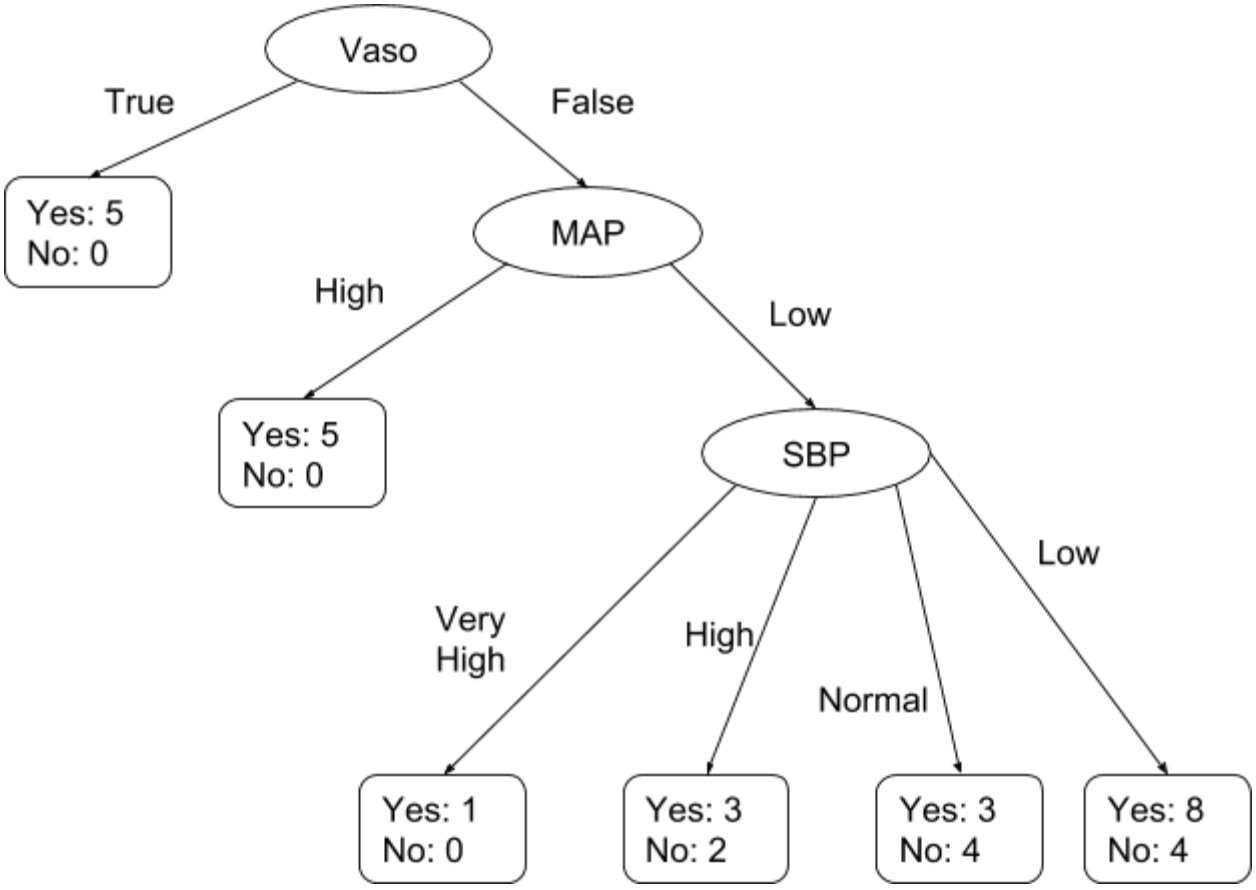
Khantil Choksi - khchoksi
Aman Chauhan - achauha3

Problem 3:




(a) Post-pruning based on optimistic errors

(i) Optimistic Errors:

| Before Splitting | After Splitting |
|---|---|
| $\frac{10}{25}$ | $\frac{9}{25}$ |

(ii) Based upon the optimistic erros, the subtree should not be pruned. As after splitting based on SBP, the optimistic error reduces. Therefore, we keep the given decision tree retained.

(iii) Classified provided testing dataset (“hw2q3_test.csv”):

| Vaso | MAP | SBP | Sepsis Shock | Predicted Class (Based on given retained decision tree) |
|---|---|---|---|---|
| FALSE | Low | High | Yes | Yes |
| FALSE | Low | High | Yes | Yes |
| FALSE | Low | High | Yes | Yes |
| FALSE | Low | High | No | Yes |
| TRUE | High | High | Yes | Yes |
| FALSE | Low | Low | Yes | Yes |
| FALSE | Low | Low | Yes | Yes |
| FALSE | Low | Low | Yes | Yes |
| FALSE | Low | Low | Yes | Yes |
| TRUE | Low | Low | Yes | Yes |
| FALSE | High | Low | Yes | Yes |
| FALSE | Low | Normal | Yes | No |
| FALSE | Low | Normal | Yes | No |
| FALSE | Low | Normal | No | No |
| FALSE | Low | Normal | No | No |
| FALSE | Low | Normal | No | No |
| FALSE | Low | Normal | No | No |
| TRUE | High | Normal | Yes | Yes |
| FALSE | Low | Very High | Yes | Yes |
| TRUE | Low | Very High | Yes | Yes |

Performance Matrix

| | | Predicted Class | |
|---|---|---|---|
| | | Class = Yes | Class = No |
| Actual Class | Class = Yes | 13 (TP) (a) | 2 (FN) (b) |
| | Class = No | 1 (FP) (c) | 4 (TN) (d) |

Accuracy = $\frac{a+d}{a+b+c+d} = \frac{13+4}{13+2+1+4} = \frac{17}{20} = 0.85$

Recall = $\frac{a}{a+b} = \frac{13}{13+2} = \frac{13}{15} = 0.867$

Precision = $\frac{a}{a+c} = \frac{13}{13+1} = \frac{13}{14} = 0.93$

Specificity (TNR) = $\frac{TN}{N} = \frac{TN}{TN+FP} = \frac{4}{4+1} = \frac{4}{5} = 0.8$

Sensitivity (TPR) = $\frac{TP}{P} = \frac{TP}{TP+FN} = \frac{13}{13+2} = \frac{13}{15} = 0.867$

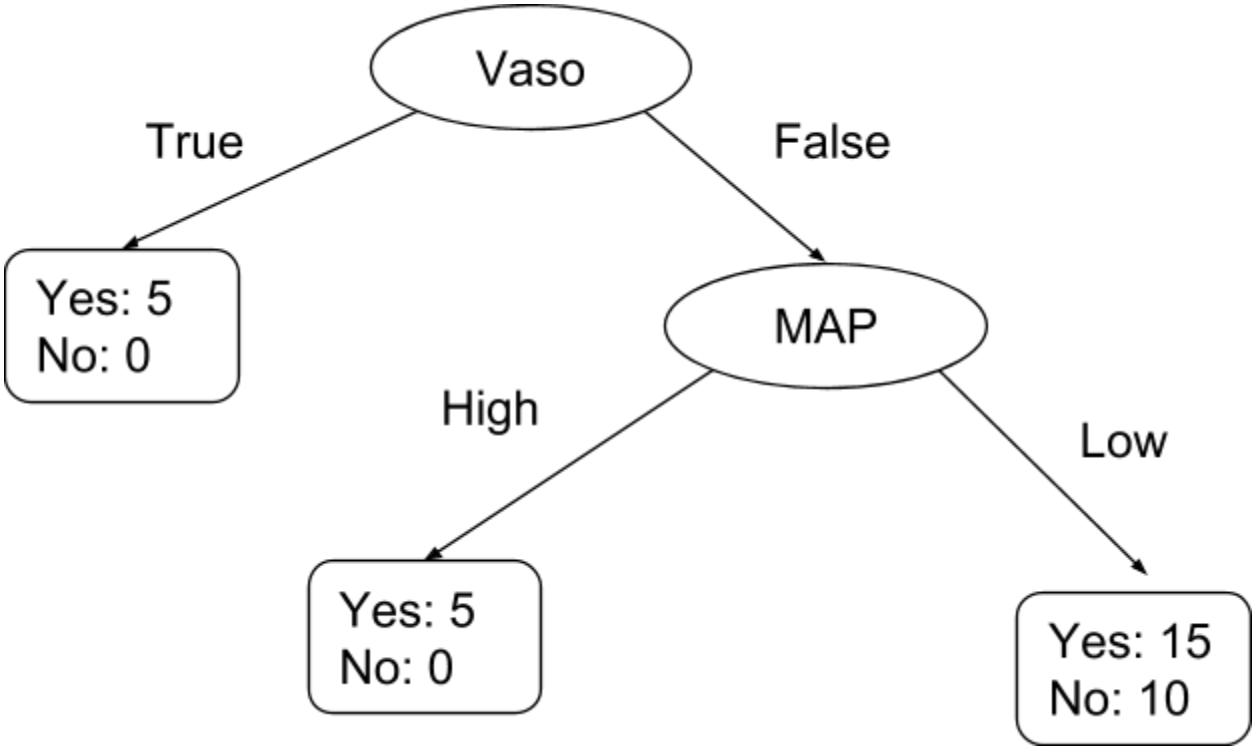
F1 Measure = $2 * \frac{precision * recall}{precision + recall} = 2 * \frac{0.93 * 0.867}{0.93 + 0.867} = 0.8965$

(b) Post-pruning based on pessimistic errors

(i) Pessimistic Errors:

| Before Splitting | After Splitting |
|---|---|
| $\frac{(10 + 1 * 0.5)}{25} = \frac{10.5}{25}$ | $\frac{(9 + 4 * 0.5)}{25} = \frac{11}{25}$ |

(ii) Based upon the pessimistic erros, the subtree should be pruned. As after splitting based on SBP, the pessimistic error increases. Therefore, we prune the given decision tree from SBP. Pruned Decision Tree is as shown below:




(iii) Classified provided testing dataset ("hw2q3_test.csv"):

| Vaso | MAP | SBP | Sepsis Shock | Predicted Class (Has Septic Shock) |
|---|---|---|---|---|
| FALSE | Low | High | Yes | Yes |
| FALSE | Low | High | Yes | Yes |
| FALSE | Low | High | Yes | Yes |
| FALSE | Low | High | No | Yes |
| TRUE | High | High | Yes | Yes |
| FALSE | Low | Low | Yes | Yes |
| FALSE | Low | Low | Yes | Yes |
| FALSE | Low | Low | Yes | Yes |
| FALSE | Low | Low | Yes | Yes |
| TRUE | Low | Low | Yes | Yes |

| | | | | |
|---|---|---|---|---|
| FALSE | High | Low | Yes | Yes |
| FALSE | Low | Normal | Yes | Yes |
| FALSE | Low | Normal | Yes | Yes |
| FALSE | Low | Normal | No | Yes |
| FALSE | Low | Normal | No | Yes |
| FALSE | Low | Normal | No | Yes |
| FALSE | Low | Normal | No | Yes |
| TRUE | High | Normal | Yes | Yes |
| FALSE | Low | Very High | Yes | Yes |
| TRUE | Low | Very High | Yes | Yes |

Performance Matrix

| | | Predicted Class | |
|---|---|---|---|
| | | Class = Yes | Class = No |
| Actual Class | Class = Yes | 15 (TP) (a) | 0 (FN) (b) |
| | Class = No | 5 (FP) (c) | 0 (TN) (d) |

Accuracy = $\frac{a+d}{a+b+c+d}$ = $\frac{15+0}{15+0+5+0}$ = $\frac{15}{20}$ = 0.75

Recall = $\frac{a}{a+b}$ = $\frac{15}{15+0}$ = $\frac{15}{15}$ = 1

Precision = $\frac{a}{a+c}$ = $\frac{15}{20}$ = $\frac{15}{20}$ = 0.75

Specificity (TNR) = $\frac{TN}{N}$ = $\frac{TN}{TN+FP}$ = $\frac{0}{0+5}$ = $\frac{0}{5}$ = 0

Sensitivity (TPR) = $\frac{TP}{P}$ = $\frac{TP}{TP+FN}$ = $\frac{15}{15+0}$ = $\frac{15}{15}$ = 1

F1 Measure = 2 * $\frac{precision * recall}{precision + recall}$ = 2 * $\frac{0.75 * 1}{0.75 + 1}$ = 0.857

**(c) Performance Comparison**

| | Before Pruning | After Pruning |
|---|---|---|
| Accuracy | 0.85 | 0.75 |
| Recall | 0.867 | 1 |
| Precision | 0.93 | 0.75 |

Here, as the problem mentioned that the sepsis deaths could be prevented with early diagnosis and intervention. So, it is very important to measure TP and TNs are very costly for this example. So, we more care about recall. So, after pruning we have higher recall. So problem 3 (b) (ii) decision tree would be better model for the task of septic shock prediction.