## **NSL-KDD Intrusion Detection Steps**

Here's a summarized list of the steps involved in building an intrusion detection model using the NSL-KDD dataset (Kaggle):

**1. Data Acquisition and Preprocessing**

* Load the NSL-KDD dataset (training and testing sets).
* Define features, categorical features, and attack labels.
* Check data information (data types, missing values).
* Analyze descriptive statistics for numerical features.
* Explore class distribution (number of instances per attack type).

**2. Data Visualization**

* Visualize the distribution of attack types and services.
* Analyze the distribution of the 'duration' feature.
* Plot the frequency of flags in the training data.

**3. Feature Engineering**

* Create new categorical features from numerical features (e.g., binning 'duration').
* Analyze the relationship between protocols and attack classes.
* Analyze the relationship between services and attack classes.
* Create a cross-tabulation table to explore protocol usage per service.

**4. Feature Selection**

* Identify features with high correlations with the encoded intrusion label.
* Select features based on chi-squared test results (binary classification).
* Select the most informative numerical features using feature selection techniques.

**5. Data Preparation for Classification**

* Create binary and multi-class versions of the data (normal vs. abnormal).
* Encode categorical features using one-hot encoding.
* Prepare separate datasets for training and testing.
* Split data into features and target variables.

**6. Model Training and Evaluation**

* Train and evaluate different models like Extra Trees Classifier, Support Vector Classifiers (SVC), and K-Nearest Neighbors (KNN).
* Use metrics like accuracy, precision, recall, F1-score to assess model performance.

**7. Conclusion**

* KNN achieved the highest accuracy due to:
  + Clear separation between normal and abnormal classes.
  + Less susceptibility to the "curse of dimensionality."