Answer a)

There are many popular statistical techniques to identify the outliers in a given distribution of data. The choice of the technique depends on:

a) The distribution of data in the selected features(parametric or non-parametric).

b) Dimension of the data (like univariate or multivariate), and any specific pre-processing requirements.

c) the nature of the environment (like point outliers, contextual outliers and collective outliers).

For the sake of simplicity and with reference to the cleveland dataset, we would consider a univariate point outlier detection method. One such popular technique is the **Interquartile Range method.** The criteria of such technique is as follows.

(Xi > (Q3 + k * IQR)) V (Xi < (Q1 - k * IQR)) where,

Xi : An individual data point of the selected variable/feature
Q1: The lower-quartile of the feature vector
Q3: The upper-quartile of the feature vector
IQR: The interquartile range, which is Q3 - Q1
K: A constant greater than zero which is the interquartile multiplier and set to 1.5

The effects of outliers on the data are as follows:

1. Effect on the mean and standard deviation which are not robust against outliers.
2. Makes the distribution skewed towards the outliers.

Hence, for further predictive modeling and analysis, the outliers can have a significance effect.

However, median is robust against outliers unlike mean because median essentially divides the sorted dataset into two halves and it is the middle element. Hence, it doesn’t matter how much the data is skewed, median will always take the middle element. But, the value of mean which is essentially the arithmetic average of all the data points heavily depends on the data point values. So, if the data point values have lots of outliers, the arithmetic average will be changed significantly. For the same reason, standard deviation is heavily affected because of outliers unlike median absolute deviation.

Answer b)

The random variable X can take on the following values X = {10, 15, 17, 19, 20, 22, 28, 30}

i) According to the question, it is observed that pr(MMSE is at least 20) is given by the following:
pr(X>=20) = pr(X=20) + pr(X=22) + pr(X=28) + pr(X=30)

Now, substituting the values of the probabilities from the given table, we get:

pr(X >=20) = 0.03 + 0.07 + 0.02 + 0.07
= 0.19
Hence, the probability that MMSE is at least 20 is **0.19**

ii) According to the question, it is observed that pr(MMSE is at least 15 and at most 22) is given by the following:
pr(X>=15 and X<=22) = pr(X=15) + pr(X=17) + pr(X=19) + pr(X=20) + pr(X=22)
= 0.23 + 0.28 + 0.09 + 0.03 + 0.07
= 0.7
Hence, the probability that MMSE is at least 15 and at most 22 is **0.7**

Answer c)

The random variable can be modelled with Binomial distribution since all the testing are done independently with only two possible outcomes i.e. either they are tested correctly or incorrectly.

The following data are given in the question.
Probability that it is false positive p = 1% i.e. 0.01…………………………………………………..(i)
Number of independent trials n = 30………………………………………………………………(ii)

From (i):
Probability that it is not false positive is q = 1-p = 0.99……………………………………………(iii)

We have to find out the probability of at least 2 false positives .i.e. In other words probability of 1 - (probability of exactly 0 false positive + probability of exactly 1 false positive)

Applying the binomial distribution formula:
pr(x) = nCx . (p)**x. (q)**(n-x), where '**' denotes power and C denotes combination symbol

pr(0) = 30C0 . (0.01)**0 . (0.99)**30 = 0.73970……………...(iv)

pr(1) = 30C1 . (0.01)**1 . (0.99)**29 = 0.22415……………………………………………………..(v)

Hence, the probability of at least 2 false positives for 30 samples is 1 - pr(0) - pr(1).
Substituting the values of (iv) and (v):
Desired Probability = 1 - 0.73970 - 0.22415 = **0.03615**