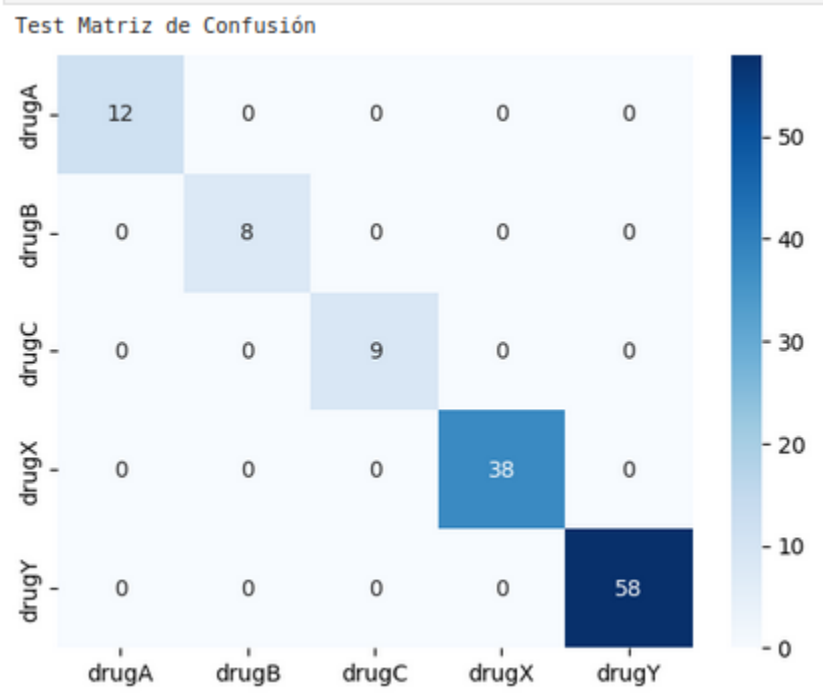


En los valores se aprecia que el punteo para el entrenamiento y el testeo es el mismo, lo cual indica que no hay sobre ajuste, aunque es muy bueno para ser cierto, dando indicios que puede estar mal planificado el tema de la clasificación de las variables y que haya dado valores predeterminado y no predictions siendo un hecho que puede resultar muy negativo para el analisis.

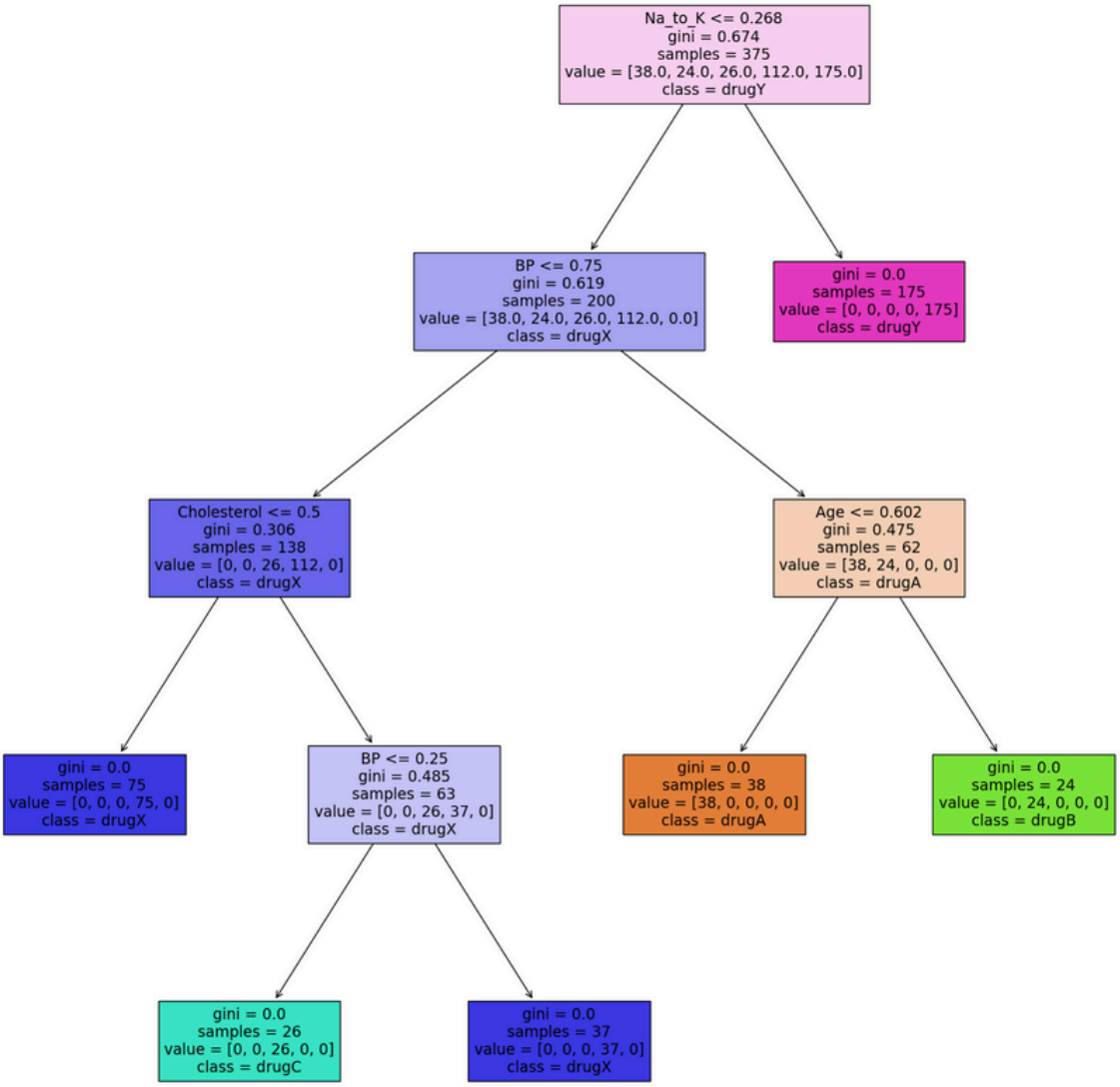
```
Accuracy en Train:  1.0
Accuracy en Test:  1.0
```

```
[28]: confusion_matrix_plot(y_train_pred, y_train_encoded, 'Entrenamiento')
```





Para estos resultados de pre poda, se ve que no hay diferencia alguna entre la pre poda y el árbol crudo. Este es el arbol crudo



Donde se aprecia que tiene una similitud entre los pre punning y post punning, no habiendo variación. Como tal, se sigue planteando la posibilidad de tener un error pues no se cree que es algo muy perfecto, 1/1 es muy bueno para ser verdad. Lo cual da ciertas alarmas como analistas de que puede haber algo mal realizado durante el análisis.