ASSIGNEMENT – 1

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.

1. Introduction

In machine learning, building an accurate predictive model requires balancing two important concepts: Bias and Variance. These two factors directly affect how well a model learns from training data and how accurately it predicts unseen data. Understanding bias and variance helps in identifying problems like underfitting and overfitting, and in selecting the most appropriate model complexity.

This report explains bias, variance, their combinations, the bias–variance trade-off, and identifies the best model condition for optimal performance.

2. Bias

Definition

- Bias refers to the error introduced due to overly simple assumptions made by a machine learning model to approximate real-world data.

Explanation

- Bias measures how far the model's predictions are from the actual values.
- A model with high bias pays very little attention to training data.
- It oversimplifies the problem.
- It fails to capture important patterns.

Characteristics of High Bias

- Oversimplified model
- Ignores complex relationships
- High training error
- High testing error

Example

Using a straight line (linear regression) to fit highly curved data.

3. Variance

Definition

Variance refers to how much the model's predictions change when trained on different training datasets.

Explanation

- Variance measures model sensitivity to fluctuations in training data.
- A high variance model learns too much from training data.
- It captures noise along with patterns.
- It performs poorly on unseen data.

Characteristics of High Variance

- Very complex model
- Low training error
- High testing error
- Sensitive to small data changes

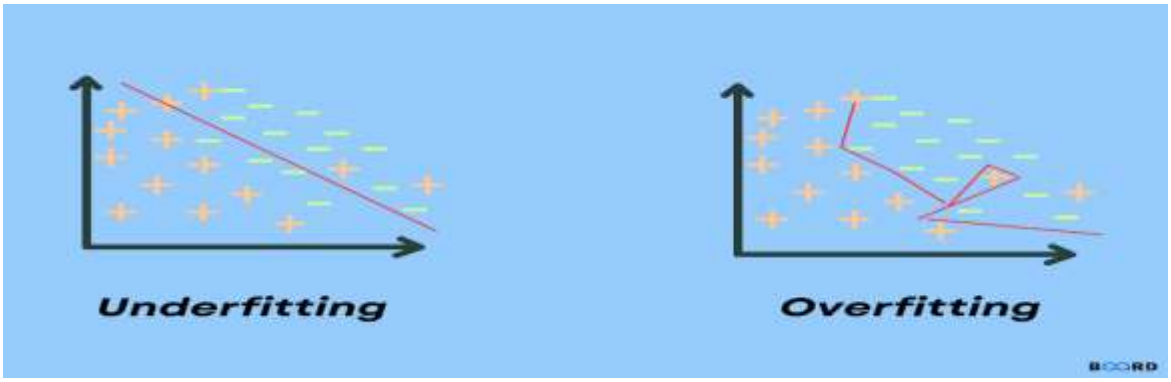
Example

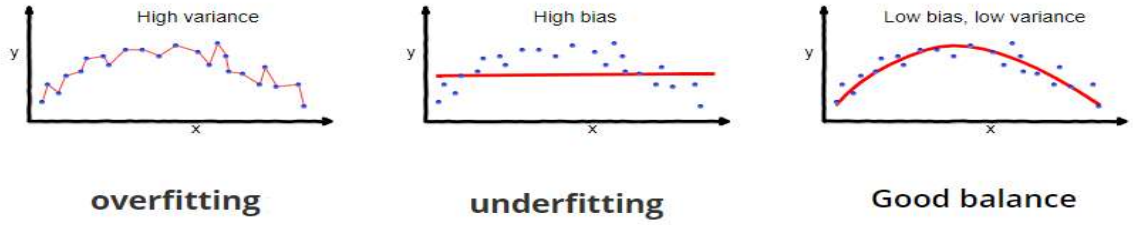
Using a very deep decision tree that perfectly fits training data but fails on new data.

4. Bias–Variance Combinations

A. High Bias & Low Variance

- Model is too simple.
- Consistent but inaccurate predictions.
- Fails to learn patterns properly.
- Leads to Underfitting.

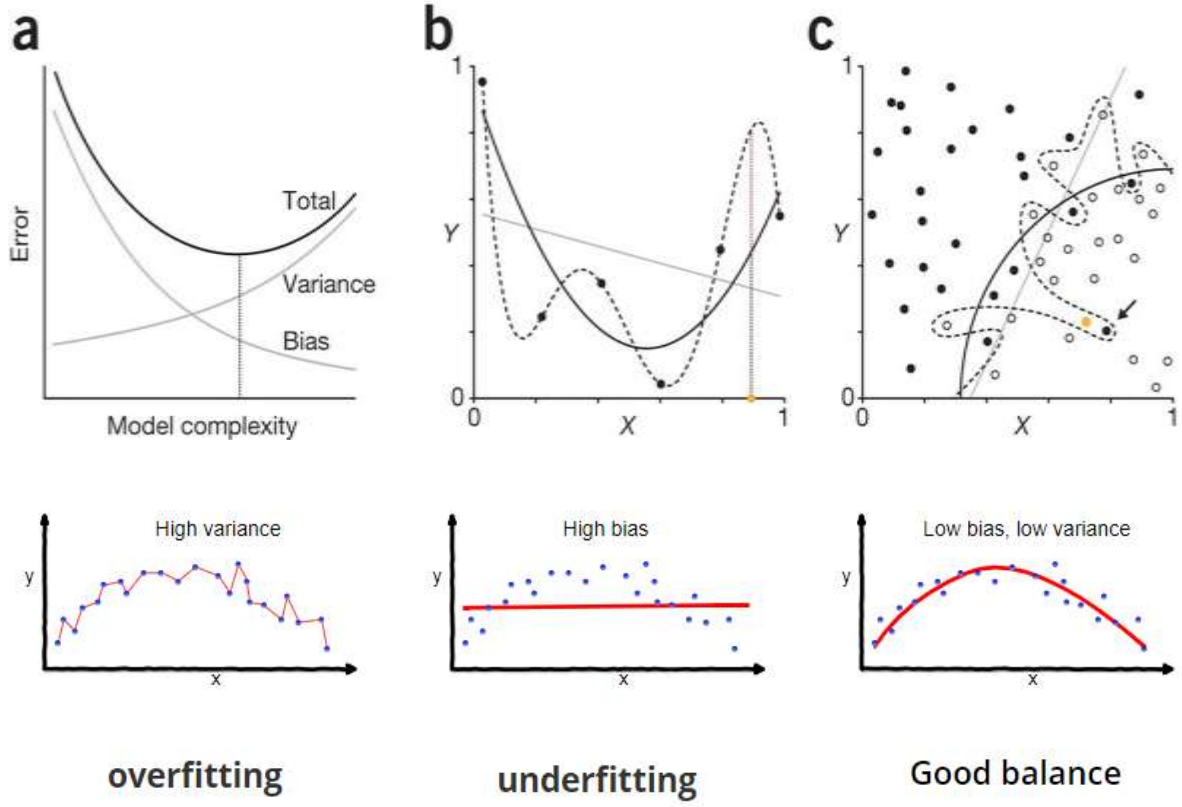




Example: Linear model for non-linear data.

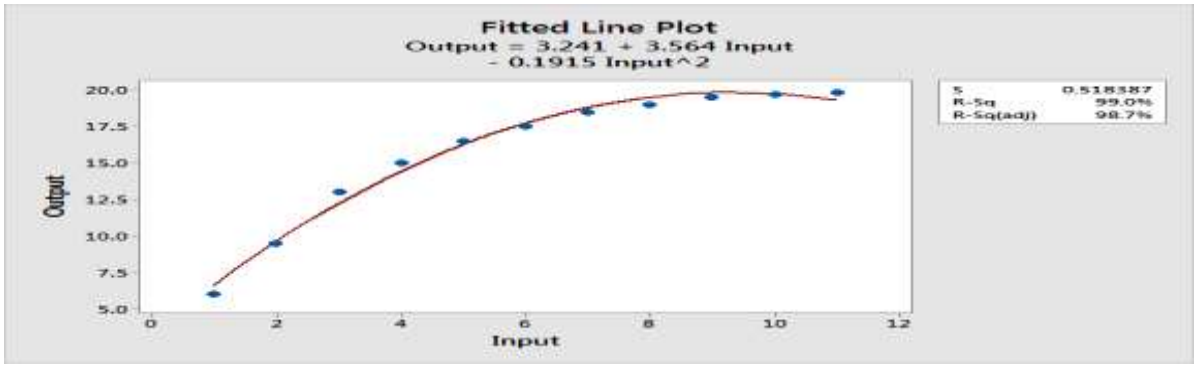
B. Low Bias & High Variance

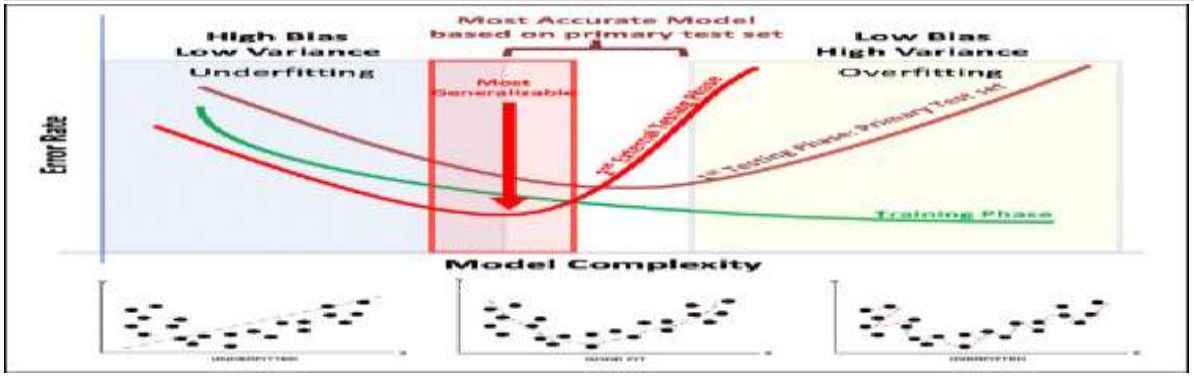
- Model is too complex.
- Learns training data very well.
- Sensitive to noise.
- Leads to Overfitting



Example: Very deep neural network on small dataset.

C. Low Bias & Low Variance





- Ideal model condition.
- Captures patterns correctly.
- Stable predictions across datasets.
- Good training and testing performance.

This is the best combination for a machine learning model.

D. High Bias & High Variance

- Worst case scenario.
- Model is both inaccurate and unstable.
- Poor performance on both training and testing data.

Points are scattered and far from centre → inaccurate and inconsistent.

5. Underfitting & Overfitting

Underfitting

Underfitting occurs when the model is too simple to capture the underlying structure of data.

Causes:

- High bias
- Insufficient model complexity
- Too few features

Effects:

- Poor performance on training data
- Poor performance on testing data
-

Overfitting

Overfitting occurs when the model learns training data too well, including noise and outliers.

Causes:

- High variance
- Too many features
- Very complex model

Effects:

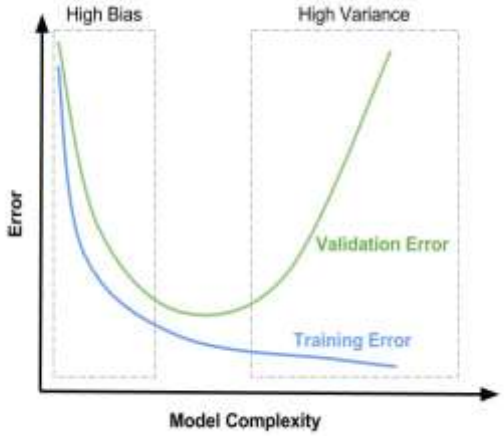
- Excellent training accuracy
- Poor testing accuracy

6. Bias–Variance Trade-off

The Bias–Variance Trade-off describes the relationship between model complexity and prediction error.

- Increasing model complexity decreases bias.
- Increasing model complexity increases variance.
- Optimal performance lies in balancing both.

Bias–Variance Trade-off Diagram



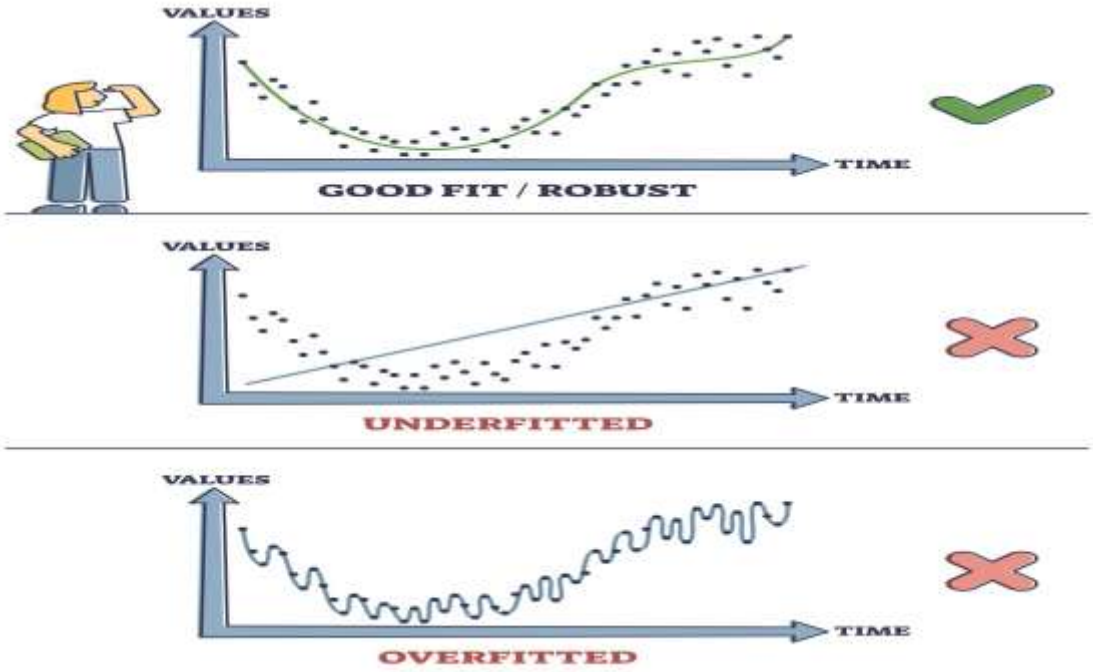


Diagram Explanation

- X-axis → Model Complexity
- Y-axis → Error
- Left side → High Bias (Underfitting)
- Right side → High Variance (Overfitting)
- Middle point → Optimal Model

At the optimal point:

- Bias is low
- Variance is controlled
- Total error is minimized

7. Mathematical Understanding

Total prediction error can be represented as:

$$TotalError = Bias^2 + Variance + IrreducibleError$$

- **Bias²** → Error due to wrong assumptions
- **Variance** → Error due to model sensitivity
- **Irreducible Error** → Noise present in data

Goal: Minimize Bias² + Variance.

8. Best Model Combination

Low Bias & Low Variance is the best combination.

Because:

- Model captures actual patterns.
- Predictions are stable.
- Good generalization.
- Balanced complexity.
- Minimum total error.

This condition ensures the model performs well on both training and unseen data.

9.Summary Table

Condition	Bias	Variance	Result
High Bias, Low Variance	High	Low	Underfitting
Low Bias, High Variance	Low	High	Overfitting
Low Bias, Low Variance	Low	Low	Best Model
High Bias, High Variance	High	High	Poor Model

10. Conclusion

- In machine learning, bias and variance are the two primary sources of prediction error that directly influence model performance.
- Bias and variance are the two main sources of prediction error in machine learning models. Together, they determine how well a model learns from data and how accurately it predicts new outcomes.
- High bias leads to **underfitting** because the model is too simple to capture the true relationship between input and output variables. Such models make strong assumptions and ignore important patterns in the dataset.
- High variance leads to **overfitting** because the model becomes too complex and sensitive to small changes in the training data. It captures noise and random fluctuations instead of the real underlying structure.
- The **bias–variance trade-off** explains that reducing bias usually increases variance, and reducing variance usually increases bias. Therefore, model performance depends on finding the right balance between simplicity and complexity.
- The best model achieves **low bias and low variance**, meaning it accurately captures the data pattern while remaining stable across different datasets.
- A balanced model performs well on both training and testing data, showing strong **generalization ability**. Generalization is the true measure of a successful machine learning model.
- Proper model selection, hyperparameter tuning, cross-validation, feature selection, regularization (L1/L2), pruning, and early stopping help control bias and variance effectively.
- Increasing dataset size can reduce variance, while increasing model complexity (carefully) can reduce bias.
- The total prediction error can be expressed as:
  Total Error = Bias² + Variance + Irreducible Error
  The main goal is to minimize Bias² and Variance while understanding that irreducible error cannot be removed.
- Minimizing total error ensures better generalization, improved accuracy, stable predictions, and reliable real-world performance.