

Assignment 1

Q). 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

Domain Chosen: Diabetes Prediction (Healthcare Domain)

1.Introduction:

- In Machine Learning, Bias and Variance are two types of errors that affect model performance.
- To understand this concept clearly, we consider the domain of Diabetes Prediction, where the goal is to predict whether a patient has diabetes based on:
 - Glucose level
 - Blood pressure
 - BMI
 - Age
 - Insulin level

This is a classification problem.

2. Model Chosen: Decision Tree Model

We use a Decision Tree classifier. Now we analyze how bias and variance affect this model.

3. High Bias and Low Bias

a). High Bias

High bias occurs when the model is too simple and cannot capture the true relationship in data.

Example in Diabetes Domain:

If we use a very small decision tree (only one split like:

"If Glucose $> 120 \rightarrow$ Diabetic, else Not Diabetic"),

Then:

- It ignores other important features like BMI, Age, Insulin.
- Training accuracy is low.
- Testing accuracy is also low.

This is called Underfitting. High Bias + Low Variance.

b). Low Bias:

Low bias occurs when the model captures the pattern in training data correctly.

Example: If the decision tree considers:

- Glucose
- BMI
- Age
- Insulin
- Blood pressure

Then: It learns relationships properly & Training error becomes low.

Bias becomes low because the model is not oversimplified.

4. High Variance and Low Variance:

a). High Variance:

High variance occurs when the model becomes too complex and sensitive to training data.

Example:

If we build a very deep decision tree:

- It perfectly classifies all training patients.
- Training accuracy = 100%.
- But when new patient data comes, prediction is wrong.

This is called Overfitting. Low Bias + High Variance.

Reason: The model memorizes training data instead of learning general patterns.

b). Low Variance:

Low variance occurs when the model performs similarly on training and testing data.

Example: If we prune the decision tree properly:

- Training accuracy is good.
- Testing accuracy is also good.
- The model generalizes well.

Variance becomes low.

5. Model Identification: In this explanation, when we use a very deep decision tree, the model results in:

✓ Low Bias

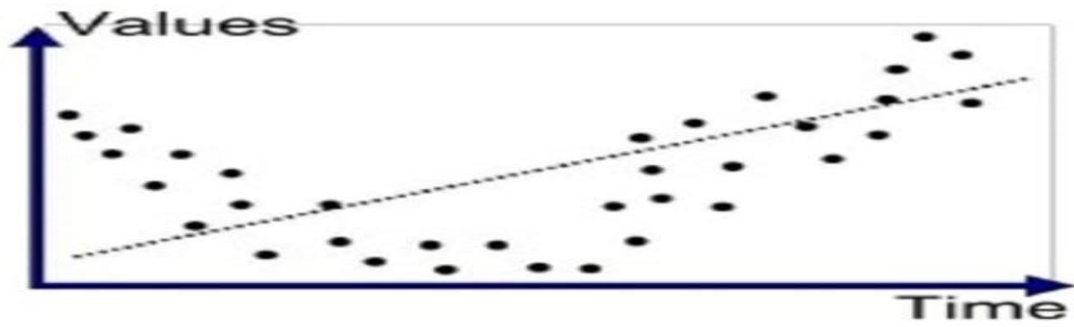
✓ High Variance

✓ Overfitting

Therefore the chosen model (Deep Decision Tree) leads to Overfitting.

6. Neat Diagram (Conceptual)

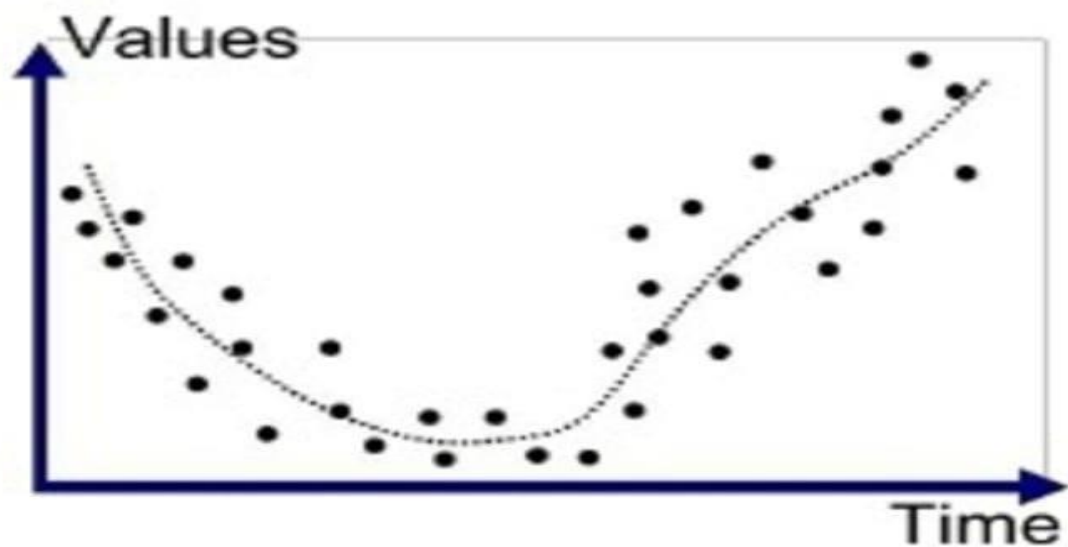
a). Underfitting (High Bias)



Underfitted

Model is too simple → High Bias

b).Overfitting (High Variance)

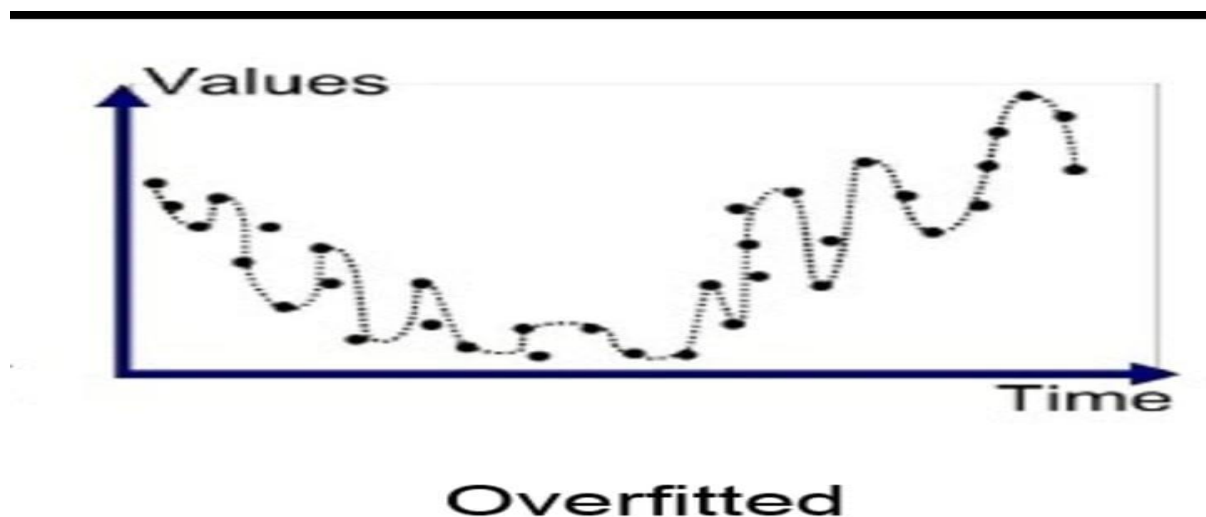


Good Fit/Robust

(Complex zig-zag boundary touching every point)

Model too complex → High Variance

c).Best Fit (Balanced)



Balanced model \rightarrow Low Bias + Low Variance

7. Conclusion

In the Diabetes Prediction domain:

- A small decision tree \rightarrow High Bias \rightarrow Underfitting
- A very deep decision tree \rightarrow High Variance \rightarrow Overfitting
- A properly pruned tree \rightarrow Low Bias + Low Variance \rightarrow Best Fit

Thus, understanding Bias and Variance helps us choose the correct model complexity and improve prediction accuracy.