NSL-KDD Dataset Preprocessing Documentation

1. Missing Values

- The dataset was inspected for missing values across all columns.
- No missing values were found in the NSL-KDD dataset.
- Predefined strategy (in case of missing values in future datasets):
  - For numerical features → Replace with median.
  - For categorical features → Replace with mode.

2. Duplicate Removal

- Duplicate rows were identified and removed.
- This step ensures unbiased model training and prevents overfitting due to repeated samples.
- Final dataset size after removal: **(rows × columns depending on data shape after processing).**

3. Encoding Categorical Features

- The dataset contained categorical features such as:
  - protocol_type
  - service
  - flag
- These features were converted into numerical form using **One-Hot Encoding**.
- Target labels outcome and category were transformed using **Label Encoding**.

4. Feature Scaling

- All numerical features were scaled using **MinMaxScaler** to normalize their values between [0,1].
- This scaling ensures fair contribution of features during model training, especially those with large ranges (e.g., src_bytes, dst_bytes).

5. Preprocessing Summary

The preprocessing steps applied were:

1. Checked for missing values.
2. Removed duplicate rows.
3. Encoded categorical variables.
4. Scaled numerical features.

The resulting dataset is clean, consistent, and ready for machine learning model training.