1 Exercise 1

1. This is the answer to Part 1. Due to the nature of the question, some of the descriptions of items within part 1 will also apply to the commentary of other parts. For example, I will not re-describe the positivity rates in subsequent parts even though these are essential components of the ROC and DET curves.

   - The following statistics are calculated
     - Precision - This corresponds with when an item is classified as being true (or postive), what percentage of the time is that the correct classification. This is defined as $P := \frac{(TP)}{(TP)+(FP)} \in [0, 1]$ where TP is True Positives, FP is False Positives.
     - Recall - This corresponds with what what percentage of all positives are actually positive. This is defined as $R := \frac{(TP)}{(TP)+(FN)} \in [0, 1]$ where TP is True Positives, FN is False Negatives.
     - F-Measure - The F_1-Measure is the harmonic mean of precision and recall. It's a performance indicator designed to weigh the trade off between precision and recall and consolidate the information contained from both into a single value. This is defined as $F_1 := 2\frac{(P)(R)}{P+R} \in [0, 1]$ where P is Precision and R is Recall.
     - True Positive Rate - This is equivalent to recall.
     - False Positive Rate - This is the number of false positives over the total number of negatives. It is used to identify the probability that a sample is falsely classified as possitive when it is in fact negatives.
     - True Negative Rate - This corresponds with when an item is classified as being negative, what percentage of the time is that the correct classification.
     - False Negative Rate - This is the number of false negatives over the total number of positives. It is used to identify the probability that a sample is falsely classified as negative when it is in fact positive.
   - I calculate the Precision and Recall for each threshold value. For readability purposes, in my dataframe and plots I set each value as a percentage and round to four decimal places (four is sufficient to distinguish all values within this dataset). This is sensible as the values lie within $[0, 1] \subset \mathrm{R}$ and thus converting to a percentage involves linearly scaling the dataset by 100. In calculating the precision, recall, and f-measure at each threshold, we observe that the precision decreases as the threshold increases, while conversely the recall increases as the threshold increases. This is ultimately because the lower threshold causes lower false positives decreasing the denominator (i.e. of those classified positively, only a low number are false positives meaning true positives are a much higher percent). For recall conversely, the low threshold causes fewer positive flags resulting in a low percentage of all 'positives' being correctly classified, and therefore increasing the threshold results in a higher value. The f-measure is the harmonic mean of precision and recall and therefore factors and balances both measures. In this instance we use the f1-measure which 'equally' accounts for both.
   - In this dataset, we see the best threshold, by F1-measure, is 20 at the F-measure is maximised when the Threshold is 20.

	Threshold	Precision_AtThreshold	Recall_AtThreshold	FMeasure_AtThreshold	TruePositiveRate	FalsePositiveRate	TrueNegativeRate	FalseNegativeRate
0	1	90.91	20.0	32.7869	20.0	2.0	98.0	80.0
1	5	90.91	50.0	64.5164	50.0	5.0	95.0	50.0
2	10	85.71	60.0	70.5868	60.0	10.0	90.0	40.0
3	15	80.00	80.0	80.0000	80.0	20.0	80.0	20.0
4	20	74.58	88.0	80.7361	88.0	30.0	70.0	12.0
5	25	69.23	90.0	78.2604	90.0	40.0	60.0	10.0
6	30	65.52	95.0	77.5530	95.0	50.0	50.0	5.0
7	35	61.54	96.0	75.0011	96.0	60.0	40.0	4.0
8	40	58.08	97.0	72.6562	97.0	70.0	30.0	3.0
9	50	55.06	98.0	70.5067	98.0	80.0	20.0	2.0

Figure 1: Question 1 Dataframe

2. I generate the following plot for the ROC curve. As outlined above, I have scaled everything to be out of 100 so it can be a percentage plot. The ROC curve is a plot of the True Positive Rate vs False Positive Rate across various thresholds, while the AUC is the integral of the ROC (i.e. total area under the curve). We see our ROC curve is in the upper left with an AOC of approximately 0.66 which suggests a decent ability to separate between classes. We see the furtherest perpendicular distance to the line y=x (which would correspond with an AUC of 0.5, i.e. a random classifier) occurs roughly at the point (20,80) which corresponds with when the threshold is 20 and F1 measure is maximised.

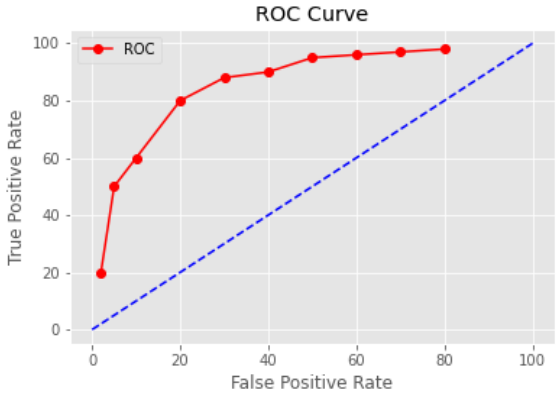

Figure 2: ROC Curve

3. I generate the following plot for the DET curve; first as a log-scaled curve and secondly not using a log scale. As outlined above, I have scaled everything to be out of 100 so it can be a percentage plot.

   The DET curve is a plot of the False Positive and False Negative rates. DET curves are often used in place of ROC curves as while ROC curves typically only differ in the upper left quadrant, DET curves can help provide information at which point the trade off between false positives and false negatives is acceptable. This is often presented using log-scale values and therefore I have presented this view on the right chart in addition to an un-scaled view on the left in the image below. We can see the trade-off between minimising the false positive and false negative rate in the graph with the best 'balance' between the two appearing to visually occur approximately at the threshold value where the false positive and false negative rate are approximately 20%. This values also corresponds with a threshold value of 20. If however somebody valued minimising false positives, a lower point might be taken and the analyst could identifiy the expected consequential false negative rate of a threshold value which would result in the false positive rate.

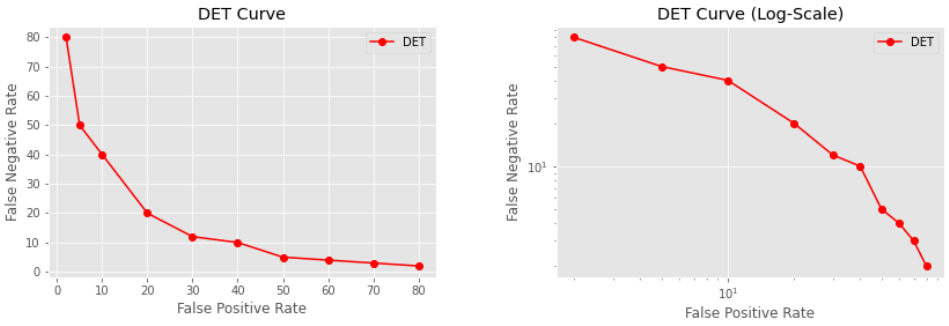

Figure 3: LOG Scaled DET Curve