Day 13 – Model Comparison & Feature Engineering

What Does Training Multiple Models Mean?

- *Training more than one ML algorithm on the same dataset*
- *Helps us understand which model performs best*
- *Different models learn patterns differently*

Common Models Used

- *Linear Regression*
- *Decision Tree*
- *Random Forest / Gradient Boosted Trees*
- *Logistic Regression (for classification)*

Why Train Multiple Models?

- *No single model is best for all problems*
- *Compare accuracy, error, and performance*
- *Choose the most reliable model for production*

Hyperparameter Tuning

- *Settings defined before training the model*
- *Control how the model learns*
- *Examples:*
  - *learning rate*
  - *max depth*
  - *number of trees*

Why Hyperparameter Tuning Is Important

- *Improves model performance*
- *Reduces underfitting and overfitting*
- *Helps the model generalize better on new data*

Hyperparameter Tuning in Spark

- *Use ParamGridBuilder*
- *Use CrossValidator or TrainValidationSplit*
- *Track results using MLflow*

What Is Feature Importance?

- *Measures how much each feature contributes to predictions*
- *Helps understand model behavior*
- *Improves trust and explainability*

Why Feature Importance Matters

- *Identify most impactful features*
- *Remove irrelevant or noisy columns*
- *Improve model efficiency and accuracy*

Feature Importance in Spark ML

- *Supported by tree-based models*
- *Available using featureImportances*
- *Useful for feature selection and insights*

What Is a Spark ML Pipeline?

- *A sequence of data processing and modeling steps*
- *Ensures consistency in training and prediction*
- *Similar to real-world ML workflows*

Pipeline Components

- *Data preprocessing (StringIndexer, VectorAssembler)*
- *Model training (Regression / Classification model)*
- *All steps combined into one pipeline*

Benefits of Spark ML Pipelines

- *Clean and reusable code*
- *Easy to maintain and scale*
- *Reduces manual errors in ML workflows*