BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini bertujuan untuk membangun sebuah website yang memiliki kemampuan mendeteksi URL phishing yang berpotensi menyerang PT Bank Central Asia, Tbk . Dalam pengembangan ini, digunakan data URL phishing yang didapatkan langsung dari PT Bank Central Asia, Tbk , sehingga sistem deteksi dapat dioptimalkan dengan data yang relevan dan mencerminkan ancaman yang spesifik terhadap institusi tersebut. Dengan begitu, website ini tidak hanya akan memiliki kecerdasan dalam mendeteksi URL phishing secara umum tetapi juga dirancang untuk mendeteksi pola serangan yang mungkin mengincar PT Bank Central Asia, Tbk secara langsung.

Agar mencapai tingkat akurasi yang tinggi, penelitian ini juga akan memanfaatkan algoritma machine learning dengan performa terbaik, yang telah terbukti unggul dalam klasifikasi phishing URL pada berbagai studi terdahulu. Melalui pemilihan algoritma yang optimal, seperti SVM atau Random Forest yang sering menghasilkan akurasi tinggi dalam deteksi phishing, sistem diharapkan mampu mengidentifikasi URL phishing secara cepat dan akurat. Dengan menggabungkan data aktual dari PT Bank Central Asia, Tbk dan teknologi machine learning yang teruji, website ini diharapkan menjadi alat yang andal dan efektif untuk mendeteksi dan melawan ancaman phishing yang bisa merugikan perusahaan serta para penggunanya.

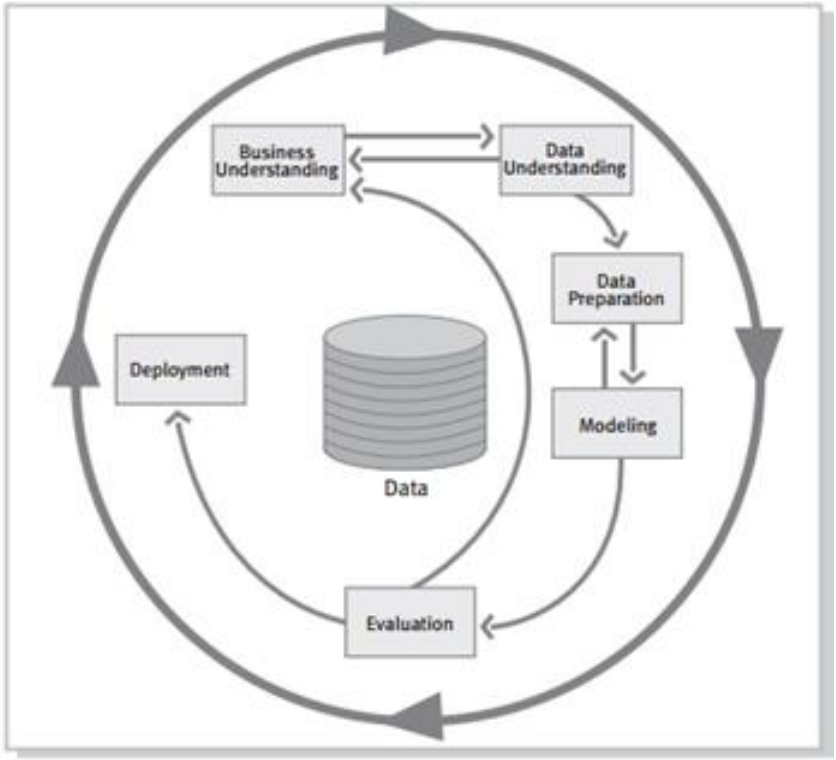
3.2 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif karena fokus utamanya adalah mengukur dan membandingkan kinerja algoritma machine learning berdasarkan metrik yang terukur secara objektif, seperti akurasi, precision, dan

recall. Data yang digunakan dianalisis secara statistik untuk menentukan model dengan performa terbaik dalam mendeteksi URL phishing [23]. Pendekatan kuantitatif dipilih untuk menghasilkan kesimpulan berbasis angka yang dapat diandalkan, relevan, dan dapat direplikasi, tanpa melibatkan eksplorasi aspek subjektif atau naratif.

3.2.1 Alur Penelitian

Berikut adalah penjelasan gambar alur penelitian CRISP-DM (Cross-Industry Standard Process for Data Mining) yang relevan dengan topik pembuatan model untuk mendeteksi URL phishing di perusahaan PT Bank Central Asia, Tbk :



Gambar 3. 1 Alur Penelitian [23]

3.2.1.1 Pemahaman Bisnis

Tahap ini merupakan langkah awal yang sangat penting dalam proses pengembangan sistem deteksi phishing. Pada tahap ini, fokus utamanya adalah memahami masalah bisnis yang ingin diselesaikan dan bagaimana sistem yang akan dikembangkan dapat memberikan solusi yang relevan. Dalam penelitian ini, PT Bank Central Asia, Tbk (BCA) diidentifikasi memiliki kebutuhan untuk mengembangkan sistem yang mampu mendeteksi URL phishing secara efektif. Hal ini menjadi penting karena phishing merupakan salah satu bentuk kejahatan siber yang sering menyasar lembaga keuangan seperti bank. Phishing dapat menyebabkan kerugian besar, baik dari segi finansial maupun reputasi perusahaan. Oleh karena itu, sistem deteksi phishing diharapkan dapat memberikan perlindungan tambahan kepada nasabah dengan meminimalkan risiko akses ke situs palsu yang mencoba mencuri informasi sensitif. Tujuan utama dari penelitian ini adalah untuk menghasilkan model yang akurat, andal, dan dapat diintegrasikan dengan mudah ke dalam sistem PT Bank Central Asia, Tbk , sehingga dapat memberikan nilai tambah dalam operasional mereka [22].

3.2.1.2 Data Understanding

Pada tahap ini, data menjadi fokus utama untuk dianalisis lebih mendalam. Data yang digunakan adalah kumpulan URL phishing yang diperoleh dari PT Bank Central Asia, Tbk . Analisis awal dilakukan untuk memahami pola, distribusi, dan kualitas data tersebut. Pemahaman ini mencakup identifikasi karakteristik URL phishing, seperti panjang URL, penggunaan simbol atau karakter khusus, serta pola domain yang mencurigakan. Selain itu, data juga dibandingkan dengan URL yang sah (legitimate) untuk menemukan perbedaan signifikan yang dapat digunakan sebagai dasar pengembangan fitur dalam model. Analisis ini penting untuk memastikan bahwa data

memiliki representasi yang cukup untuk menggambarkan berbagai skenario phishing. Jika terdapat ketidakseimbangan kelas dalam data, strategi seperti oversampling atau undersampling akan direncanakan untuk mengatasi masalah tersebut. [23].

3.2.1.3 Data preparation

Tahap ini mencakup berbagai langkah untuk membersihkan dan mempersiapkan data agar siap digunakan dalam proses pemodelan. Data yang mentah sering kali memiliki banyak tantangan, seperti duplikasi, informasi yang hilang, atau data yang tidak relevan. Oleh karena itu, proses pembersihan dilakukan untuk memastikan kualitas data. Setelah itu, fitur diekstraksi dari URL, misalnya panjang URL, jumlah tanda garis miring ("/"), atau keberadaan IP address dalam domain. Fitur-fitur ini diharapkan mampu merepresentasikan pola-pola unik yang terdapat pada URL phishing. Data kemudian dibagi menjadi data pelatihan dan data pengujian dengan proporsi yang sesuai, sehingga model dapat dilatih dan diuji secara terpisah untuk menghindari overfitting [23].

3.2.1.4 Modeling

Tahap ini merupakan inti dari penelitian, di mana berbagai algoritma machine learning diuji untuk membangun model yang optimal. Algoritma yang digunakan dalam penelitian ini meliputi Support Vector Machine (SVM), Random Forest, Naïve Bayes, dan Decision Tree. Setiap algoritma memiliki karakteristik dan kekuatan masing-masing. Misalnya, SVM dikenal memiliki performa yang baik dalam memisahkan data non-linear, sementara Random Forest unggul dalam menangani data yang kompleks dengan fitur-fitur yang saling berinteraksi. Selain memilih algoritma, dilakukan juga pengujian terhadap berbagai parameter (hyperparameter tuning) untuk

meningkatkan performa model. Hasil pemodelan akan dibandingkan berdasarkan metrik evaluasi untuk memilih model terbaik..

3.2.1.5 Evaluation

Model yang telah dikembangkan kemudian dievaluasi menggunakan metrik-metrik kinerja seperti akurasi, precision, recall, dan F1-score. Metrik-metrik ini digunakan untuk mengukur seberapa baik model dalam mendeteksi URL phishing dibandingkan dengan URL yang sah. Selain itu, evaluasi juga dilakukan untuk memastikan bahwa model memenuhi tujuan bisnis dan teknis. Jika hasil evaluasi menunjukkan bahwa model belum memenuhi standar yang diharapkan, maka akan dilakukan iterasi ulang, baik pada tahap pemilihan fitur, pemilihan algoritma, maupun tuning parameter. Evaluasi yang komprehensif penting agar model dapat diandalkan ketika diimplementasikan di lingkungan nyata..

3.2.1.6 Deployment

Tahap terakhir adalah menerapkan model yang telah dipilih ke dalam sistem yang dapat diakses oleh pengguna akhir. Dalam penelitian ini, model diimplementasikan dalam sistem berbasis web menggunakan framework Flask. Framework ini memungkinkan integrasi model machine learning dengan antarmuka pengguna yang sederhana dan responsif. Sistem yang dihasilkan akan memiliki fitur untuk mendeteksi URL phishing dengan mudah, di mana pengguna cukup memasukkan URL yang ingin diperiksa dan sistem akan memberikan hasil deteksinya.

Dengan adanya sistem ini, PT Bank Central Asia, Tbk dapat memberikan perlindungan tambahan kepada nasabahnya dan meningkatkan kepercayaan mereka terhadap keamanan layanan yang disediakan. Tahap deployment ini juga mencakup pemantauan performa model secara berkelanjutan untuk memastikan bahwa model tetap akurat seiring waktu..

3.2.2 Metode Data Mining

Data mining adalah metode analisis data yang bertujuan untuk mengekstrak informasi berharga dari kumpulan data yang besar dan kompleks. Proses ini melibatkan teknik-teknik seperti klasifikasi, klastering, asosiasi, regresi, dan deteksi anomali untuk menemukan pola, hubungan, atau tren yang tersembunyi di dalam data. Data mining umumnya terdiri dari beberapa tahapan, termasuk pembersihan data (data cleaning), transformasi data, serta proses analisis dengan algoritma tertentu. Metode ini diterapkan di berbagai bidang, seperti pemasaran, keuangan, kesehatan, dan keamanan siber, untuk membantu pengambilan keputusan berdasarkan data yang lebih tepat dan informatif. Selain itu, data mining sering kali melibatkan penggunaan teknologi machine learning dan statistik untuk meningkatkan akurasi dan efisiensi dalam pemrosesan data besar, sehingga memungkinkan organisasi untuk memperoleh wawasan yang lebih dalam dan memprediksi kejadian di masa mendatang. CRISP-DM, SEMMA, dan KDD adalah tiga kerangka kerja utama yang sering digunakan dalam proses data mining untuk mengarahkan peneliti atau analis dalam mengolah data menjadi informasi yang berguna.

Table 3. 1 Perbandingan Framework data mining

Aspek	CRISP-DM	KDD	SEMMA
Dokumentasi / Kemudahan Penggunaan [54]	- Dokumentasi lengkap dan komprehensif. - Dikenal sebagai standar industri. - Mudah diikuti dan diterapkan.	- Dokumentasi terbatas dibanding CRISP-DM. - Proses lebih akademik. - Kurang detail dalam fase implementasi.	- Didukung oleh SAS Institute. -Terdokumentasi dengan baik dalam konteks SAS. - Lebih mudah bagi pengguna SAS.
Kelebihan	- Fleksibel dan dapat diterapkan dalam berbagai proyek data mining. - Berfokus pada pemahaman bisnis dan data. - Memiliki fase evaluasi yang kuat.	- Mendorong pendekatan eksploratif dalam analisis data. - Mencakup tahap pra-pemrosesan dan transformasi data yang jelas.	- Struktur yang jelas dan terorganisir dengan baik. -Memfasilitasi pemodelan statistik dengan baik. -Dapat diintegrasikan langsung dengan alat SAS.
Kekurangan	- Bisa terlalu umum bagi beberapa proyek spesifik. -Membutuhkan pemahaman mendalam untuk implementasi yang efektif.	-Kurang didokumentasikan untuk penggunaan industri. - Fokus lebih pada aspek teknis daripada bisnis .	-Sangat tergantung pada perangkat SAS. - Kurang fleksibel dibandingkan dengan CRISP-DM. - Kurang detail dalam fase pemahaman bisnis.

Penelitian ini menggunakan framework data mining CRISP-DM untuk pembuatan model klasifikasi URL phishing dan non-phishing pada PT Bank Central Asia, Tbk . Pemilihan framework CRISP-DM dilakukan karena mencakup seluruh tahap dari pemahaman bisnis hingga deployment. Tujuan akhir dari model yang akan dibuat adalah implementasi dalam bentuk website, sehingga framework CRISP-DM dianggap paling sesuai. Sebaliknya, framework KDD dan SEMMA hanya fokus pada proses pengolahan data tanpa

mencakup tahap deployment, sehingga kurang cocok untuk kebutuhan penelitian ini.

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan metode pengumpulan data melalui sistem pada perusahaan PT Bank Central Asia, Tbk yang mendeteksi url phising, Hal ini bertujuan untuk mendapatkan hasil yang akurat terhadap url phishing yang mengatasnamakan PT Bank Central Asia, Tbk . Proses pengumpulan data dilaksanakan selama periode satu tahun, dimulai dari tanggal 1 Januari 2023 hingga 1 Juli 2024. Jumlah URL Phishing yang digunakan dalam penelitian ini sebanyak 4078 URL.

3.4 Teknik Analisis Data

Penelitian ini akan berfokus dari mendeteksi URL Phishing, Python yang merupakan bahasa pemograman yang akan digunakan dan juga akan menggunakan beberapa model klasifikasi machine learning. Tekniknya adalah SVM, yang dibandingkan dengan tiga teknik machine learning lainnya, yaitu Random Forest,Decision Tree, dan Naive bayes untuk mendapatkan hasil yang optimal dan yang nantinya akan digunakan dalam implementasi website.