Question 1)

The bias-variance tradeoff deals with the tradeoff between two sources of error that affects the performance of models. Models with high bias make strong assumptions about the data and are too simplistic. This is likely because the hypothesis space made available by the classification method is insufficient. Models with high bias are likely to underfit the data and are unlikely to capture the underlying patterns. They are likely to perform poorly on both training and testing data.

Variance refers to error that is a result of a model's sensitivity to specific data it was trained on. This is likely because the hypothesis space is too large for the training data and the hypothesis may not be accurate on new data. Models with high variance are likely to overfit, which means they can fit the training data very closely, but may not generalize well to new data. High variance may be a result of the use of a very complex model that fits the noise in the data.

The tradeoff is that as you decrease bias, you often increase variance and as you decrease variance, you often increase bias. The goal is to strike a balance between bias and variance. To reduce bias, you can increase the hypothesis space such as by increasing the model complexity. By doing so the models can capture more intricate patterns in the data, helping to reduce underfitting. To reduce variance you can employ resampling techniques like bagging. Combining many individual models reduces the overall variance and improves generalization. By averaging the predictions of many trees, the variance decreases without significantly increasing the bias. As the number of trees increases, the variance decreases. Ultimately there are generally no analytical techniques to find the optimal bias-variance trade-off. It requires a trial and error process. To find the right balance you need to experiment with different model complexities and measure error.

Question 2)

If we're assuming that class 1 is the positive class then:

Precision for class 1 = $\frac{TP}{TP+FP} = \frac{50}{50+40} = \frac{50}{90}$ = 0.555

Recall for class 1 = $\frac{TP}{TP+FN} = \frac{50}{50+30}$ = $\frac{50}{80}$ = 0.625

F1 score for class 1 = 2 * ($\frac{5/9 * 5/8}{5/9+5/8}$) = 0.588

If we're assuming that class 2 is the positive class then:

Precision for class 2 = $\frac{TP}{TP+FP}$ = $\frac{60}{60+30} = \frac{60}{90} = 0.6667$

Recall for class 2 = $\frac{TP}{TP + FN}$ = $\frac{60}{60 + 40}$ = $\frac{60}{100}$ = 0.6

F1 score for class 2 = 2 * ($\frac{6/9 \ * 6/10}{6/9 + 6/10}$) = 0.632

Question 3

Total Entropy $= -\left(\frac{6}{10}\log\left(\frac{6}{10}\right) + \frac{4}{10}\log\left(\frac{4}{10}\right)\right)$

$= 0.971$

Calculate IG for Outlook

$= 0.971 - \frac{4}{10}\left(-\left(\frac{3}{4}\log\left(\frac{3}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right)\right)\right)$

$-\frac{2}{10}.0 \quad -\frac{4}{10}\left(-\left(\frac{3}{4}\log\left(\frac{3}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right)\right)\right)$

$= 0.322$

IG for Temperature

$= 0.971 - \frac{3}{10}\left(-\left(\frac{2}{3}\log\left(\frac{2}{3}\right) + \frac{1}{3}\log\left(\frac{1}{3}\right)\right)\right)$

$-\frac{3}{10}\left(-\left(\frac{2}{3}\log\left(\frac{2}{3}\right) + \frac{1}{3}\log\left(\frac{1}{3}\right)\right)\right)$

$-\frac{4}{10}\left(-\frac{3}{4}\log\left(\frac{3}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right)\right)$

$= 0.096$

IG for Humidity

$$0.971 - \frac{5}{10}\left(-\left(\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}\right)\right)$$

$$- \frac{5}{10}\left(-\left(\frac{4}{5}\log\frac{4}{5} + \frac{1}{5}\log\frac{1}{5}\right)\right)$$

$$= 0.125$$

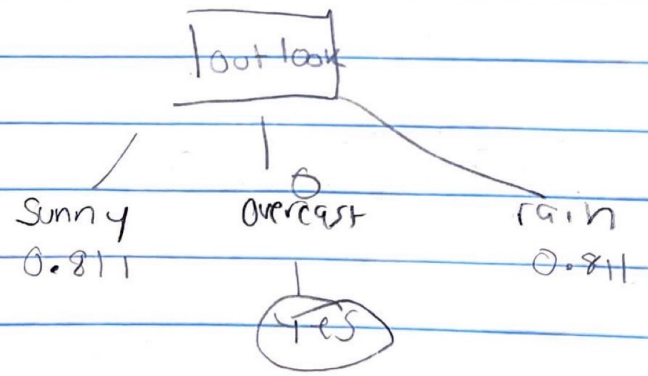
IC for Wind

$$0.971 - \frac{7}{10}\left(-\left(\frac{5}{7}\log\frac{5}{7} + \frac{2}{7}\log\frac{2}{7}\right)\right)$$

$$- \frac{3}{10}\left(-\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right)\right)$$

$$= 0.091$$

We get highest info gain from outlook so outlook will be the root node. This is the resulting tree:



. If it's, overcast PlayTennis is always a yes so it's directly a yes

IG for Sunny

IG temperature

$$0.811 - \frac{2}{4}\left(-\left(\frac{2}{2}\log\frac{2}{2} + \frac{0}{2}\log\frac{0}{2}\right)\right)$$

$$- \frac{1}{4}\left(-\left(\frac{1}{1}\log\frac{1}{1} + \frac{0}{1}\log\frac{0}{1}\right)\right)$$

$$- \frac{1}{4}\left(-\left(\frac{1}{1}\log\frac{1}{1} + \frac{0}{1}\log\frac{0}{1}\right)\right)$$

$$= 0.811 - 0 = 0.811$$

IG Humidity

$$0.811 - \frac{3}{4}\left(-\left(\frac{3}{3}\log\frac{3}{3} + \frac{0}{3}\log\frac{0}{3}\right)\right) -$$

$$\frac{1}{4}\left(-\left(\frac{1}{1}\log\frac{1}{1} + \frac{0}{1}\log\frac{0}{1}\right)\right)$$

$$= 0.811 - 0 = 0.811$$

IG Wind

$$0.811 - \frac{3}{4}\left(-\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right)\right)$$

$$- \frac{1}{4}\left(-\left(\frac{1}{1}\log\frac{1}{1} + \frac{0}{1}\log\frac{0}{1}\right)\right) = 0.023$$

There is a tie in IG between temperature and humidity, so I'm going to split arbitrarily and am choosing humidity. I'm adding humidity under sunny. The final tree will be shown on the last page for this problem

Rainy

IG for temperature

$$0.811 - \frac{2}{4}\left(-\left(\frac{2}{2}\log\frac{2}{2} + \frac{0}{2}\log\frac{0}{2}\right)\right)$$

$$- \frac{2}{4}\left(-\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)\right)$$

$$= 0.811 - 0 - 0.75 = 0.061$$

IG for Humidity

$$0.811 - \frac{1}{4}\left(-\left(\log\frac{1}{1}\,\frac{1}{1} + \frac{0}{1}\log\frac{0}{1}\right)\right)$$

$$- \frac{3}{4}\left(-\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right)\right)$$

$$= 0.025$$

IG Wind

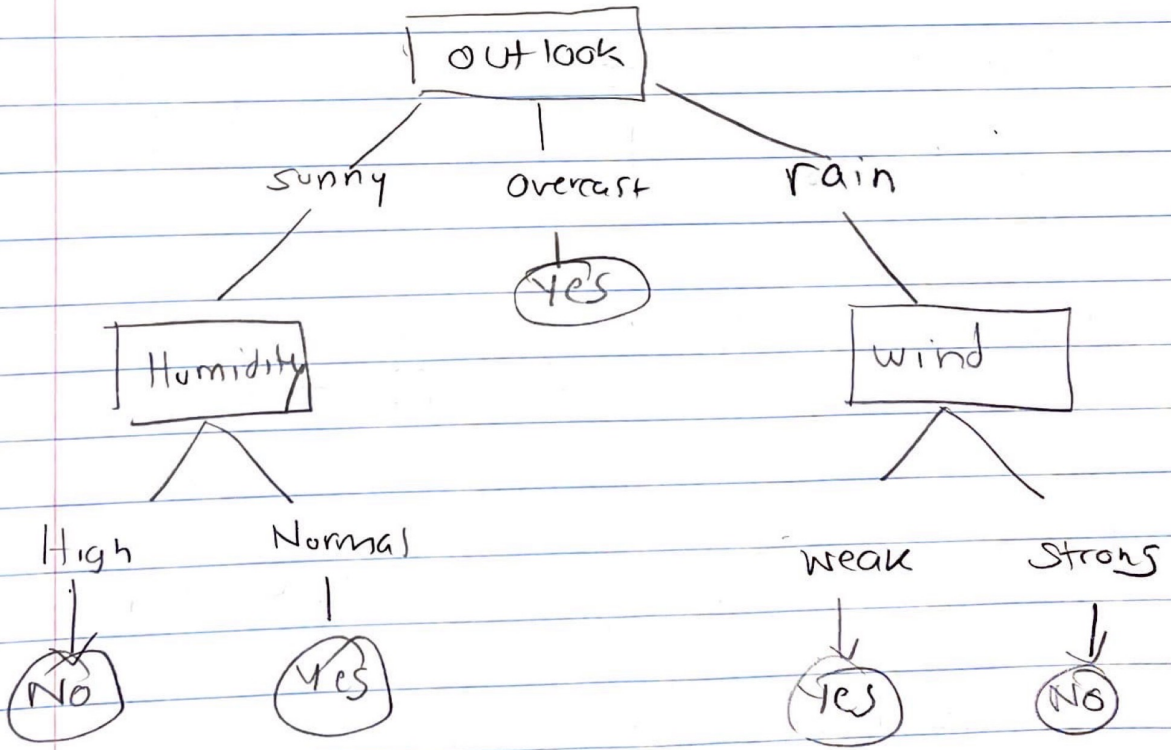
$$0.811 - \frac{3}{4}\left(-\left(\frac{3}{3}\log\frac{3}{3} + \frac{0}{3}\log\frac{0}{3}\right)\right)$$

$$- \frac{1}{4}\left(-\left(\frac{1}{1}\log\frac{1}{1} + \frac{0}{1}\log\frac{0}{1}\right)\right)$$

$$= 0.811$$

Wind has the highest information gain, so I'm adding wind under rain.

Final Decision tree



There are 3 Nos and 0 yes when humidity is high and 1 yes and 0 nos when humidy is normal so the decision is No when humidy is high and yes when humidity is normal.

There are 3 yes and 0 no when it's rainy and the wind is weak and 1 no and 0 yes when it's rainy and the wind is strong, so the decision is a yes when it's rainy and the wind is weak and no when it's rainy and the wind is strong.

Question 4

	Class 1	Class 2
Classifier 1 (class 1)	40/70	30/70
Classifier 2 (class 1)	20/40	20/40
Classifier 3 (class 2)	0/10	10/10

class 1 score: $\frac{40}{70} \cdot \frac{20}{40} \cdot \frac{0}{10} = 0$

class 2 score: $\frac{30}{70} \cdot \frac{20}{40} \cdot \frac{10}{10} = 0.214$

Class 2 has a higher score, hence a higher probability and so class 2 is the final decision.