Project Progress Summary

- **Data Preparation & Feature Selection:** We selected the CSE-CIC-IDS2018 dataset, performed feature selection using a RandomForest model to identify the **top 30 most impactful features**, and created a smaller, optimized filtered_dataset.csv to accelerate all subsequent work.

- **Supervised Modeling:** We developed and compared four models (e.g., RandomForest, LightGBM). We successfully resolved a **class imbalance** using a hybrid SMOTE and undersampling approach. After tuning hyperparameters to reduce the final model's file size, we selected **RandomForest** as the best-balanced classifier for its high accuracy and low false-positive rate.

- **Unsupervised Modeling:** We evaluated three unsupervised models for anomaly detection. After diagnosing initial poor performance. Need to better train it, since currently it has a lot of false positives.

Conclusion we came up with : Running both models in parallel combines a specialist (supervised model) for accurately identifying known attacks with a generalist watchdog (unsupervised model) for catching novel, unknown anomalies, ensuring comprehensive threat coverage.

Next Step

- Decrease False positives and increase Accuracy.
- Integrate somehow the supervised and unsupervised models to run parallelly.
- Integrate the saved, balanced RandomForest model into the LangChain agent framework to create a functional threat analysis tool.