**Question: Explain dimension reduction in machine learning.**

In machine learning especially classification problems, often there are too many factors which can be included while making classifications. These factors are basically variables called features.

Higher is the number of features, the harder it is to understand the training set then train the model on it. This may also result in the model on the training set performing poorly. At times, higher number of features will computational power especially if Techniques such as ANN are used to train the model.

At times, many of the features can be correlated, and hence not required. At times there may be features, which may not be significant in determining the outcome of the classification but only increase the complexity of the model.

Therefore we might need to remove such features.

Example:

classifying whether the e-mail is spam or not.

Features involved:

* Title of the email,
* if the e-mail uses a template,
* common words in the body of the email.

While in the above example it looks only 3 features but after text mining, there can be more than 100, 2- grams, 3 grams formed from the content of the email. Hence it becomes more important to understand what n-grams are more in the email body and have more importance. Accordingly, the number of features can then be reduced to 15-20 to train the model.

components of dimensionality reduction:

* **Feature selection:** Select a few features from a larger set of features.
* **Feature extraction:** Create new features which result in fewer features but represent the same information as the original ones.

**Methods of Dimensionality Reduction**

* Principal Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)
* Factor Analysis
* Sometimes we can identify features to be removed by simple data visualisation by identifying correlated features or features which have a lot of missing data.