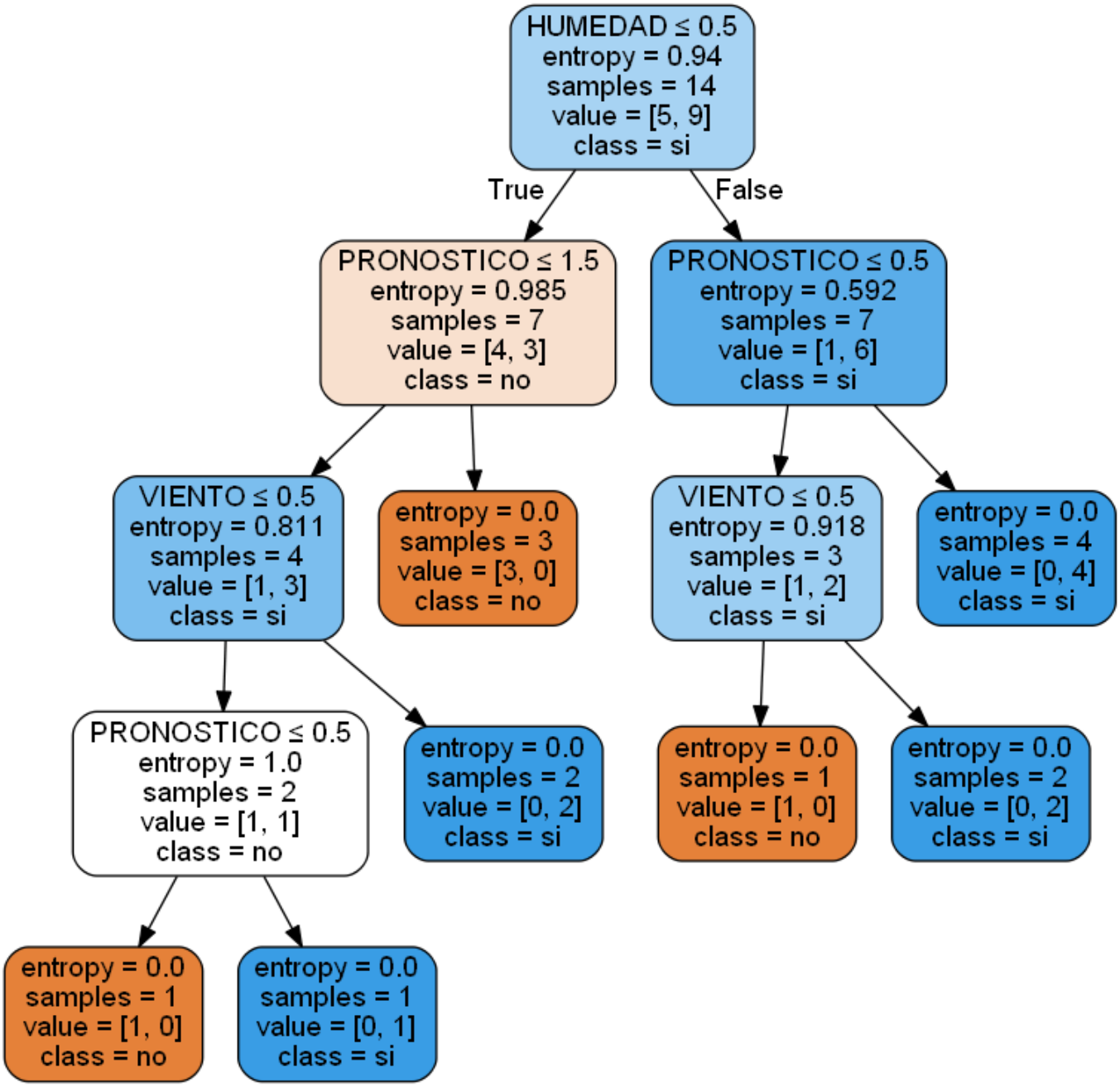
Tp04HerramientasParaElAnalisisDeInformacion

1. A partir del dataset presentado a continuación, y teniendo en cuenta las fórmulas de entropía y ganancia de información calcule y diagrame el árbol de decisión que le permita decidir si comer asado o no en función del clima:

Respuesta: el arbol de decision generado es el siguiente



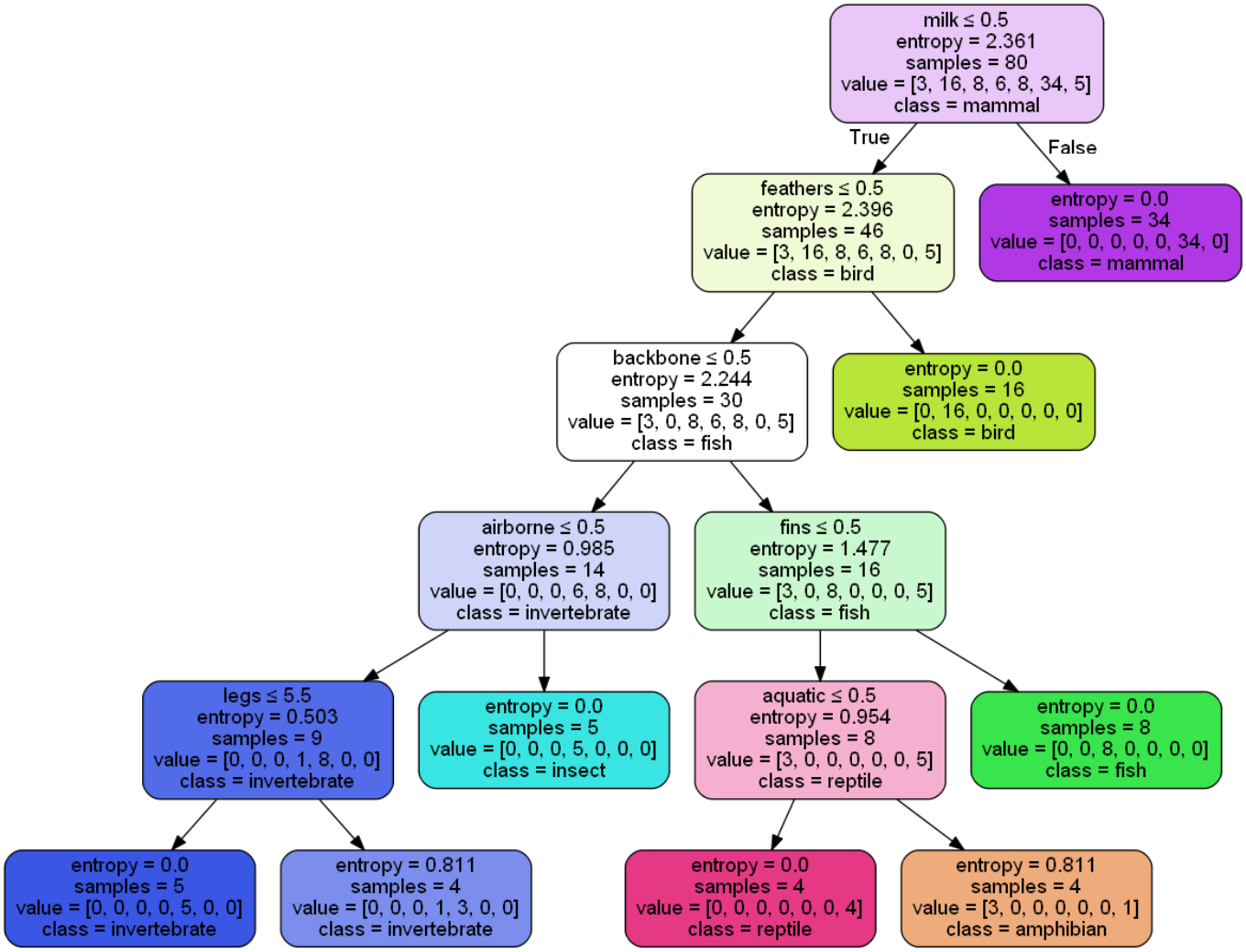
2. Trabaje con el dataset de Scikit Learn “wine”:

```
a. Utilice el metadata que provee la librería, ¿Cuál es el tema que aborda el
dataset?
```

```
Respuesta: El dataset usa el analisis quimico para determinar el origen de los
vinos
haciendo un import del dataset se puede observar su descripcion completa
![descripcion_dataset_wine](code\punto2\descripcion_wine.txt)

b. Genere el árbol de decisión que permita clasificar los diferentes tipos de vino
utilizando un muestreo con proporciones de 80% para entrenamiento y 20% para
testeo.

[arbol punto 2b](code\punto2\arbol_punto2bTP05.png)

c. Explore la solución dada y las posibles configuraciones para obtener un nuevo
árbol que clasifique “mejor”. Documente las conclusiones.
```

3. Ahora, analice el archivo zoo.csv2:

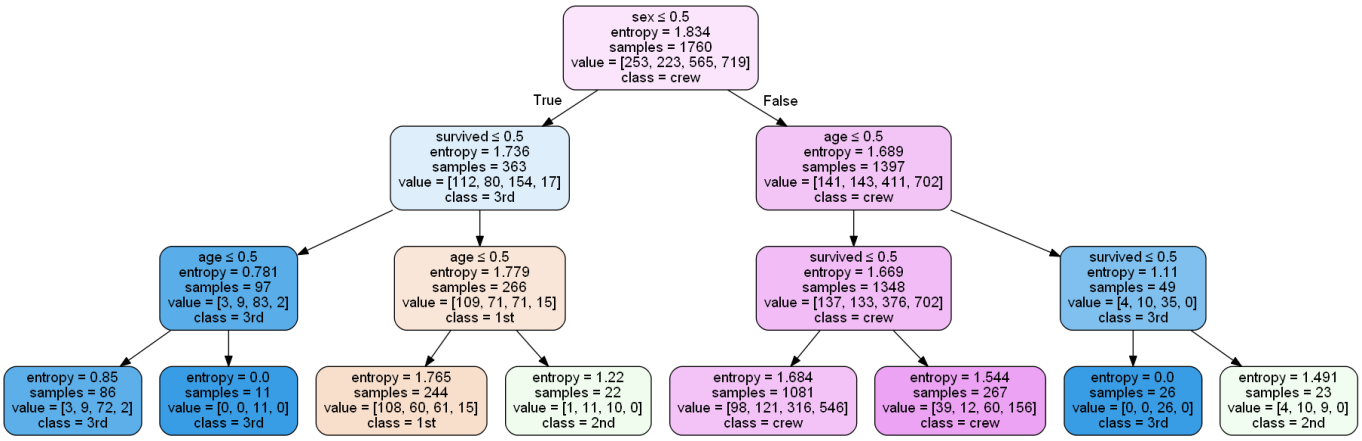
a. Genere el árbol de decisión que permita inferir el tipo de animal en función de sus características. Explique someramente que resultado se obtiene en términos del árbol y en términos de la eficiencia del mismo.



• ¿Varía ese resultado si se elimina el atributo “animal”? ¿Por qué? • Cuantos niveles posee el árbol generado? ¿Qué atributos debemos modificar si deseamos realizar una poda del mismo? Modifique esos atributos para que el árbol generado conste de 4 niveles. ¿Afecta la eficiencia de la clasificación esta modificación?

4. Se provee la base de datos de los pasajeros del famoso barco que se hundiera en su viaje inaugural (archivo titanic-en.csv) con los siguientes atributos y valores posibles:

- Class {"1st","2nd","3rd","crew"}
- Age {"adult","child"}
- Sex {"male","female"}
- Survived {"yes","no"} Genere el árbol de clasificación, explore la solución dada y las posibles alternativas para obtener un nuevo árbol que clasifique “mejor”.



5. Un Banco de Portugal realizó una campaña de marketing en busca de clientes de plazos fijos basada en llamados telefónicos. Se provee el dataset3 real (bank-full.csv) con más 45000 instancias y el detalle (bank-names.txt) de los datos registrados de cada una de las personas contactadas por la entidad bancaria.

a. Realice las tareas necesarias para poder procesar el dataset en Scikit-Learn.

Fue necesario cambiar los tipos de datos a string ´´´ data['balance'] = data['balance'].astype('string') data['pdays'] = data['pdays'].astype('string') ´´´ Con visual-studio reemplace los ";" por ","

y la fila de "target names" le tuve que sacar la comillas dobles

b. Luego, genere el árbol de decisión, y optimice los resultados, con el objetivo de explicar cuáles son las características más importantes que permiten identificar a una persona que accederá o no al plazo fijo. Documente los resultados.

el arbol resultante es el siguiente

6. Guarde los archivos resultantes de las actividades prácticas en una carpeta denominada tp0301- que a su vez tenga un directorio por cada uno de los puntos de este trabajo, comprima la carpeta y envíelo al equipo docente.

Medidas de evaluación para técnicas de clasificación:

En función de la clasificación realizada, complete las siguientes actividades: a. Accuracy. 1. Escoja un modelo y calcule el accuracy del mismo. 2. ¿Cómo se interpreta la métrica anterior? 3. ¿Qué aporta el accuracy? b.

Recall/Precision. 1. Ahora, sobre el mismo modelo de a), calcule las métricas recall y precisión para ambos modelos. 2. ¿Cuál es la diferencia entre ambas? 3. ¿Qué aspectos aborda cada una? c. Matiz de confusión: ¿En qué casos el modelo clasifica mal?