Herhalingsles classificeren

1) Gegeven onderstaande dataset die gebruikt kan worden om het ziektebeeld te bepalen op basis van een aantal symptomen.

Training	fever	vomiting	diarrhea	shivering	Classification
d_1	no	no	no	no	healthy (H)
d_2	average	no	no	no	influenza (I)
d_3	high	no	no	yes	influenza (I)
d_4	high	yes	yes	no	salmonella poisoning (S)
d_5	average	no	yes	no	salmonella poisoning (S)
d_6	no	yes	yes	no	bowel inflammation (B)
d_7	average	yes	yes	no	bowel inflammation (B)

Bepaal welk attribuut de hoogste informatiewinst geeft om een eerste splitsing op uit te voeren. Maak gebruik van entropie als onzuiverheidsmaat.

2) Gegeven onderstaande dataset die net als in de eerste oefening gebruikt wordt om een ziektebeeld te bepalen op basis van symptomen.

Training	N running nose	C coughing	R Reddened skin	F Fever	Classification
d_1	+	+	+	-	positive (ill)
d_2	+	+	-	-	positive (ill)
d_3	-	-	+	+	positive (ill)
d_4	+	-	-	-	negative (healthy)
d_5	-	-	-	-	negative (healthy)
d_6	-	+	+	-	negative (healthy)

a. Bepaal alle kansen die wij nodig hebben om m.b.v. de naive bayes classifier te bepalen of een persoon "`ill`" is of "`healthy`". Maak gebruik van de m-schatting met $m = 4$ en $p = 0{,}5$.
b. Classificeer volgende records m.b.v. de naive bayes classifier
$d_7 = (\bar{N}, C, \bar{R}, F)$
$d_8 = (N, \bar{C}, \bar{R}, F)$

3) Veronderstel dat wij een model getraind hebben om documenten te classificeren, de mogelijke (system) class labels zijn Pos (positive), Neg (negative) en Neu (neutral). Wij testen onze classifier (gold) op 10 documenten. Bepaal precision, recall, accuracy en F_1 voor elk class label afzonderlijk (gold = voorspelde klasse, system = werkelijke klasse).

Documents	*gold class*	*system class*
d_1	*Pos*	*Pos*
d_2	*Pos*	*Pos*
d_3	*Pos*	*Pos*
d_4	*Pos*	*Neu*
d_5	*Neg*	*Neg*
d_6	*Neg*	*Neu*
d_7	*Neg*	*Neg*
d_8	*Neu*	*Pos*
d_9	*Neu*	*Neu*
d_{10}	*Neu*	*Neu*