Question 1:

Total number of students = 150

Chances of each student carrying coronavirus = 0.02

Chances of each student not having coronavirus = (100 − 2)%

= 98% = 0.98

⇒ Probability of entire class being coronavirus free = $(0.98)^{150}$

= 0.048

Chances of Omicron variant = 95% = 0.95

Chances of each person not having omricon variant =

= 1 − Chances of each student carrying omricon

= 1 − (95% of 2%)

= 1 − (0.95 × 0.02)

= 1 − 0.019

= 0.981

⇒ Probability of entire class being Omicron free = $(0.981)^{150}$

= 0.056

Question 2:

a) You randomly select $n \times m_i/m$ elements from each group.
We refer to this method of sampling as stratified sampling.

(b) You randomly select n elements from the data set, without regard for the group to which an object belongs.
This type of sampling is called simple random sampling.

Difference between Stratified and Simple Random Sampling:
Simple Random Sampling:
- Elements are randomly chosen from the entire dataset in this sampling without any restrictions.
- It is employed when there is a lack of data on the data set. For instance, when there is just one trait or when there are too many for classification.
- It provides a summary or generalization of the data.

Stratified Random Sampling:
- Here, we choose elements at random from each group. Since we are choosing elements from each group, we end up with elements that have a variety of distinct qualities.
- Each element that is selected from each group should only belong to that group.

Which is better?
In my opinion, stratified sampling is better because of the following reasons:
- Even if we are selecting components without bias and getting a general sense of the dataset when using basic random sampling, there are times when we may overlook significant samples. Think of a chocolate factory, for instance, where a new chocolate bar needs to be created based on feedback from 1000 clients. There is no guarantee that the entire population is accurately represented because customers are chosen at random in random sampling.
- Since samples from all categories are taken into account in stratified sampling, the total data can be better understood or represented without being missing any significant details. Additionally, this will aid in showing the variations between each feature. Both in terms of time and space, this is effective.

Question 3

i) $x = (1, 0, 0, 1, 1, 1)$ $\quad y = (1, 1, 1, 0, 0, 1)$

Cosine Similarity, $\cos(x, y) = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|} = \frac{1\times1 + 0\times1 + 0\times1 + 0\times1 + 0\times1 + 1\times1}{\sqrt{1^2+1^2+1^2+1^2} \times \sqrt{1^2+1^2+1^2+1^2}}$

$= \frac{2}{2 \times 2} = \frac{1}{2} = 0.5$

Euclidean distance $= d(x, y) = \sqrt{(x-y)(x-y)} = \sqrt{1^2+1^2+1^2+1^2}$

$= 2$

Correlation, $corr(x, y) = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$

$\bar{x} = \frac{1+1+1+1}{6} = \frac{2}{3} = 0.67$

$\bar{y} = \frac{1+1+1+1}{6} = \frac{2}{3} = 0.67$

$$corr(x, y) = \frac{(1-0.67)(1-0.67) + (0-0.67)(1-0.67) + (0-0.67)\times(1-0.67) + (1-0.67)(0-0.67) + (1-0.67)\times(0-0.67) + (1-0.67)(1-0.67)}{\sqrt{(1-0.67)^2 + (0-0.67)^2 + (0-0.67)^2 + (1-0.67)^2 + (1-0.67)^2 + (1-0.67)^2} \times \sqrt{(1-0.67)^2 + (1-0.67)^2 + (1-0.67)^2 + (0-0.67)^2 + (0-0.67)^2 + (1-0.67)^2}}$$

$$= \frac{0.1089 + (-0.2211) + (-0.2211) + (0.2211) + (-0.2211) + 0.1089}{\sqrt{1.334} \times \sqrt{1.334}}$$

$= -0.495$

Jaccard similarity $= \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$

$= \frac{2}{2+2+2}$

$= \frac{1}{3} = 0.3334$

d) $x = (1, -2, 0, 2, 0, -3)$ $y = (-1, 2, -1, 0, 0, -1)$

Cosine Similarity, $\cos(x,y) = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|} = \frac{1\times-1 + -2\times2 + 0\times-1 + 2\times0 + 0\times0 + -3\times-1}{\sqrt{1^2+2^2+2^2+3^2} \times \sqrt{1^2+2^2+1^2+1^2}}$

$= \frac{-2}{3\sqrt{2}\times\sqrt{7}} = -0.178$

Euclidean distance, $d(x,y) = \sqrt{(x-y)(x-y)} = \sqrt{(-1+1)^2+(-2-2)^2+1^2+2^2+0^2+(-3+1)^2}$

$= \sqrt{4+16+1+4+4} = \sqrt{29} = 5.39$

Correlation, $corr(x,y) = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

$\bar{x} = \frac{1-2+2-3}{6} = -0.33$ $\bar{y} = \frac{-1+2-1-1}{6} = -0.167$

$corr(x,y) = \frac{(1+0.33)(-1+0.167) + (-2+0.33)(2+0.167) + (0+0.33)(-1+0.167) + (2+0.33)(0+0.167) + (0+0.33)(0+0.167) + (-3+0.33)(-1+0.167)}{\sqrt{(1+0.33)^2+(-2+0.33)^2+(0+0.33)^2+(2+0.33)^2+(0+0.33)^2+(-3+0.33)^2} \times \sqrt{(1+0.167)^2+(2+0.167)^2+(-1+0.167)^2+(0+0.167)^2+(0+0.167)^2+(-1+0.167)^2}}$

$= \frac{-2.33}{4.16 \times 2.614} = -0.214$

Question 4:

PCA analyses:

- A technique for reducing dimensionality is PCA. The correlation between characteristics is found using PCA. The associated properties become independent of one another following PCA. The data starts overfitting when there are too many features in the dataset, but PCA helps prevent it by limiting the number of features.

- We can see that the variability is linear from the graph that was drawn for the two dimensions.

- We can also see that when variance is high, data points are scattered and clusters develop in PCA.

- PCA is not an effective strategy for dimensionality reduction, in my opinion. It is clear from the scatter plot that it incorrectly identifies cases of n and t2d. Even though PCA aims to account for as much variance across the dataset's features as possible, it can still leave out some data when compared to the original features.

- t-SNE is a result in a good dimensionality reduction of this dataset. The local and overall structure of the dataset is maintained via t-SNE. In t-SNE, the points that are comparable in high dimensions are also similar in low dimensions. t-SNE is also effective with both linear and non-linear data. It is clear from the scatterplot graph that the data is better visualized with t-SNE.

Question 5:

Curse of dimensionality

The curse of dimensionality states that as the size of the dataset rises, the model's performance begins to fall rather than rise.

Here we are computing the Euclidean distance between points x and y whose values are random integers generated from 0 to 500.

This is carried out for various dimensions from 2 to 50. We can see that while there are fewer dimensions, the distance is greater, and as there are more dimensions, the Euclidean distance begins to decrease. This is really effectively illustrated in the graph.

This is the problem with dimensionality: it gets harder and harder to extract key information as the dimension increases.