Data Science Models

1. Linear/Probabilistic Model

Logistic Regression (LR): This model uses a linear equation to predict the probability of a given network flow belonging to a specific class (e.g., HTTP, DNS, P2P). It's a fundamental baseline model.

2. Ensemble Model

Random Forest (RF): This is an ensemble of Decision Trees. Each tree is trained on a random subset of the data and features. The final classification decision is made by taking a **majority vote** from all the individual trees.

3. Distance-Based Model

K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm. When a new, unlabeled packet/flow comes in, it classifies it by looking at the 'K' closest, already-labeled data points (neighbors) in the feature space and assigning the class that is most common among those neighbors.

4. Kernel-Based Model

Support Vector Machine (SVM): SVM works by finding an optimal hyperplane that maximizes the margin (the distance) between the data points of different classes in the feature space. For complex data, it uses the "kernel trick" to map data into higher dimensions to make separation easier.

5. Tree-Based Model

Gradient Boosting Classifier (GBC) or XGBoost: This is another powerful ensemble method, but unlike Random Forest, it builds trees sequentially. Each new tree attempts to correct the errors made by the previous trees, gradually improving the overall prediction accuracy.