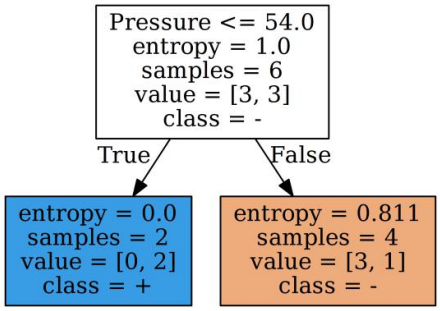
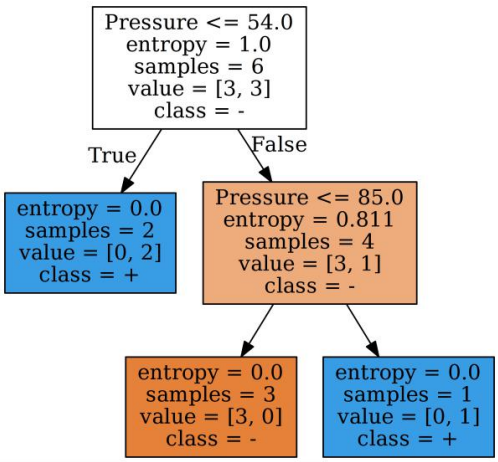
ENTREGA 3.
ÁRBOLES DE DECISIÓN.

Patricia Aguado Labrador

PARTE I:

Presión	40	48	60	72	80	90
Clase	+	+	-	-	-	+





PARTE II:

1. Porque el archivo de datos contiene atributos numéricos y el clasificador ID3 solo trabaja con atributos nominales.


Relation: Thoracic_Surgery_Data

No.	1: DGN	2: PRE4	3: PRE5	4: PRE6	5: PRE7	6: PRE8	7: PRE9	8: PRE10	9: PRE11	10: PRE14	11: PRE17	12: PRE19	13: PRE25	14: PRE30	15: PRE32	16: AGE	17: **Risk1Yr**
	Nominal	Numeric	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Nominal

2. Al observar el árbol vemos que hay dos atributos del dataset que no aparecen en él. Podríamos prescindir de los atributos PRE19 y PRE32.
Al eliminar los atributos que no resultan importantes en la clasificación obtenemos los mismos resultados, salvo el árbol que en este caso es más pequeño, se reduce el número de hojas en 11 y el tamaño del árbol en 12.

```
=== Summary ===

Correctly Classified Instances         120               75      %
Incorrectly Classified Instances        40               25      %
Kappa statistic                          0.0208
Mean absolute error                      0.2586
Root mean squared error                  0.4423
Relative absolute error                 99.5075 %
Root relative squared error            119.7396 %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,154    0,134    0,182      0,154    0,167      0,021    0,593     0,197     T
                 0,866    0,846    0,841      0,866    0,853      0,021    0,593     0,870     F
Weighted Avg.    0,750    0,730    0,734      0,750    0,741      0,021    0,593     0,760

=== Confusion Matrix ===

   a   b   <-- classified as
   4  22 |   a = T
  18 116 |   b = F
```

3. Tanto en la parte I como en el ejercicio 2 de la parte II, se realiza una discretización implícita basada en la entropía, es decir ninguna de ellas es más eficiente ya que es lo mismo pero con diferentes datos. Por otro lado, la técnica que puede resultar más eficiente es la de realizar la discretización en los datos directamente, ya que el clasificador puede producir demasiados intervalos al plantear las diferentes ramas del árbol.

4. Al aplicar el clasificador J48 sobre el archivo de datos que contiene atributos numéricos podemos observar que realiza la siguiente discretización implícita sobre 3 atributos:
- AGE: x<=62, x>62, x<=65, x>65, x<=70, x>70
- PRE4: x<=2.66, x>2.66, x<=2.88, x>2.88
- PRE5: x<=2.04, x>2.04

Para discretizar estos atributos y poder aplicar ID3 los convertiremos en nominales de la siguiente forma:
- AGE pasará a estar dividido en 4 posibles valores:
    62 o menos
    De 63 a 65
    De 66 a 70
    Más de 70
- PRE4 pasará a estar divido en 3 posibles valores:
    2.66 o menos
    De 2.67 a 2.88
    Más de 2.88
- PRE5 pasará a tener 2 posibles valores:
    2.04 o menos
    Más de 2.04

Al aplicar ID3 obtenemos los siguientes resultados:

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         124               77.5    %
Incorrectly Classified Instances        33               20.625  %
Kappa statistic                          0.1127
Mean absolute error                      0.225
Root mean squared error                  0.4615
Relative absolute error                 90.628  %
Root relative squared error            129.4144 %
UnClassified Instances                   3                1.875  %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,895    0,792    0,862      0,895    0,878      0,114    0,579     0,860     F
                 0,208    0,105    0,263      0,208    0,233      0,114    0,541     0,186     T
Weighted Avg.    0,790    0,687    0,771      0,790    0,780      0,114    0,573     0,757

=== Confusion Matrix ===

   a   b   <-- classified as
 119  14 |   a = F
  19   5 |   b = T
```

Al comparar estos resultados, clasificador ID3 sobre el dataset discretizado, con los ofrecidos por el clasificador J48 sin poda sobre el dataset realizando una discretización implícita, vemos que la tasa de acierto aumenta en un 2.5%. Podemos ver también que el número de instancias mal clasificadas es menor, aunque ID3 deja 3 instancias sin clasificar.