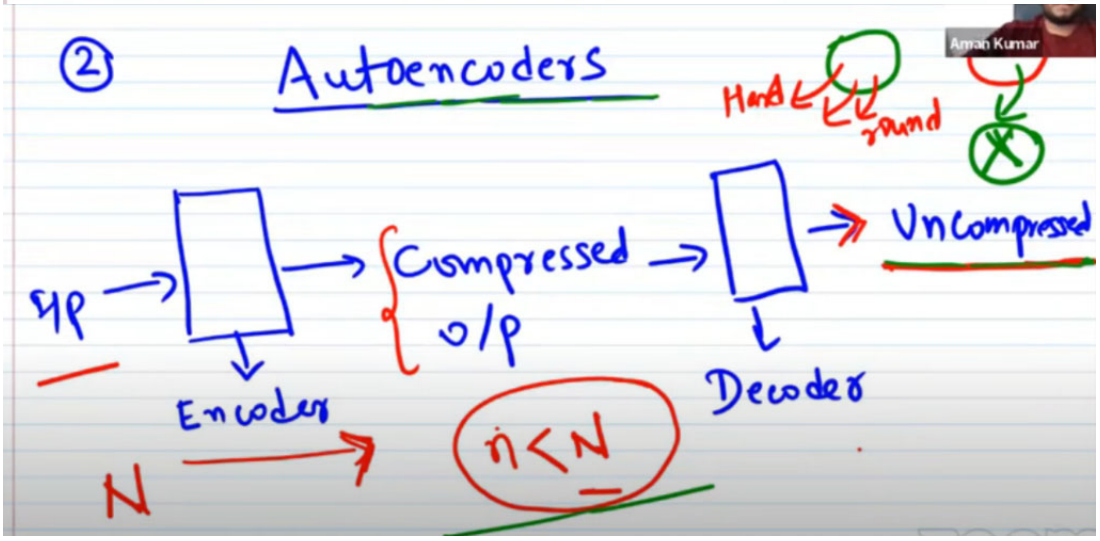
Session 7

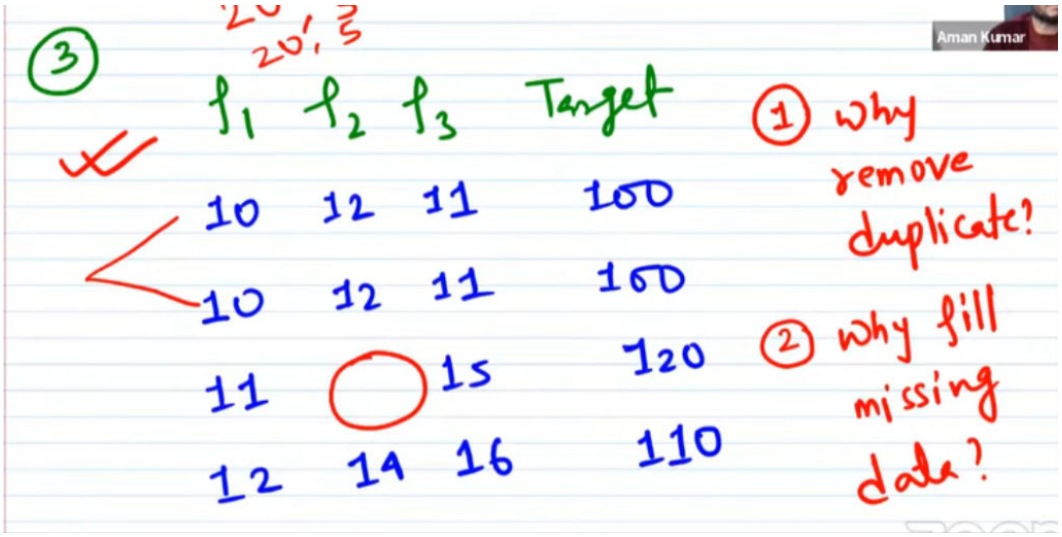
Wednesday, April 12, 2023 5:21 PM



Autoencoders are a type of neural network that can learn to compress and decompress data, typically used for unsupervised learning tasks. The basic idea behind an autoencoder is to encode the input data into a lower-dimensional representation, or "code," and then decode the code back into the original input data. By doing so, the autoencoder learns to capture the most important features or patterns in the input data.

Autoencoders consist of two main components: an encoder and a decoder. The encoder takes in the input data and transforms it into a code, while the decoder takes in the code and reconstructs the original data. The goal of training an autoencoder is to minimize the difference between the input data and the reconstructed data, which encourages the model to learn a good encoding and decoding strategy.

Autoencoders have many practical applications, including image compression, anomaly detection, and generative modeling. They have also been used as a building block in more complex deep learning architectures, such as variational autoencoders and denoising autoencoders.



Removing duplicates from a dataset helps to improve the quality of the data and can help to prevent overfitting when creating a model. Here are some reasons why removing duplicates is important:

1. Reduces bias: If a dataset contains duplicate examples, these examples may be overweighted during training, leading to biased model predictions.
2. Improves model accuracy: By removing duplicates, you ensure that each example in the dataset is unique and contributes to the learning process. This can lead to a more accurate model.
3. Saves computational resources: Removing duplicates reduces the size of the dataset, which can save storage space and reduce the computational resources required to train the model.
4. Improves generalization: Duplicate data can cause a model to overfit to specific examples in the dataset, leading to poor generalization to new, unseen data. Removing duplicates can help to prevent this problem and encourage the model to learn more general patterns in the data.

Overall, removing duplicates is an important step in preparing a dataset for machine learning, as it helps to ensure that the model learns from a high-quality, diverse set of examples.

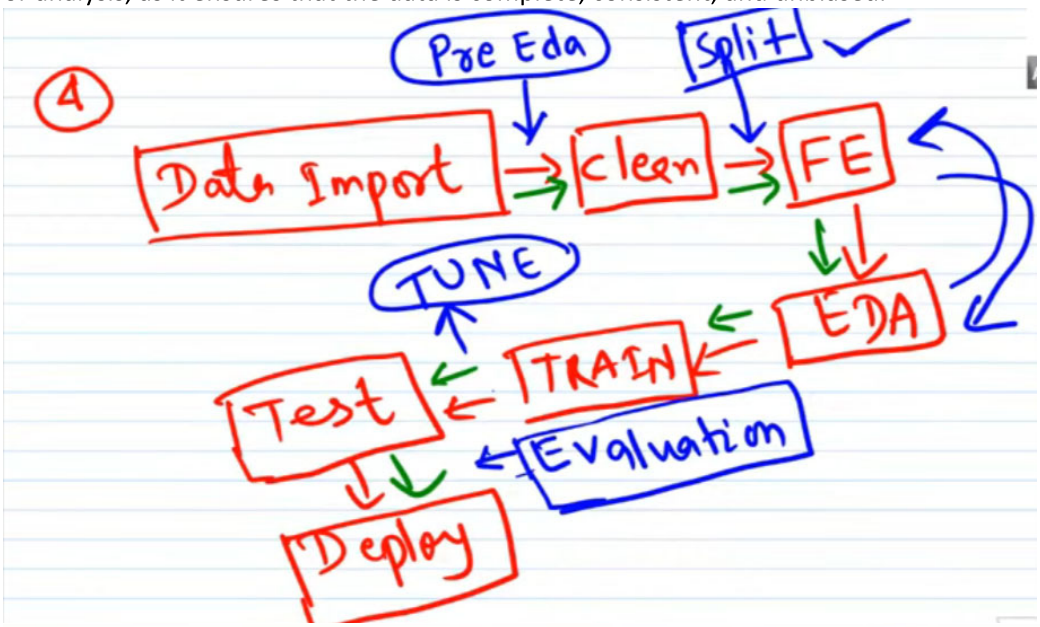
Filling missing data or filling in the gaps in a dataset is important because many machine learning algorithms cannot handle missing values. Missing data can lead to biased predictions or inaccurate models if not handled properly.

Here are some reasons why filling missing data is important:

a. Avoids bias: If missing values are not filled, it could lead to biased model predictions, especially if the missing data is not distributed randomly.

b. Improves accuracy: Filling in the missing data provides more complete information to the model, which could lead to a more accurate prediction.
c. Enables better analysis: Missing data can impact statistical analysis and make it difficult to identify patterns or trends in the data.
d. Consistency: Filling in missing data ensures that all data points in the dataset have the same format and structure, making the dataset consistent.
e. There are different techniques to fill missing data such as using the mean or median of the known values, using interpolation, backfilling, forward-filling, and using machine learning methods such as KNN imputation, decision trees, and regression models.

Overall, filling missing data is an important step in preparing a dataset for machine learning or analysis, as it ensures that the data is complete, consistent, and unbiased.



Test data should never be exposed in learning process.