# BAB 3
# METODOLOGI PENELITIAN

## 3.1 Studi Literatur

Pada tahap ini, dilakukan penelusuran sumber-sumber informasi dan teori yang sudah ada untuk mendukung proses penelitian. Tahap ini sangat penting untuk memahami penelitian sebelumnya yang mencakup masalah yang sama. Teori-teori yang didapat dari studi literatur digunakan sebagai dasar untuk memperdalam pemahaman terhadap penelitian yang sedang berlangsung, dengan cara mencari artikel, jurnal, dan buku dari sumber *online* yang terpercaya.

## 3.2 Pengumpulan Dataset

Pada tahap ini dilakukan perolehan dataset *network traffic log* berjudul "*Malware Detection in Network Traffic Data*" [25] yang berisi label yang menjelaskan hubungan antara aliran yang terkait dengan aktivitas berbahaya atau kemungkinan berbahaya untuk memberikan informasi yang lebih mendalam kepada peneliti dan analis malware jaringan. Dataset ini didapatkan dari Kaggle.

## 3.3 *Preprocessing*

Pada tahap ini, dilakukan *preprocessing*. Tujuan dari *preprocessing* adalah agar dataset dapat disiapkan dan lebih efektif untuk digunakan dengan model SVM dan Naive Bayes. *Preprocessing* meliputi langkah-langkah seperti *data cleaning*, yaitu menghapus kolom-kolom fitur yang kurang relevan, *handling missing values*, dan *one-hot encoding*.

## 3.4 Pembagian Data

Dataset dibagi menjadi tiga. 70% digunakan untuk *training*, 15% untuk *validation* dan 15% terakhir digunakan untuk *testing*. Analisis empiris telah menunjukkan bahwa hasil terbaik dicapai jika kita mengalokasikan 20-30% dari data asli untuk *testing*, dan menggunakan sisa 70-80% untuk *training*. Pembagian ini umum digunakan dan juga membuat hasil akurasi model menjadi *valid*, dalam arti hasilnya tidak dilebih-lebihkan [26].

### 3.5 *Training* dan *Validation*

*Fit* model NB dan SVM ke dalam *train split* dan *val split* menggunakan *library* sklearn. Setelah itu dilakukan *report* klasifikasi yang menunjukkan skor *precision, recall, f1-score* dan akurasi.

### 3.6 *Testing dan Evaluasi*

Tahap ini adalah tahap di mana dataset yang sudah melalui *preprocessing* dan sudah di-*train* di-*test* dengan menggunakan NB dan SVM untuk mengetahui tingkat akurasinya. Evaluasi dilakukan dengan *confusion matrix* dan ROC (*Receiver Operating Characteristic*) AUC (*Area Under Curve*). Tahap ini sangat penting untuk menilai seberapa kuat model yang dipakai untuk menentukan keakuratan model dalam mendeteksi jaringan bahaya melalu *traffic log*.