Labs 9–12 Summary: Visualization, Spark MLlib, MLflow, and Deployment

Z5008 Big Data Lab · Even Semester 2026 · Dr. Innocent Nyalala

1. Distributed ML Pipelines and Experiment Tracking

`pyspark.ml.Pipeline` chains `StringIndexer`, `VectorAssembler`, `StandardScaler`, and estimators without leaking fold statistics. MLflow logs parameters, evaluation metrics (F1, ROC-AUC), and model artifacts across runs into a versioned Model Registry (`Staging` → `Production`).

2. Containerized Serving and CI/CD

Models package as stateless REST services inside Docker containers with pinned dependencies, automated smoke tests, and rolling Kubernetes updates.