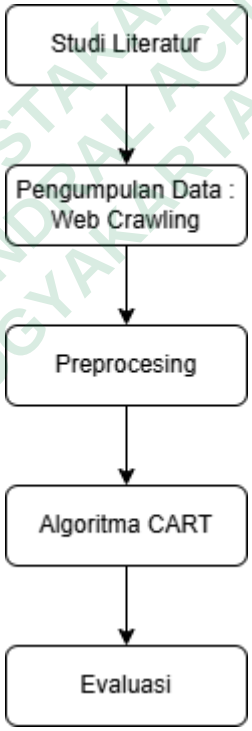
BAB 3
METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah klasifikasi dengan menggunakan algoritma CART (*Classification and Regression Tree*) untuk membangun pohon keputusan (*Decision Tree*) dalam mendeteksi situs *phising*. Model ini dikembangkan untuk menentukan tingkat akurasi dalam mengklasifikasikan situs *phising* dan *non-phising* berdasarkan fitur-fitur yang diekstrak dari dataset. Penelitian ini teridiri dari beberapa tahapan yang membentuk alur kerja sistematis yang dapat dilihat pada gambar 3.1 berikut.



Gambar 3.1 Alur Tahapan Sistematis

Berdasarkan alur penelitian tersebut berikut penjelasan pada tiap tahapan:

1. Studi Literatur
   a. Mengkaji penelitian terdahulu mengenai deteksi dan *klasifikasi webshite phising*.

b. Mengidentifikasi fitur-fitur yang umum digunakan dalam mendeteksi situs *phising*, seperti *URL, SSL/TLS*, metadata, dan faktor lainnya.

2. Pengumpulan Data.
   a. Menggunakan teknik crawling untuk mengumpulkan data dari berbagai situs web, baik yang aman maupun yang dicurigai sebagai *phising*.
   b. Sumber data berasal dari daftar situs *phising* yang dikumpulkan dari layanan *PhisTank.*
   c. Data yang dikumpulkan meliputi fitur *URL, HTML, domain,* dan faktor lainnya.
3. *Preprocessing*,
   a. Menghapus duplikasi dan *Handling Missing Values*, memeriksa dataset untuk menemukan nilai yang hilang (missing values) pada fitur-fitur tertentu.
   b. Encoding Data *Kategorikal*, Ubah data *kategorikal* (seperti "Ya/Tidak") menjadi numerik (misalnya, 0 dan 1).
   c. Normalisasi, normalisasi atau standarisasi pada fitur *numerik.*
   d. Pemisahan dataset menjadi fitur (X) dan label (Y), label 1 (*phising*) dan 0 (*legitimate*).
   e. Membagi dataset menjadi training (80%) dan testing (20%) untuk menghindari overfitting pada model.
4. Algoritma *CART*

   Berikut merupakan cara kerja algoritma *CART* untuk klasifikasi *phising:*
   a. Menghitung Gini impurity atau entropy untuk menentukan pemisahan fitur terbaik. Formula Gini Impurity dapat dilihat pada persamaan (6) berikut :

      $$Gini\ Index = 1 - \sum_{i=1}^{c} pi^2 \qquad (6)$$

      1. Jika Gini = 0, semua data dalam satu kelas (ideal)
      2. Jika Gini mendekati 0,5, data bercampur antara *phising* dan *legitimate.*
   b. Membangun pohon keputusan secara rekursif
   c. Melakukan pruning untuk menghindari overfitting

d. Pelatihan dan pengujian model

5. Evaluasi

Pada Tahapan ini model dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk mengukur klasifikasi situs *phising* dan non-*phising*. Berikut evaluasi dengan menggunakan beberapa metrik dalam klasifikasi, yaitu:

a. Akurasi *(Accuracy)*

Akurasi merupakan rasio jumlah prediksi yang benar terhadap total jumlah data yang diuji. Menunjukkan seberapa sering model membuat prediksi yang tepat. Formula penghitungan prediksi akurasi dapat dilihat pada persamaan (7), berikut:

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Jumlah\ Total\ Data} \quad (7)$$

b. Presisi *(Precision)*

Presisi mengukur seberapa banyak dari semua situs yang diprediksi sebagai *phising* oleh model, benar-benar merupakan situs *phising*. Formula pengukuran prediksi presisi dapat dilihat pada persamaan (8), berikut:

$$Precision = \frac{True\ Positive(TP)}{True\ Positive\ (TP) + False\ Positive(FP)} \quad (8)$$

Nilai presisi yang tinggi berarto model jarang mengklasifikasikan situs yang bukan *phising* sebagai *phising*.

c. Recall (Sensitivity / True Positive Rate)

Recall mengukur seberapa banyak dari seluruh situs *phising* yang benar berhasil dikenali oleh model. Formula pengukuran dengan recall dapat dilihat pada persamaan (9), berikut:

$$Recall = \frac{True\ Posotive(TP)}{True\ Positive(TP) + False\ Negative(FN)} \quad (9)$$

Nilai recall yang tinggi berarti model mampu menangkap sebagian besar situs *phising* yang ada.

d. *F1-Score*

F1-score adalah harmonisasi antara precision dan recall. Digunakan ketika ingin menyeimbangkan keduanya, terutama jika terdapat

ketidakseimbangan data. Formula dengan *F1-score* dapat dilihat pada persamaan (10), berikut:

$$F1 - Score = 2x \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

F1-score mendekati 1 berarti model memiliki keseimbangan yang baik antara presisi dan recall

3.1 Bahan dan Alat Penelitian

Dalam penelitian ini, dataset yang digunakan berasal dari *PhisTank*. *PhisTank* adalah layanan berbasis *web* yang menyediakan daftar situs *phising* yang telah didentifikasi dan diverifikasi secara berkala. Platform ini digunakan oleh peneliti keamanan siber, perusahaan, dan pengembang sistem keamanan untuk mendeteksi serta mencegah serangan *phising*. *PhisTank* secara otomatis mengumpulkan, menganalisis, dan memperbarui daftar situs *phising* dengan menggunakan algoritma deteksi berbasis data. Keunggulan *PhisTank* adalah kemampuannya dalam menyediakan data *phising* yang selalu diperbarui, sehingga pengguna dapat memperoleh informasi terbaru mengenai tren serangan *phising*. Dengan menggunakan *PhisTank*, penelitian di bidang keamanan siber dapat lebih akurat dalam mengembangkan sistem deteksi *phising* berbasis penambangan data dan *web crawling*.

Untuk memastikan data dari *PhisTank* disimpan dan diproses dengan aman, terutama jika melibatkan data sensitif, diperlukan beberapa langkah keamanan seperti, data harus dienkripsi baik saat disimpan maupun selama transmisi menggunakan protokol seperti AES-256 dan TLS untuk mencegah akses tidak sah. Manajemen akses yang ketat juga diperlukan, termasuk penerapan prinsip least privilege, autentikasi multi-faktor (MFA), dan audit akses berkala. Penyimpanan data, baik di cloud maupun on-premise, harus memenuhi standar keamanan tinggi seperti sertifikasi ISO 27001 atau SOC 2, serta dilengkapi dengan firewall dan sistem deteksi intrusi. Selama pemrosesan, data sensitif sebaiknya diisolasi, di-anonimisasi, atau di-pseudonimisasi untuk mengurangi risiko kebocoran. Selain itu, backup rutin dan pemulihan data harus disiapkan untuk memastikan data dapat dipulihkan jika terjadi insiden. Kepatuhan terhadap regulasi privasi data seperti

GDPR atau CCPA juga wajib dipenuhi, termasuk pelaporan insiden keamanan jika diperlukan.

Fitur yang dikumpulkan dalam penelitian ini meliputi fitur URL (panjang URL, jumlah subdomain, karakter spesial dalam URL, penggunaan HTTPS atau HTTP), fitur domain (usia domain, informasi WHOIS, lokasi server, jumlah redirect), serta fitur HTML & JavaScript (penggunaan iframe, hidden input fields, script mencurigakan, keberadaan formulir login).

Alat yang digunakan dalam penelitian ini adalah komputer dengan spesifikasi cukup untuk menjalankan sistem operasi dan perangkat lunak pengembangan serta koneksitas Internet. Sistem Operasi dan program-program aplikasi yang dipergunakan dalam dalam penelitian ini adalah:

1. Sistem Operasi: Windows 10
2. Processor: Intel Core i5-4200U CPU 2.29 Ghz
3. RAM: 8 GB
4. Penyimpanan: HDD 500 GB
5. Software: Python, GoggleColab.

3.2 Jalan Penelitian

Jalan penelitian mencakup penjelasan lengkap dan mendalam tentang proses yang diambil selama pelaksanaan penelitian dan pengumpulan data. Pada penelitian ini terdapat beberapa tahapan yang dilakukan, penelitian ini terdiri dari 5 tahap, yaitu:

1. Tahap Studi Literatur

   Kajian terhadap penelitian terdahulu yang berkaitan dengan deteksi dan klasifikasi situs *phising*. Fokus utama adalah mengidentifikasi fitur-fitur umum yang digunakan dalam pendeteksian, seperti karakteristik URL, penggunaan SSL/TLS, metadata.

2. Tahap Pengumpulan Data.

   Data dikumpulkan dengan teknik *web crawling* dari berbagai situs web, baik yang aman maupun yang dicurigai sebagai *phising*. Sumber utama berasal

dari layanan *OpenPhish*, yang menyediakan daftar situs *phising* terkini dan telah diverifikasi

3. Tahap *Preprocessing*,

   Pre-processing merupakan tahapan yang bertujuan untuk menyeleksi data dari *missing values* sehingga mendapatkan data yang bersih dan siap digunakan.

4. Tahap Implementasi Algoritma *CART*

   Model klasifikasi dibangun menggunakan algoritma CART (*Classification and Regression Tree*) untuk mendeteksi apakah suatu situs web termasuk *phising* atau bukan. Algoritma CART membentuk pohon keputusan berdasarkan fitur yang telah dipilih, dengan memanfaatk an teknik pemisahan berbasis Gini *Index* atau *Entropy* untuk menentukan node terbaik dalam pohon keputusan.

5. Tahap Evaluasi

   Pada Tahapan ini model dievaluasi menggunakan metrik akurasi,*precision*, *recall*, dan *F1-score* untuk mengukur seberapa baik model dalam mengklasifikasikan situs *phising* dan non-*phising*.