

ONE PAGE SUMMARY

IT & Software > Machine Learning > Bias and Fairness

Introduction to Bias and Fairness

Understanding Bias and Fairness in Machine Learning

DEFINITION

Bias in machine learning refers to systematic errors that lead to unfair outcomes. Fairness ensures that algorithms treat all individuals equitably, without discrimination.

KEY CONCEPTS

Types of Bias

Bias can be introduced through data, algorithms, or human decisions, affecting outcomes.

Fairness Metrics

Metrics like demographic parity and equal opportunity help assess fairness in models.

Algorithmic Discrimination

When algorithms favor one group over another, it leads to unequal treatment.

Mitigation Strategies

Techniques like re-sampling, re-weighting, or modifying algorithms can reduce bias.

Ethical Considerations

Understanding the ethical implications of bias and fairness is crucial for responsible AI.

EXAMPLES

- Facial recognition systems misidentifying people of color more frequently than white individuals.
- Job recruitment algorithms favoring candidates from certain demographics over others.

MEMORY TIPS

- ★ Think 'B-Fair' to remember Bias and Fairness.
- ★ Use the acronym 'FAME' for Fairness, Assessment, Mitigation, Ethics.

COMMON MISTAKES

- ✗ Confusing bias with variance in model performance.
- ✗ Overlooking the importance of diverse training data.

QUICK RECAP

Bias and fairness are critical in machine learning to ensure equitable outcomes. Understanding types of bias, fairness metrics, and mitigation strategies helps create responsible AI systems.