Assignment-1

Variance and Bias (Diagram, overfit, underfit)-For best fit model should we have low bias or high variance, low bias or low variance, high bias or high variance, low bias or high variance

Solution :

➢ Introduction

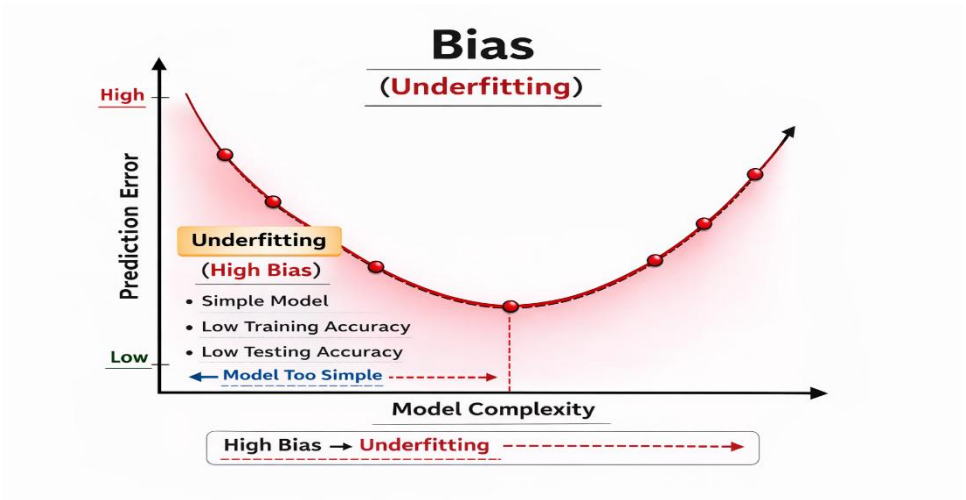
Machine Learning (ML) is a branch of Artificial Intelligence (AI) that enables computers to learn patterns from data and make predictions or decisions without being explicitly programmed. The success of a machine learning model depends on how well it can learn from training data and how accurately it can predict outcomes on new, unseen data. Two key concepts that strongly influence the performance of machine learning models are **bias** and **variance**.

Bias and **variance** are two major sources of error in machine learning. Bias represents the error caused by overly simple assumptions in the learning algorithm, while variance represents the error caused by excessive sensitivity of the model to changes in the training data. Both bias and variance affect the accuracy, stability, and reliability of predictions.

If a model has **high bias**, it fails to capture the complexity of the data and leads to **underfitting**. On the other hand, if a model has **high variance**, it learns noise and unnecessary details from the training data, leading to **overfitting**. Therefore, a good machine learning model should maintain a proper balance between bias and variance to minimize prediction errors.

1. Concept of Bias

Bias refers to the systematic error introduced into a machine learning model due to simplifying assumptions. It measures how far the predicted values are from the actual values because the model is too simple to represent the underlying data patterns.In simple words, bias occurs when a model **does not learn enough from the training data**, resulting in inaccurate predictions.



❖ **Causes of Bias**

Bias can occur due to several reasons, including:

- Choosing a very simple model.
- Ignoring important input features.
- Making incorrect assumptions about the data distribution.
- Applying excessive regularization.
- Poor feature selection and extraction.

❖ When bias is high:

- The model performs poorly on training data.
- The model performs poorly on testing data.
- The model fails to learn important patterns.
- Underfitting occurs.

❖ **Example of Bias**

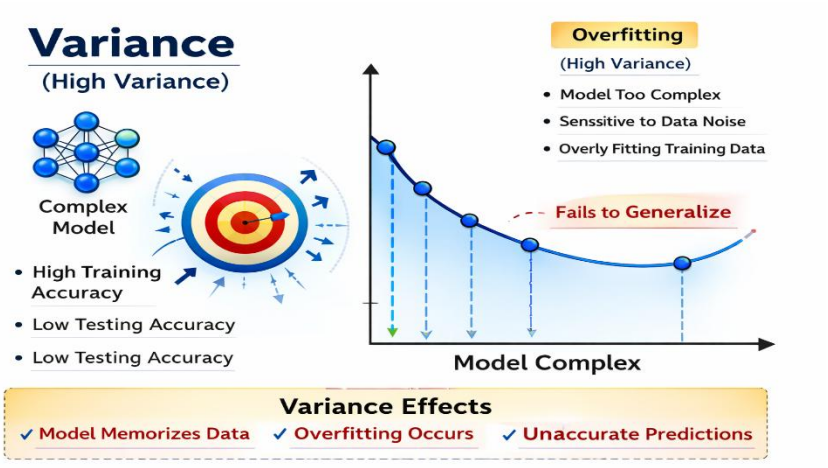
If a linear regression model is applied to data that follows a curved or non-linear pattern, the model cannot represent the curve properly. As a result, predictions become inaccurate, and bias becomes high.

❖ **Characteristics of High Bias Models**

- Simple structure
- Low training accuracy
- Low testing accuracy
- Poor learning capability
- Underfitting

2. Concept of Variance

Variance refers to the degree to which a model's predictions change when it is trained on different training datasets. A model with high variance becomes extremely sensitive to small changes in the data and often learns noise instead of useful patterns.In simple terms, variance



❖ **Causes of Variance**

Variance increases due to:

- Very complex models.
- Training on small datasets.
- Using too many features.
- Lack of regularization.
- Overfitting noise in the data.

❖ **Effects of High Variance**

High variance results in:

- Very high training accuracy.
- Very low testing accuracy.
- Poor generalization.
- Overfitting.

❖ **Example of Variance**

A deep decision tree trained on a small dataset often memorizes the training samples. While it performs extremely well on training data, it fails to generalize on new data, resulting in high variance.

❖ **Characteristics of High Variance Models**

- Complex structure
- High training accuracy
- Low testing accuracy
- Overfitting
- Poor generalization

3. Underfitting and Overfitting

3.1 Underfitting

Underfitting occurs when a model is too simple to capture the actual relationship between input and output data.

❖ **Causes of Underfitting:**

- High bias
- Simple algorithms
- Limited features
- Excessive regularization

❖ **Effects of Underfitting:**

- Low training accuracy
- Low testing accuracy

- Poor predictive performance

❖ **Example:**
Using a straight-line model to represent complex curved data.

3.2 Overfitting

Overfitting occurs when a model learns the training data too closely, including noise and random variations, leading to poor performance on unseen data.

❖ **Causes of Overfitting:**
- High variance
- Very complex models
- Too many features
- Small training dataset

❖ **Effects of Overfitting:**
- Very high training accuracy
- Very low testing accuracy
- Poor generalization
- Unstable predictions

❖ **Example:**
A deep neural network that memorizes training examples instead of learning meaningful patterns.

4. Bias–Variance Tradeoff

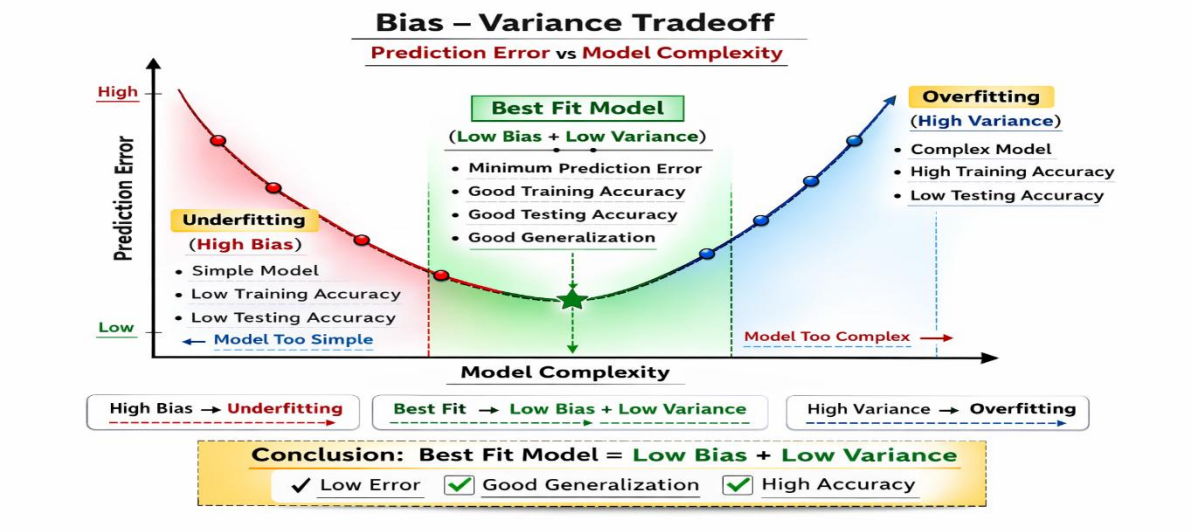
The **bias–variance tradeoff** describes the balance between bias and variance. Reducing bias often increases variance, while reducing variance often increases bias. The goal is to achieve an optimal balance that minimizes total prediction error.

❖ **Total Error Formula**

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

❖ **Explanation**
- Simple models → High bias, low variance → Underfitting
- Complex models → Low bias, high variance → Overfitting
- Balanced models → Low bias and low variance → Best fit

5. Diagram: Bias–Variance Relationship



❖ Diagram Explanation

- The **left side** shows underfitting caused by high bias.
- The **right side** shows overfitting caused by high variance.
- The **middle region** represents the best-fit model where bias and variance are balanced.

6. Explanation of Best-Fit Model

The **best-fit model** is the one that learns the actual patterns from data while also generalizing well to new, unseen data.

- **Low Bias** ensures the model captures complex patterns.
- **Low Variance** ensures stable predictions and good generalization.

Thus, **low bias and low variance** together result in high accuracy, reliable predictions, and minimal error.

7. Why Other Options Are Incorrect

❌ Low Bias and High Variance

This leads to **overfitting**, where the model memorizes data and performs poorly on new data.

❌ High Bias and High Variance

This is the **worst combination**, leading to both poor learning and unstable predictions.

❌ High Bias and Low Variance

This results in **underfitting**, where the model is too simple to learn meaningful patterns.

8. Comparison Table: Bias vs Variance

Feature	Bias	Variance
Meaning	Error due to assumptions	Error due to sensitivity
Model Type	Simple	Complex
Main Problem	Underfitting	Overfitting
Training Accuracy	Low	High
Testing Accuracy	Low	Low
Control Methods	Increase complexity	Regularization, more data

9. Applications of Bias–Variance Tradeoff

- Medical diagnosis systems
- Stock market prediction
- Weather forecasting
- Face recognition
- Recommendation systems

➢ Conclusion

Bias and variance are two fundamental sources of error in machine learning models. High bias leads to underfitting, while high variance leads to overfitting. Both situations result in poor predictive performance.

The **best-fit model** is achieved when **both bias and variance are low**, ensuring that the model learns meaningful patterns and generalizes well to new data. This balance, known as the **bias–variance tradeoff**, is essential for building accurate, reliable, and efficient machine learning systems.

Therefore, the **correct choice for the best-fit model is: Low Bias and Low Variance**.