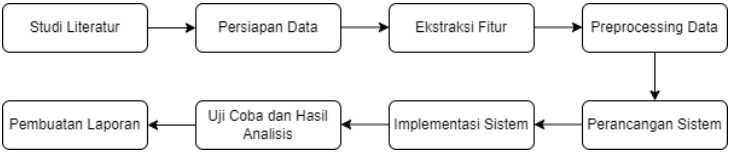
BAB 3
Metodologi Penelitian

3.1 Diagram Alur Metodologi Penelitian

Sistem yang diusulkan pada penelitian ini terdiri dari dua sistem utama, yaitu Sistem Evaluasi dan Implementasi Sistem, dimana implementasi sistem mengandung proses penting, yaitu fitur ektraksi, sementara pada Sistem Evaluasi akan ada menjadi proses pengoptimalan parameter dari algoritma klasifiasi yang akan digunakan. Algoritma yang akan digunakan dalam penelitian ini adalah *Logistic Regression*, *K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting* dan *Catboost* sebagai perbandingan kinerja sistem.

Tahapan yang akan dilakukan dalam penelitian ini dapat dilihat pada gambar berikut:



Gambar 3.1 Diagram Alur Metodologi Penelitian

Berdasarkan diagram pada diatas, secara umum penelitian dapat digambarkan sebagai berikut:

1. Studi Literatur merupakan tahapan awal dari penelitian ini. Tahapan ini dilakukan untuk mengumpulkan penelitian yang berkaitan dengan metode yang digunakan kepada ektraksi fitur dan klasifikasi.

2. Persiapan data merupakan langkah yang dilakukan untuk mendapatkan dataset yang akan digunakan dalam mengklasifikasikan *website*. Dataset yang digunakan bersumber dari *Kaggle*.
3. Ektraksi fitur dilakukan untuk mengekstraksi fitur terdapat di situs web berdasarkan dataset yang diperoleh dari *Kaggle*. Hasil ektraksi fitur ini kemudian akan digunakan untuk mendeteksi situs web *phishing*.
4. *Preprocessing* Data dilakukan dalam bentuk menganalisis data yang akan digunakan untuk memilih data berdasarkan hasil ekstraksi ciri yang telah dilakukan sebelumnya.
5. Perancangan sistem pada penelitian ini dilakukan untuk menentukan algoritma yang akan dinguakan dalam klasifikasi situs web phising. Algoritma-algoritma yang digunakan dalam mengklasifikasi adalah *Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting* dan *Catboost* sebagi perbandingan kinerja sistem. Pada tahapan ini dibuat *flowchart* yang berkaitan dengan alur kerja sistem.
6. Implementasi sistem dilakukan sesuai dengan *flowchart* yang telah dibuat ditahap sebelumnya. Pada penelitian ini, sistem dibuat dengan menggunakan bahasa pemrograman *python* dan *framework flask*.
7. Pengujian sistem dilakukan untuk mengetahui keakuratan sistem yang dibuat, sistem akan diuji dengan beberapa percobaan dan bentuk URL yang berbeda.
8. Tahapan akhir dalam penelitian ini adalah menulis laporan penelitian dalam bentuk laporan.

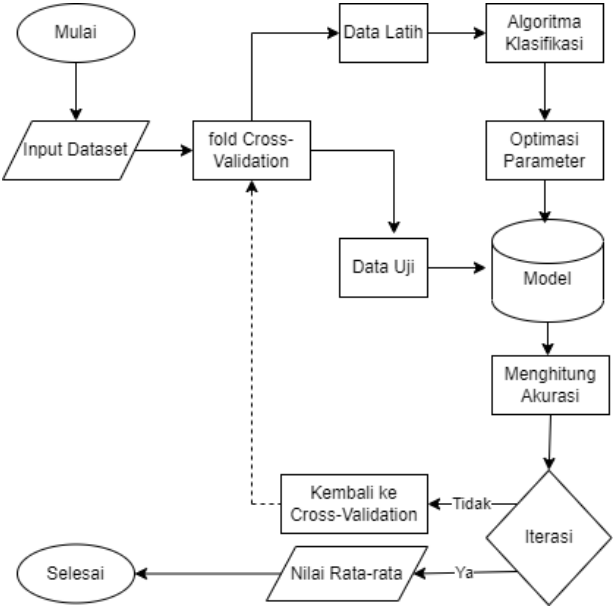
3.2 Tahapan – tahapan Diagram Alur Metofologi Penelitian

Pada tahapan penelitian ini dilakukan untuk mengektrak fitur yang terdapat pada situs web dengan memasukkan URL ke dalam sistem dankemudian sistem akan mengkase URL dan kemudian istem akan mengakses URL dan melakukan fitur ekstraksi di situs *website*. Karena untuk mengakses situs web yang akan dilakukan deteksi harus terhubung kedalam internet. Sistem ini secara otomatis menjadi sistem online.

Sistem yang dibuat terdiri dari dua bagian yaitu sistem pelaksanaan dan sistem evaluasi. Implementasi sistem dapat digunakan ketika pengguna penginputkan URL kedalam form. Sedangkan sistem evaluasi adalah sistem dibuat untuk menganalisis kinerja dari sistem implementasi.

3.2.1 Sistem Evaluasi

Sistem evaluasi merupakan sistem yang dirancang sebagai untuk evaluasi sistem produksi. Evaluasi terhadap sistem ini dilakukan dengan cara mengevaluasi dataset yang digunakan dalam implementasi sistem. *Flowchart* sistem evalusi dapat dilihat pada gambar berikut:



Gambar 3.2.1 Flowchart Sistem Evaluasi

Pada tahapan diatas, proses pembuatan model deteksi phishing dimulai dari input dataset. Dataset yang digunakan berisi informasi tentang URL situs web beserta label kelas yang mengidentifikasinya sebagai situs web *phishing* atau bukan. Setelah dataset diinput, proses selanjutnya adalah melakukan *fold Cross Validation*.

Cross Validation adalah metode untuk mengevaluasi kinerja model dengan membagi dataset menjadi beberapa bagian yang disebut sebagai fold. Dari beberapa fold tersebut, satu fold digunakan sebagai data pengujian dan sisanya digunakan sebagai data latih. Proses ini dilakukan beberapa kali dengan menggunakan setiap fold sebagai data pengujian. Hal ini dilakukan untuk mengevaluasi kinerja model dengan menghitung rata-rata akurasi dari setiap fold.

Setelah proses *fold Cross Validation* selesai, data tersebut dilatih menggunakan algoritma klasifikasi. Algoritma yang digunakan dapat berbeda-beda tergantung pada kebutuhan dan karakteristik dari dataset yang digunakan. Selanjutnya, dilakukan optimasi parameter untuk meningkatkan kinerja model.

Pengujian untuk pembuatan model. Pengujian ini dilakukan dengan membandingkan kinerja model yang telah dilatih dengan dataset pengujian yang tidak digunakan sebelumnya. Hal ini dilakukan untuk mengevaluasi kinerja model dengan data yang belum pernah dilihat sebelumnya.

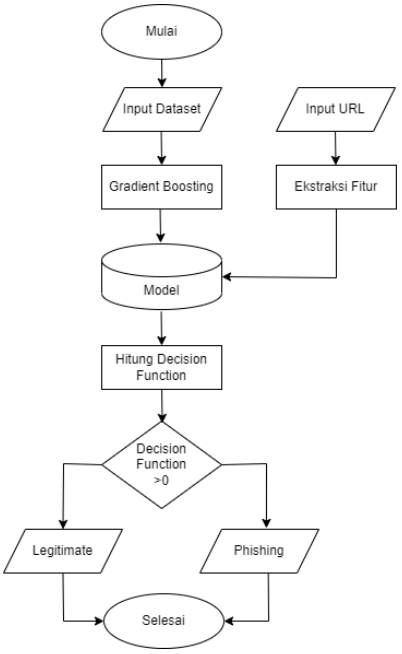
Setelah proses pengujian selesai, dilakukan penghitungan akurasi dari model yang telah dibuat. Akurasi adalah seberapa baik model dapat mengklasifikasikan data. Semakin tinggi nilai akurasi, semakin baik kinerja model. Namun, dalam beberapa kasus, akurasi tidak cukup untuk menentukan kinerja model yang baik.

Selanjutnya, dilakukan tahapan iterasi. Tahapan ini dilakukan untuk meningkatkan kinerja model dengan cara mengubah beberapa parameter atau mencoba algoritma klasifikasi yang berbeda. Proses iterasi ini dilakukan sampai kita dapat menentukan nilai rata-rata akurasi yang diinginkan. Jika nilai rata-rata akurasi sudah dapat ditentukan, maka program akan selesai dan model yang telah dibuat dapat digunakan untuk melakukan deteksi *phishing*.

3.2.2 Sistem Implementasi

Sistem implementasi merupakan sistem yang dirancang untuk digunakan dan diimplementasikan dalam

melakukan deteksi website *phishing*. Hal ini yang membedakan sistem implementasi dengan sistem evaluasi adalah sistem ini dirancang untuk digunakan oleh pengguna dengan input berupa URL. Selain ini perbedaannya terdapat pada proses ekstraksi yang sangat penting dalam menentukan kinerja sistem. Untuk lebih jelasnya dapat dilihat pada *flowchart* sistem berikut:



Gambar 3.2.2 Flowchart Sistem Implementasi

Sistem implementasi dalam deteksi *website phishing* ini dimulai dengan mengumpulkan dataset yang akan digunakan dalam proses pelatihan model. Data ini dikumpulkan dari berbagai sumber dan diperoleh dalam bentuk URL yang telah diklasifikasikan sebagai *phishing* atau legitimate. Setelah dataset diinputkan, maka data tersebut diolah dengan menggunakan algoritma *Gradient Boosting*. Algoritma ini

dipilih karena dapat menangani masalah klasifikasi dengan baik dan memiliki tingkat akurasi yang tinggi.

Setelah model pelatihan selesai, maka kita dapat menggunakan algoritma tersebut untuk digunakan saat menbangun *website* yang bertujuan untuk mendeteksi *website phishing*. Dalam tahap ini, kita akan membuat sebuah form input yang digunakan untuk memasukkan URL yang akan diuji. Selanjutnya, kita akan melakukan ekstraksi fitur dari URL yang telah diinputkan. Fitur-fitur ini akan digunakan sebagai input dalam model yang telah dibuat sebelumnya.

Setelah fitur-fitur di ekstrak, maka data tersebut akan masuk kedalam model yang telah dibuat. Model ini akan menghasilkan sebuah decision *function* yang digunakan untuk menentukan apakah suatu *website* merupakan *website phishing* atau *legitimate*. *Decision function* ini akan mengevaluasi fitur-fitur yang telah diinputkan dan memberikan hasil dalam bentuk prediksi yang dapat digunakan untuk menentukan apakah suatu *website* merupakan *website phishing* atau *legitimate*. Proses deteksi website phishing akan selesai setelah *decision function* ini dijalankan dan hasilnya ditampilkan kepada pengguna.

3.3 Metodologi Data Science

Metodologi data science untuk proyek deteksi halaman website phishing menggunakan algoritma Machine Learning Gradient Boosting Classifier adalah sebagai berikut:

1. Definisi masalah: Pemahaman akan masalah deteksi halaman website phishing dan tujuan dari proyek ini adalah untuk mengembangkan sebuah sistem yang dapat mendeteksi halaman website phishing dengan tingkat akurasi tinggi.

2. Pengumpulan data: Mengumpulkan data halaman website phishing dan legitimate dari berbagai sumber seperti database, file spreadsheet, atau melalui scraping dari website. Data ini digunakan untuk melatih dan menguji model.
3. Analisis data: Melakukan analisis data untuk mengekstrak fitur yang relevan dari setiap halaman website yang diambil. Fitur ini dapat meliputi informasi seperti URL, jumlah link internal dan eksternal, dan kata-kata yang digunakan dalam halaman website.
4. Pemodelan: Membuat model dengan menggunakan algoritma Machine Learning Gradient Boosting Classifier. Model ini dibangun dengan menggunakan data latih yang telah dikumpulkan sebelumnya dan digunakan untuk memprediksi apakah suatu halaman website merupakan phishing atau legitimate.
5. Evaluasi: Evaluasi model yang dibuat dengan menguji model dengan menggunakan data uji yang telah dikumpulkan sebelumnya. Model di evaluasi dengan menggunakan metrik seperti akurasi, recall, dan precision.
6. Deployment: Implementasi model dalam sebuah aplikasi atau sistem yang digunakan oleh pengguna akhir. Aplikasi ini dapat digunakan untuk memasukkan URL yang akan diuji dan menampilkan hasil prediksi dari model.