Learning Log

Challenges Faced:
Imbalanced Dataset: Phishing emails were the minority, requiring SMOTE and class weights.

Feature Selection: Needed to experiment with metadata features like URLs, HTML tags, and capitalization.

Model Tuning: Finding the right hyperparameters for XGBoost was time-consuming.

Resources Consulted:
scikit-learn documentation

YouTube tutorials on phishing detection and XGBoost tuning

Blogs on Towards Data Science and Medium (phishing detection case studies)

NLTK and spaCy official docs for lemmatization and stopword removal

SHAP tutorials for understanding feature contributions

Key Takeaways:
Recall is critical in security-related ML problems where false negatives can cause real harm.

Combining NLP and metadata gives a more comprehensive view of potential phishing emails.

Model interpretability through SHAP values can uncover surprising indicators of phishing.

Automation via pipelines and modular code helped scale experiments smoothly.