BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian ini berfokus pada deteksi *phishing* pada *URL website* dengan merangkum identifikasi dan klasifikasi *URL* yang berpotensi sebagai situs *phishing*. Proses analisis melibatkan penelitian mendalam terhadap berbagai aspek *URL*, seperti struktur domain, konten halaman *web*, dan karakteristik khusus yang dapat menjadi tanda serangan *phishing*. Penelitian ini bertujuan untuk mengembangkan metode atau model yang dapat secara otomatis dapat membedakan *URL* yang sah dari yang berbahaya, sehingga dapat meminimalkan risiko pengguna terkena serangan *phishing* dan meningkatkan keamanan saat *browsing web.*

Pengidentifikasian pola-pola mencurigakan dalam *URL* menjadi fokus utama, termasuk fitur-fitur yang sering terkait dengan situs *phishing*, seperti *URL* menyesatkan, penggunaan domain palsu, dan karakteristik tertentu dalam kode halaman *web*. Dengan demikian, penelitian ini diarahkan untuk mengembangkan metode deteksi yang efektif dan efisien, yang pada akhirnya akan memberikan perlindungan yang lebih baik kepada pengguna internet dan memperkuat keseluruhan sistem perlindungan keamanan web.

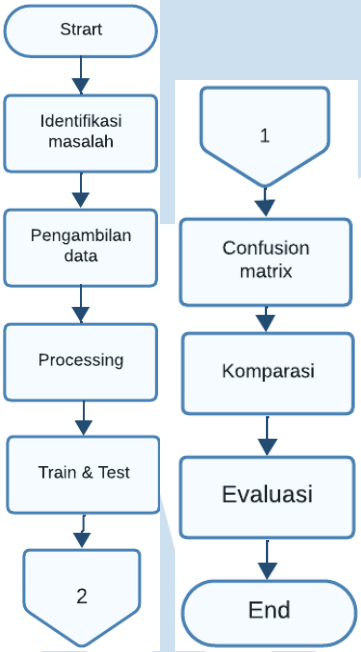
3.2 Metode Penelitian

3.2.1 Alur Penelitian

Alur penelitian adalah rencana kerja yang disusun untuk mengarahkan jalannya penelitian dan memastikan pelaksanaannya terstruktur disebut sebagai alur penelitian. Alur ini bertindak sebagai panduan yang mengatur proses penelitian. Berdasarkan gambar 3.1, penelitian ini dimulai dengan mengidentifikasi permasalahan yang ada dan semakin meningkat presentasenya di Indonesia yaitu mengenai *phishing* pada *url website*, oleh karena itu peningkatan keamanan *url website* menjadi tujuan utama dalam penelitian ini. Setelah mengidentifikasi isu yang

ada, langkah selanjutnya adalah mencari data phising melalui situs dataset resmi yaitu Kaggle.

Berdasarkan *dataset* yang didapatkan, metode untuk menyelesaikan masalah ditetapkan dengan menggunakan algoritma *Random Forest*, *Support Machine Vectore*, dan *K-Nearest Neighbor* untuk menentukan hasil akurasi terakurat. Penetapan metode atau algoritma ini didasarkan pada studi literatur yang relevan dan disesuaikan dengan kebutuhan dan karakteristik data yang ada. Setelah menentukan pendekatan, proses pengolahan data dimulai dengan langkah-langkah yang diatur berdasarkan kerangka kerja *CRISP-DM*. Proses penelitian ini sampai pada tahap evaluasi yang nantinya dapat dilanjutkan oleh peneliti lain menjadi *website* maupun *mobile app.*



Gambar 3. 1 Flowchart Alur Penelitian

3.2.2 Metode *Data Mining*

Metodologi yang digunakan dalam penelitian ini mengadopsi model *data mining* dengan pendekatan metode klasifikasi. Dalam pelaksanaannya, penelitian ini menerapkan kerangka kerja *CRISP-DM* yang dikenal sebagai kerangka kerja pengembangan *Knowledge Discovery in Databases (KDD)*. Untuk melakukan analisis klasifikasi penelitian ini

menggunakan *platform* Google Colab sebagai alat utama. *CRISP-DM* dipilih sebagai metode karena pertimbangan terhadap metode *data mining* lainnya.

Tabel 3. 1 Perbandingan CRISP-DM dan KDD

Indikator	***CRISP-DM**[38]*	***Knowledge Discovery in Databases (KDD)**[38]*
Fase	Enam tahapan yang dimulai dari pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, hingga implementasi.	Lima tahapan dimulai dari pemilihan, pra-pemrosesan, transformasi, penambangan data, hingga interpretasi/evaluasi.
Kekuatan	1. Terdapat langkah akhir yang melibatkan tahap penerapan atau implementasi. 2. Proses yang terperinci dalam tahapannya. 3. Sesuai untuk proyek atau dataset yang besar. 4. Siklus berulang yang berkelanjutan. 5. Dilengkapi dengan dokumentasi dan contoh studi kasus yang serupa yang dapat mendukung jalannya proses.	1. Tidak terdapat proses akhir yang melibatkan tahap *deployment* (penerapan) namun di akhiri dengan tahap evaluasi. 2. Siklus pengulangan yang berkelanjutan. *3.* Mendukung beragam model data mining, termasuk *Neural Network.* 4. Memerlukan pemahaman sebelumnya dalam *data mining*.
Limitasi	1. Dimulai dengan memahami tujuan proyek melalui tahapan pemahaman bisnis. 2. Proses yang memakan waktu. 3. Tahap persiapan data dan pembuatan model berbeda dari *data mining* konvensional sehingga tidak termasuk dalam dokumen *CRISP-DM.*	1. Dimulai dengan tahap pemilihan data yang akan membentuk *dataset* yang akan digunakan dalam penelitian. 2. Memerlukan pemahaman sebelumnya dalam *data mining.*

Berdasarkan perbandingan dalam table diatas, penelitian ini memilih metode *CRISP-DM* karena prosesnya memiliki tahapan yang terstruktur dan jelas, yang dapat memberikan arahan yang kuat dalam jalannya penelitian. Pemilihan ini juga didasarkan pada penggunaan dataset yang

besar, di mana *CRISP-DM* memiliki panduan yang sesuai untuk melakukan proses data mining pada dataset yang besar, sehingga metode ini cocok untuk diadopsi dalam penelitian ini.

Gambar 3. 2 Flowchart Proses Pengolahan Data

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan *dataset* yang diambil dari Kaggle.com yang berasal dari Eswar Chand. Kaggle adalah *platform* daring yang menyediakan berbagai kompetisi data, sumber daya pembelajaran, dan kumpulan *dataset* untuk para ilmuwan data dan praktisi *machine learning*. *Dataset* yang digunakan merupakan data mengenai *phishing website detector*. Kaggle adalah *platform* daring yang menyediakan berbagai kompetisi data, sumber daya pembelajaran, dan kumpulan *dataset* untuk para ilmuwan data dan praktisi *machine learning*.

3.4 Variabel Penelitian

3.4.1 Variabel Independen

Variabel bebas, yang juga dikenal sebagai variabel independen, merupakan variabel yang diyakini menjadi penyebab atau faktor yang memiliki kemungkinan berdampak pada variabel lain dalam suatu penelitian atau eksperimen. Biasanya, variabel bebas ditunjukkan dengan simbol X. Secara umum, variabel bebas hadir terlebih dahulu dan dianggap mempengaruhi variabel lain yang terkait. Variable independen pada penelitian ini yaitu *Using IP, Long URL, Short URL, Symbol@, Redirecting, Prefix Suffix, Sub Domains, HTTPS, Domain Reg Len, Favicon, Non Std Port, HTTPS Domain URL, Request URL, Anchor URL, Links In Script Tags, Server From Handler, Info Email, Abnormal URL, Website Forwarding, Status Bar Cust, Disable Right Click, Using Popup Window, Iframe Redirection, Age of Domain, DNS Recording, Website Traffic, Page Rank, Google Index, Links Pointing to Page, Stats Report.*

3.4.2 Variabel Dependen

Variabel dependen adalah variabel yang terpengaruh atau variabel yang merupakan akibat dari variabel independen. Variabel dependen pada penelitian ini yaitu menderita mental disorder. Variabel tergantung atau variabel dependen adalah variabel yang dipengaruhi atau menjadi hasil dari variabel bebas. Dengan demikian, variabel ini bergantung pada variabel bebas dan besarnya dipengaruhi oleh perubahan dalam variabel independen. Koefisien perubahan dalam variabel independen memberikan gambaran seberapa besar perubahan yang mungkin terjadi dalam variabel dependen sebagai akibat dari perubahan dalam variabel independen. Variable dependen pada penelitian ini adalah class.

3.5 Teknik Analisis Data

3.5.1 *Business Understanding*

Pada tahap pertama, penelitian ini melakukan analisa terhadap kasus *phishing* di Indonesia serta meneliti data yang akan digunakan untuk

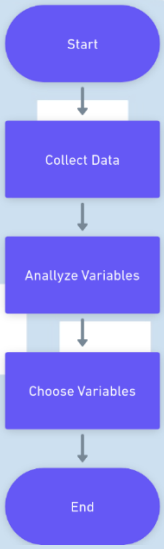
melihat kesesuaian data dengan penelitian. *Business Understanding* dalam konteks deteksi *URL web phishing* mencakup pemahaman yang mendalam tentang ancaman keamanan yang dihadapi oleh pengguna *web*. Hal ini mencangkup pemahaman terhadap kebutuhan organisasi atau *platform* terkait keamanan *online*, serta dampak yang mungkin timbul akibat serangan *phishing*. Tujuannya adalah untuk mengidentifikasi aspek-aspek utama yang terkait dengan upaya pencegahan, deteksi, dan perlindungan terhadap serangan *phishing* pada *URL web*. Pada penelitian ini, tujuan utamanya adalah mendeteksi *URL website phishing* dengan mengadopsi metode *klasifikasi Random Forest, K-Nearest Neighbors,* dan *Support Vector Machine.*

Tabel 3. 2 Penjelasan Atribut

No	Atribut	Keterangan
1	*Using IP*	Alamat IP sebagai domain.
2	*Long URL*	Panjangnya *url.*
3	*Short URL*	Pendeknya *url.*
4	*Symbol@*	Penggunaan symbol @.
5	*Redirecting*	Banyaknya pengalihan *website* yang dilakukan.
6	*Prefix Suffix*	Presentase penggunaan symbol "-".
7	*Sub Domains*	Presentase penggunaan sub domain.
8	*HTTPS*	Alamat *https.*
9	*Domain Reg Len*	Batas berlakunya domain.
10	*Favicon*	Mempunyai *favicon* dari *link* eksternal.
11	*Non Std Port*	Tidak terdapat penggunaan *port.*
12	*HTTPS Domain URL*	Pemakaian *https* ke dalam bagaian domain pada *url.*
13	*Request URL*	Presentase permintaan *url* eksternal dari keseluruhan (polinomial).
14	*Anchor URL*	Presentase penggunaan tag yang mengarah selain ke domain yang sama dari keseluruhan.

15	*Links In Script Tags*	Presentase pemakaian tag <*link*>. <*script*>, dan <meta> yang merujuk selain ke domain.
16	*Server From Handler*	Domain pemrosesan *server form handler.*
17	*Info Email*	Pemakaian fungsi “mail” dalam php.
18	*Abnormal URL*	*Url* yang tidak normal, kecocokan dengan web yang dituju.
19	*Website Forwarding*	Presentase *forwarding url.*
20	*Status Bar Cust*	Perubahan status bar ketika *mouseover* aktif.
21	*Disable Right Click*	Kondisi klik kanan pada *web.*
22	*Using Popup Window*	Presentase *popup window* dalam meminta *user* mengisi data.
23	*Iframe Redirection*	Penggunaan fungsi *iframe.*
24	*Age of Domain*	Umur domain.
25	*DNS Recording*	Terdapat catatan DNS pada domain.
26	*Website Traffic*	Presentase lalu lintas *web* dalam basis data Alexa.
27	*Page Rank*	Nilai *page rank website.*
28	*Google Index*	*Website* dalam indeks pencarian *google.*
29	*Links Pointing to Page*	Banyaknya link eksternal yang mengarah ke *website.*
30	*Stats Report*	Top *phishing* yang dibuat oleh beberapa pihak.
31	*Class*	Kelompok atribut.
32	*Id*	No atribut.

3.5.2 *Data Understanding*



Gambar 3. 3 Flowchart Data Understanding

Pada tahap kedua, dilakukan pengambilan data dari *platform* Kaggle. Proses ini dimulai dengan otomatisasi pada *platform* untuk mendapatkan semua tautan *listing* dan melakukan proses *scraping*. Kaggle dipilih sebagai sumber data karena *dataset* di Kaggle telah melalui proses kurasi dan pemrosesan awal. Ini berarti data-data tersebut telah diolah, divalidasi, dan disaring dari berbagai sumber untuk memastikan kualitas dan integritasnya. Prosesini menghasilkan data dengan 34 atribut yang diperoleh. Total data yang terdapat dalam dataset ini yaitu 11.000 lebih data. Atribut *dataset* tersebut terdiri dari 32 atribut sebagai berikut:

3.5.3 *Data Preparation*

Gambar 3. 4 Flowchart Data Preparation

Pada tahap ketiga, merupakan tahap *cleaning* dan *processing* data. Tahap ini dilakukan pengecekan *missing value* yang bertujuan untuk mengidentifikasi apakah terdapat nilai yang hilang atau tidak lengkap dalam *dataset* tersebut sehingga selanjutnya dapat dilakukan pengecekkan deskripsi data, apakah data tersebut berupa numerik atau kategorikal, dan untuk menampilkan atribut apa saja yang terdapat dalam *dataset* tersebut. Dari hasil data *preparation* didapatkan 32 atribut dengan 11.000 data. Pada penelitian ini membagi data *preparation* menjadi dua bagian yaitu *data cleansing*, dan *split data*, 1) Saat melakukan pembersihan data, atribut yang tidak diperlukan akan dihapus, dan jika terdapat *missing value* pada salah satu atribut, baris tersebut akan dihapus secara keseluruhan, Dengan melakukan proses *cleansing*, kualitas data yang digunakan dalam analisis meningkat, menghasilkan informasi yang lebih akurat dan berkualitas. Proses ini memainkan peran penting dalam tahap *data mining* karena kualitas analisis bergantung pada kualitas data yang digunakan, 2) Proses pemisahan data akan membagi data menjadi dua bagian, yakni data *training* dan data *testing* dalam perbandingan 80:20, dengan 80% untuk *training* dan 20% untuk *testing*. Setelah menyelesaikan tahap persiapan data, langkah selanjutnya adalah pemodelan data.

3.5.4 *Modeling*

Pada tahap keempat ini dilakukan pemodelan, yaitu pemilihan model dan algoritma yang akan digunakan, selanjutnya akan dilakukan penerapan algoritma yang telah ditentukan. Pada penelitian ini menggunakan pemodelan data berupa klasifikasi dengan menampilkan nilai akurasi, presisi, *recall*, dan f1 *score*. Pada langkah ini, model data yang digunakan disesuaikan dengan kebutuhan untuk mencapai hasil yang diinginkan. Pada penelitian ini dilipih 3 algoritma dengan metode klasifikasi, yaitu terdapat algoritma *Random Forest, KNN,* dan *SVM.* Algoritma-algoritma ini dipilih berdasarkan evaluasi kelebihan, kekurangan, dan informasi dari penelitian sebelumnya terkait prediksi *URL website phising* menggunakan *data mining*. Google colab dan *library* terkait akan dipakai untuk menyusun prediksi dalam pemodelan.

Tabel 3. 3 Kelebihan dan Kekurangan Algoritma

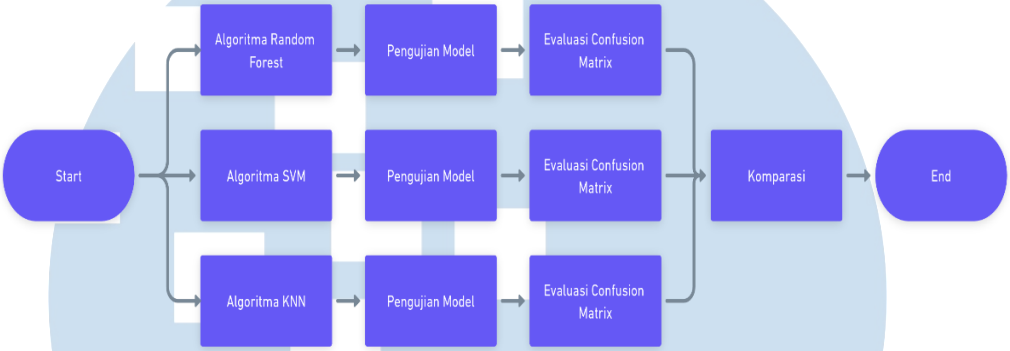
Algoritma	Kelebihan	Kekurangan
Random Forest	1. Kekuatan dalam Klasifikasi dan Regresi: Mampu menangani baik masalah klasifikasi maupun regresi. 2. Pemrosesan Data yang Besar: Dapat bekerja dengan baik pada dataset yang besar dan memiliki kemampuan untuk menangani banyak fitur.	1. Kompleksitas Model: Dalam beberapa kasus, model Random Forest cenderung terlalu kompleks dan bisa sulit diinterpretasi. 2. Waktu Komputasi: Pada dataset yang sangat besar, waktu komputasi untuk membuat model bisa menjadi lambat.
SVM	1. Efektif dalam Dimensi Tinggi: Cocok untuk *dataset* dengan dimensi tinggi dan bisa mengatasi *overfitting*. 2. Mampu Menangani *Non-linearitas*: Menggunakan *kernel trick* untuk menangani hubungan *non-linear* dalam data.	1. Sulit diatur: Memiliki beberapa parameter yang perlu diatur, yang dapat mempengaruhi performa model. 2. Kinerja Terbatas pada *Dataset* Besar: Waktu pelatihan *SVM* dapat memakan waktu pada dataset yang sangat besar.

Algoritma	Kelebihan	Kekurangan
KNN	1. Sederhana dan Mudah Dimengerti: Konsep *KNN* relatif mudah dipahami. 2. Penerapan yang Fleksibel: Cocok untuk klasifikasi dan regresi, serta bisa digunakan dengan berbagai metrik jarak.	1. Pengaruh *Outlier*: Rentan terhadap pengaruh *outlier* karena keputusan berdasarkan sejumlah tetangga terdekat. 2. Kinerja yang Kurang di Dataset Besar: Pada dataset besar, waktu komputasi dan kinerja *KNN* bisa menjadi masalah.
Naïve Bayes	1. Sederhana dan Cepat: Model yang sederhana dan efisien dalam pembelajaran. 2. Efektif pada Dataset Kecil: Bekerja dengan baik pada dataset kecil dan saat asumsi independensi berlaku.	1. Asumsi Independensi yang Sering Tidak Realistis: Asumsi bahwa semua fitur adalah independen tidak selalu berlaku pada data aktual. 2. Kinerja Terbatas pada Dataset yang Lebih Kompleks: Tidak begitu cocok untuk dataset yang sangat kompleks atau memiliki hubungan yang kompleks antara fitur.
Decision Tree	1. Interpretasi yang Baik: Model *decision tree* mudah diinterpretasi oleh manusia. 2. Pemrosesan Data yang Efisien: Bisa mengatasi dataset dengan jumlah besar dan atribut yang beragam.	1. Kecenderungan *Overfitting*: *Decision tree* cenderung *overfit* pada data yang kompleks jika tidak diatur dengan baik. 2. Kehilangan Informasi: Pada saat pemisahan node, informasi bisa hilang jika tidak ada atribut lain yang dapat memecahkan data.

Pemilihan ketiga algoritma ini didapat dari studi-studi terdahulu di mana *Random Forest* telah sering diaplikasikan dalam memprediksi *URL website phishing*. Algoritma *Random Forest* dikenal memiliki kinerja

yang solid dan menunjukkan tingkat kesalahan yang rendah pada hasil prediksinya.

3.5.5 *Evaluation*



Gambar 3. 5 Flowchart Evaluation

Pada tahap kelima ini dilakukan implementasi menggunakan phyton. Penelitian ini juga memanfaatkan metrik A (Akurasi) P (*Precision*), R (*Recall*), dan F1 *Score* untuk menilai kinerja model klasifikasi yang dibangun. *Evaluation* adalah tahap penilaian terhadap model yang telah dibuat pada tahap pemodelan. Fokus utamanya adalah memastikan apakah hasil yang diperoleh sudah sejalan dengan tujuan bisnis yang telah ditetapkan sebelumnya. Studi ini memanfaatkan sejumlah metrik performa seperti waktu pemrosesan, akurasi, sensitivitas, dan presisi untuk mengevaluasi kinerja dari setiap model yang telah dibuat. Setelah itu, model mana yang menghasilkan prediksi terbaik akan ditentukan. Dari hasil evaluasi ini, data prediksi gangguan mental akan dianalisis bersama data demografi dalam tahap deployment yang akan datang. Evaluasi kinerja model klasifikasi berfokus pada pengukuran seberapa baik sistem mampu mengklasifikasikan data [18]. Salah satu cara untuk mengukur kinerja model klasifikasi adalah dengan menggunakan confusion matrix [18].

Dalam kerangka matriks kebingungan (*confusion matrix*), terdapat empat kemungkinan hasil klasifikasi, yakni TP (*True Positive*), FP (*False Positive*), TN (*True Negative*), dan FN (*False Negative*). Matriks kebingungan sering digunakan dalam menghitung tingkat ketepatan pada

konsep *data mining*. Berikut adalah tabel matriks kebingungan yang dimaksud:

Tabel 3. 4 Hasil True dan False

Prediksi	**Hasil**	
	1	0
1	TP (*True Positive*)	FN (*False Negative*)
0	FN (*False Negative*)	TN (*True Negative*)

Precision, dikenal juga sebagai nilai prediksi positif, merupakan persentase prediksi yang benar dan dihitung dengan TP / (TP + FP). *Recall* adalah proporsi positif aktual dalam data yang diuji dan diwakili oleh TP / (TP + FN), sedangkan *F-Measure* adalah nilai rata-rata dari *Precision* dan *Recall*, dihitung sebagai 2 (P x R) / (P + R). Ketiga metrik ini memiliki rentang nilai antara 0 hingga 1. Dalam penelitian sebelumnya oleh Dongsong Zhang dan rekan-rekannya [2], mereka menduga bahwa model klasifikasi yang menggunakan fitur berbasis konten dan *URL* akan mengungguli model dengan fitur tradisional berdasarkan *Precision*, *Recall*, dan *F-Measure*. Sementara dalam penelitian ini, hipotesanya adalah bahwa fitur baru yang berfokus pada pendekatan konten dan *URL* dapat meningkatkan kinerja deteksi situs *phishing* jauh melampaui fitur dasar pada penelitian sebelumnya [2]