

# Training, Validation, and Testing in Machine Learning

Mehmet Kerem Turkcan

July 17, 2024

# Outline

- 1 Introduction
- 2 Training Data
- 3 Validation Data
- 4 Testing Data
- 5 Importance
- 6 Conclusion

# Why Split Your Data?

- Properly evaluating model performance
- Avoiding overfitting
- Ensuring generalization to new, unseen data
- Making informed decisions about model selection and hyperparameters

# Training Data

- Largest portion of your dataset (typically 60-80%)
- Used to teach the model patterns and relationships
- Model learns to minimize loss on this data
- Caution: Can lead to overfitting if used alone

# Validation Data

- Smaller portion of your dataset (typically 10-20%)
- Used to tune hyperparameters and select models
- Provides an unbiased evaluation during development
- Helps prevent overfitting to the training data

# Testing Data

- Final portion of your dataset (typically 10-20%)
- Used only once to evaluate the final model
- Simulates real-world performance on unseen data
- Crucial for estimating true model generalization

# Why is This Split Important?

- Prevents overfitting: Model generalizes beyond training data
- Enables fair model comparison: Same test set for all models
- Simulates real-world scenarios: Test on truly unseen data
- Builds trust in model performance: Unbiased evaluation

# Common Pitfalls to Avoid

- Data leakage between sets
- Overfitting to the validation set
- Using test data more than once
- Ignoring dataset shift or temporal aspects

# Key Takeaways

- Proper data splitting is crucial for model development
- Training data teaches, validation data tunes, testing data evaluates
- This approach leads to more robust and reliable models
- Always keep test data separate and use it only once