

→ **Bias:** Assumptions made by model to make function easier to learn.

- It is error rate of training data.
- When error rate is high its high bias and when it is low its low bias.

→ **Variance:** It is basically spread of data.

- The difference between the error rate of training and test data is called variance.
- If the difference is high its high variance else low variance.

→ **Training data:**

- It is subset of the original dataset that is fed to model to discover and learn patterns.

→ **Testing Data:**

- It is also a subset of the original dataset which is completely unknown to model and used to test the accuracy of the model.

→ **Training data** has more data points compared to testing data.

• Test data is completely isolated from model to prevent data leakage and avoid alteration in accuracy.

• Usually train-test data split is kept 80:20%.

→ **Overfitting:**

- A model is overfitted when model makes good prediction on training data but fails to do so on test data.

- This happens because model was unable to categorize data correctly due to noise leading it to learn from inaccurate and noise data.

• Other reasons for overfitting are:

- model too complex
 - size of training data
 - non-parametric, non-linear methods
- Symptom:

Low bias and high variance

or

High accuracy on training data but low accuracy on test data.

• Solution:

- Increase training data
- Reduce model complexity
- Early stop in training
- Regularization methods

→ **Underfitting:**

- A model is underfitted when model is unable to capture the trend underlying in data. Thus giving bad prediction in training data and may or may not good prediction on test data.

• Reasons for underfitting:

- Training data size
- Too simple model
- Noisy training data

• Symptom:

High bias and low or high variance

or

Low accuracy on training data and may or may not high accuracy on test data.

• Solution:

- Increase model complexity
- Increase train data size, features
- Increase epochs
- Reduce noise

→ **Generalized model:**

- A model which is neither overfitted or underfitted and has good accuracy on training and test data both.

• Such models have:

Low Bias and Low Variance

