Q1 - For the below data build a Naive Bayes classifier and identify the class of the given data sets noting down the probabilities as the model would save them.

Id	Age	Income	Student	Credit_Rating	Class
A	<=30	low	no	fair	??
B	31…40	high	yes	excellent	??
C	<=30	medium	yes	fair	??
D	>40	high	no	excellent	??

age	*income*	*student*	*credit_rating*	*Class:*
<=30	high	no	fair	no
<=30	high	no	excellent	no
31 . . . 40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31 . . . 40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31 . . . 40	medium	no	excellent	yes
31 . . . 40	high	yes	fair	yes
>40	medium	no	excellent	no

Q2 - For the below data what would be the prediction for play using a Naive Bayes classifier? Please use Gaussian PDF for the continuous variables

Outlook	**Temperature**	**Humidity**	**Windy**
Sunny	87	60	false
Overcast	67	64	true
Overcast	72	71	false
Overcast	80	67	false

outlook	temperature	humidity	windy	play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Q3 - For the below data

Predict the output on (High, no, no, no) and (no, yes, no, yes) using a Naive Bayes model

Fever	Vomiting	Diarrhea	Shivering	**Classification**
no	no	no	no	healty (H)
average	no	no	no	influenza (I)
high	no	no	yes	influenza (I)
high	yes	yes	no	salmonella poisoning (S)
average	no	yes	no	salmonella poisoning (S)
no	yes	yes	no	bowel inflammation (B)
average	yes	yes	no	bowel inflammation (B)

Q4 - For the below data predict the output on (-,+,-,+), (+,-,-,+) and (+,-,+,-) using both a Naive Bayes and a Bayes optimal classifier

N (running nose)	C (coughing)	R (reddened skin)	F (fever)	Classification
+	+	+	–	positive (ill)
+	+	–	–	positive (ill)
–	–	+	+	positive (ill)
+	–	–	–	negative (healthy)
–	–	–	–	negative (healthy)
–	+	+	–	negative (healthy)

Goodness of Fit

Q5. For each of the below confusion matrices compute below details

a. Accuracy
b. Precision
c. Recall
d. Sensitivity
e. Specificity
f. F1 score
g. F2 score
h. F0.5 score
i. Null error rate
j. Balanced accuracy
k. Positive prevalence
l. Negative predictive value

1.

		Observed	
		+ve	-ve
Predicted	-ve	750	2000
	+ve	250	100

2. While predicting Benign tumors

		Predicted	

		Malignant Tumors	Benign Tumors
Observed	Malignant Tumors	100	200
	Benign Tumors	50	5000

3. While detecting Negative Sentiment

		Observed	
		Negative	Non-negative
Predicted	Negative		
	Non-negative		

Q6. For the below data trying to identify Fraud, determine

a. TP
b. FP
c. TN
d. FN
e. Negative Prevalence

	Actual	Predicted
	Fraud	Fraud
	Non-fraud	Non-Fraud
	Non-Fraud	Non-Fraud
	Fraud	Non-Fraud
	Non-Fraud	Non-Fraud
	Non-Fraud	Non-Fraud
	Non-Fraud	Non-Fraud
	Fraud	Non-Fraud
	Non-Fraud	Non-Fraud

Q7. For the below model perf determine the area under the RoC curve (AUC)
TPr = eFPr+FPr3-3/2FPr-1/2