ISMET OKAN CELIK CWID:10472265

Question 1: [4 points] Explain what is the bias-variance trade-off? Describe a few techniques to reduce bias and variance respectively

In machine learning, we desire to have low bias and low variance situations. If the machine learning model is too complex, you will have a low bias – high variance, if the model is too simple you will have a high bias- low variance. These two situations are not desirable. We need to optimize that for avoiding overfitting, underfitting and to make the error rate small. When we try to minimize one of them it increases the other. We need to find a balance between bias and variance. It is called a bias-variance trade-off. Model complexity should be optimum.

Reducing Bias:

Increasing the model complexity (For example, if we think about decision three; instead of using 2 nodes, we can use 100 nodes when we use the decision tree).

Reducing Variance:

-Resampling (e.g , Random Forest)

-Using multiple models in training

-Increasing the training set

Question 2: [6 points] Assume the following confusion matrix of a classifier. Please compute its
1) precision,
2) recall, and
3) F1-score.

	Class 1	Class 2
Class 1	50 (TP)	30 (FP)
Class 2	40 (FN)	60 (TN)

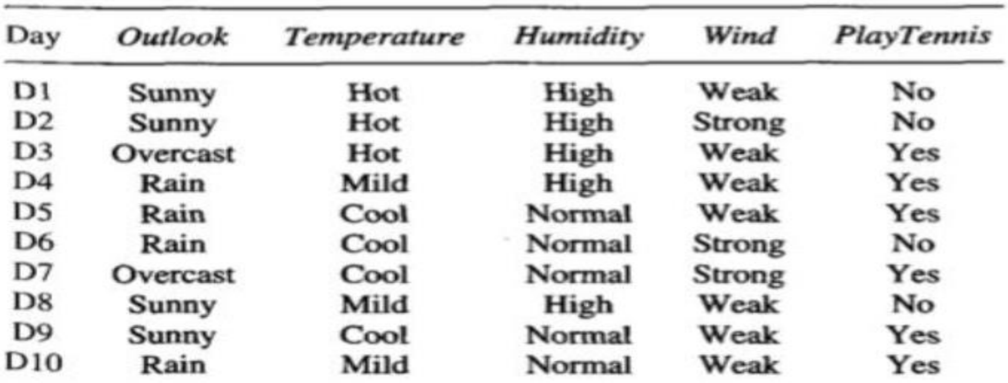
TP: True Positive
FP: False Positive
FN: False Negative
TN: True Negative

$$1)\ \text{Precision} = \frac{TP}{TP+FP} = \frac{50}{50+30} = \underline{\underline{0{,}625}}$$

$$2)\ \text{Recall} = \frac{TP}{TP+FN} = \frac{50}{50+40} = \underline{\underline{0.556}}$$

$$3)\ \text{F1-Score} = 2.\ \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2.\ \frac{0.625 \times 0.556}{0.625+0.556} = \underline{\underline{0.588}}$$

Question 3: [10 points] Build a decision tree using the following training instances (using information gain approach):



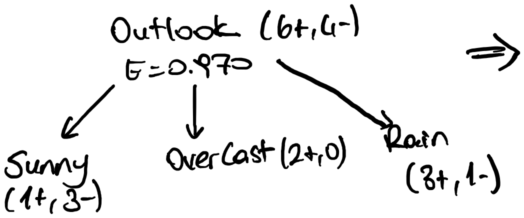
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes

Outlook $(6+, 4-) = \frac{-6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right) = 0.970$

Temperature $(6+, 4-) = 0.970$

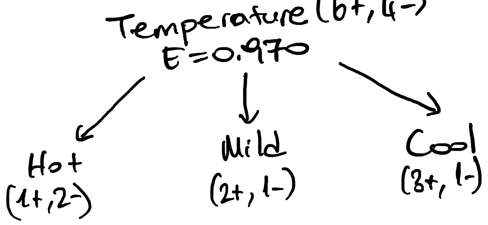
Humidity $(6+, 4-) = 0.970$

Wind $(6+, 4-) = 0.970$



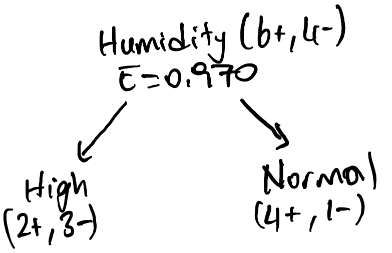
$\Rightarrow$ $\text{Gain}(S, \text{Outlook}) = 0.970 - 2\left(\frac{4}{10}\left(-\frac{3}{4}\cdot\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{3}{4}\right)\right)\right)$

$\text{Gain}(S, \text{Outlook}) = \underline{\underline{0,321}}$



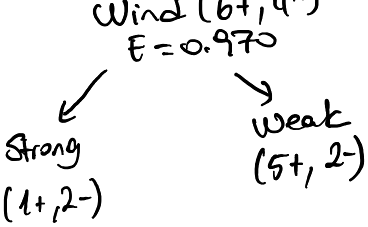
$\Rightarrow$ $\text{Gain}(S, \text{Temperature}) = 0.970 - \left(2\cdot\frac{3}{10}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) + \frac{4}{10}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)\right)$

$\text{Gain}(S.\text{Temperature}) = 0.970 - 0.551 - 0.325 = \underline{\underline{0.094}}$



$\Rightarrow$ $\text{Gain}(S, \text{Humidify}) = 0.970 - \left(\frac{5}{10}\left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right)\right) + \frac{5}{10}\left(-\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log\left(\frac{1}{5}\right)\right)\right)$

$\text{Gain}(S, \text{Humidity}) = \underline{\underline{0,124}}$

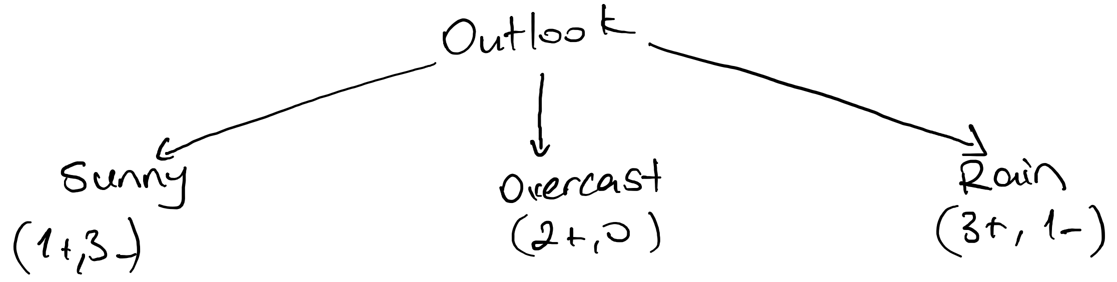


$= \text{Gain}(S, \text{Wind}) = 0.970 - \left(\frac{3}{10}\cdot\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{7}{10}\left(-\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7}\right)\right)$

$= 0.970 - 0.275 - 0.604$

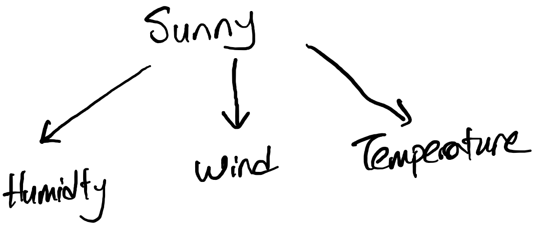
$\text{Gain}(S, \text{Wind}) = \underline{\underline{0.091}}$

Information Gain for Outlook is biggest among other values, because of that Outlook is chosen as Root Node.



$$E_{sunny} = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811$$

$E_{rain} = 0.811$



$$\text{Gain}(\text{Sunny}, \text{Humidity}) = 0.811 - \left(\frac{3}{4}(0) + \frac{1}{4}(0)\right)$$

$$= \underline{\underline{0.811}}$$

$$\text{Gain}(\text{Sunny}, \text{wind}) = 0.811 - \left(\frac{3}{4}\cdot\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{1}{4}(0)\right)$$

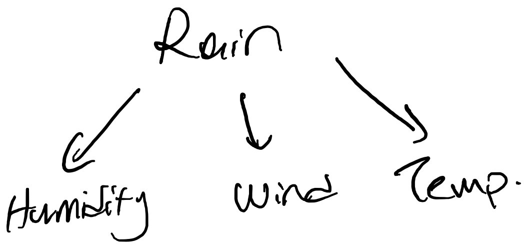
$$= 0,811 - 0.688 = \underline{\underline{0.122}}$$

$$\text{Gain}(\text{Sunny}, \text{Temperature}) = 0.811 - \left[\frac{2}{4}\cdot(0) + \frac{1}{4}(0) + \frac{1}{4}(0)\right)$$

$$= \underline{\underline{0.811}}$$

Gain(Sunny, Humidity) = Gain(Sunny, Temperature)

We can pick either Humidity or Temperature both gives the same value



$$\text{Gain}(\text{Rain}, \text{Humidity}) = 0.811 - \left(\frac{1}{4}(0) + \frac{3}{4}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)\right)$$

$$= 0.122$$

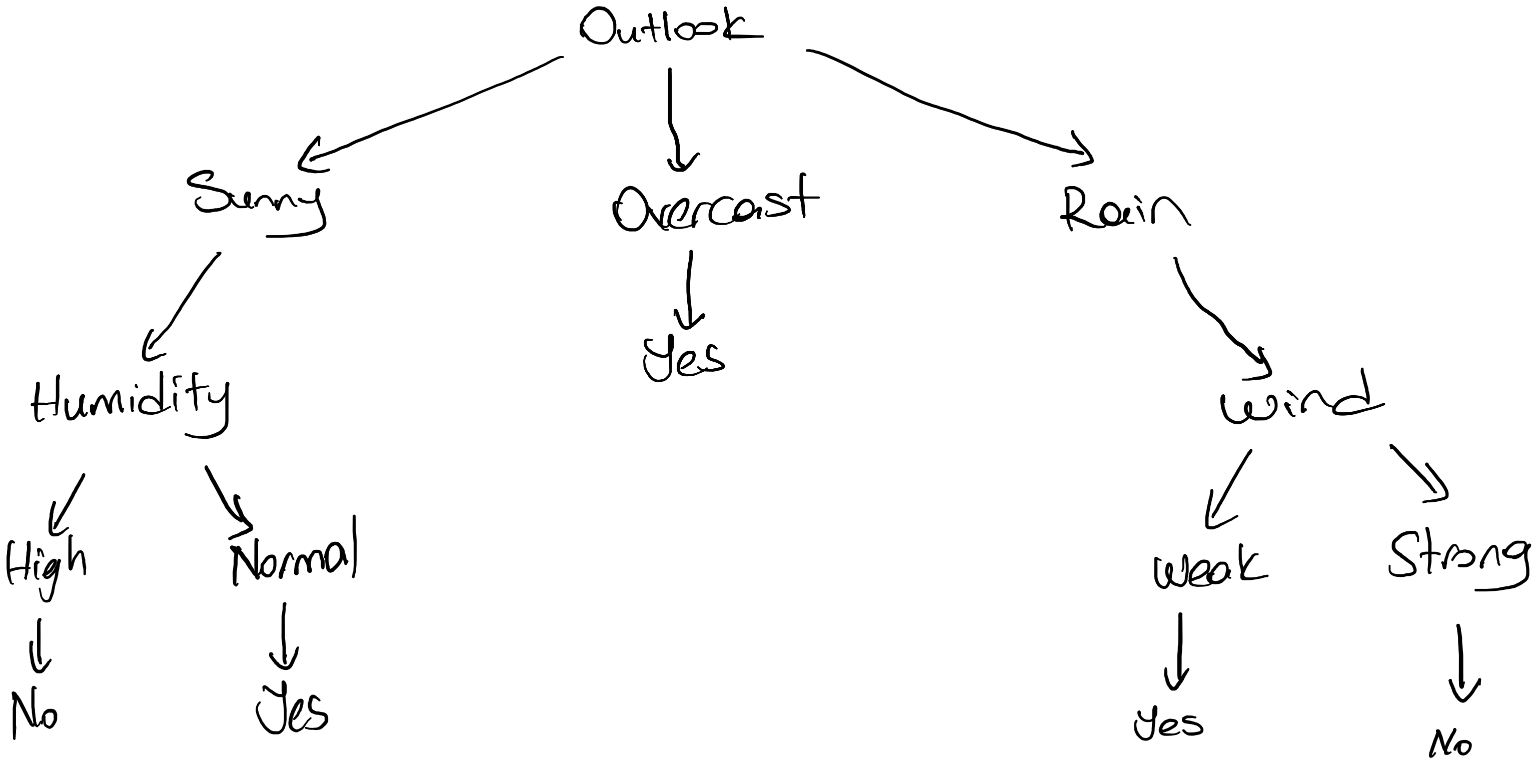
$$\text{Gain}(\text{Rain}, \text{Wind}) = 0.811 - \left(\frac{3}{4}(0) + \frac{1}{4}(0)\right)$$

$$= 0.811$$

$$\text{Gain}(\text{Rain}, \text{Temp}) = 0.811 - \left(\frac{2}{4}(0) + 0.0 + \frac{2}{4}\cdot 2\left(-\frac{1}{2}\log_2\frac{1}{2}\right)\right)$$

$$= 0.311$$

Wind gives the highest information gain



Question-4

$d_{i,j} \rightarrow$ i: Classifier , j: output of the classifier

ω_1 : Class-1 , ω_2 : Class-2

Classifier-1

	Class-1	Class-2
Class 1	40	10
Class 2	30	20

$$P(\omega_1 / d_{1,1}) = \frac{40}{70}$$

$$P(\omega_2 / d_{1,1}) = \frac{30}{70}$$

Classifier-2

	Class-1	Class-2
Class 1	20	30
Class 2	20	30

$$P(\omega_1 | d_{2,2}) = \frac{20}{40}$$

$$P(\omega_2 | d_{2,2}) = \frac{20}{40}$$

Classifier-3

	Class-1	Class-2
Class 1	50	0
Class 2	40	10

$$P(\omega_1 | d_{3,2}) = \frac{0}{10}$$

$$P(\omega_2 / d_{3,2}) = \frac{10}{10}$$

$$\text{Class-1} = \frac{40}{70} \cdot \frac{20}{40} \cdot \frac{0}{10} = 0 \qquad \text{class-2} = \frac{30}{70} \cdot \frac{20}{40} \cdot \frac{10}{10} = 0.214$$

Final Decision is Class-2 = 0.214