BAB II

LANDASAN TEORI

2.1 Tinjauan Teori

2.1.1 *Phishing*

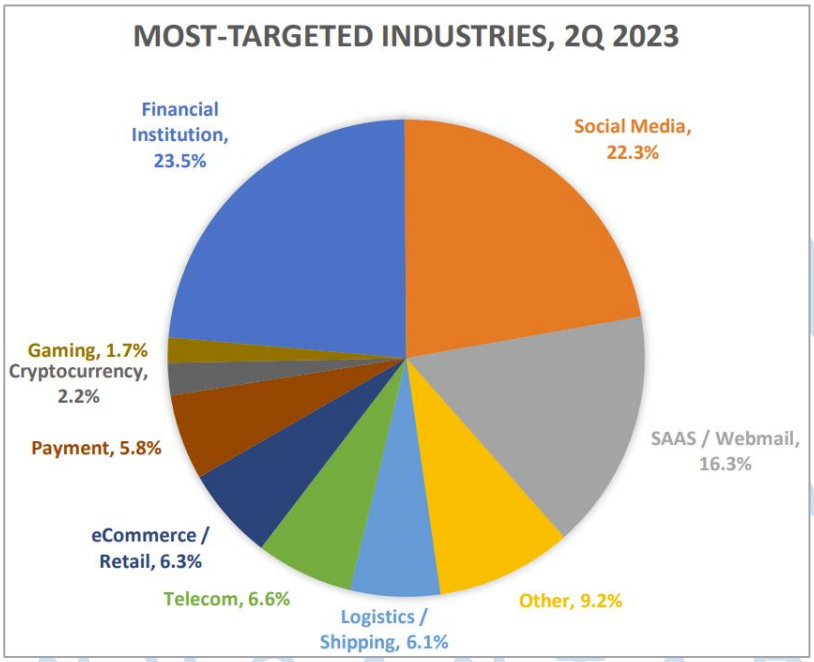
Phishing adalah suatu skema kriminal berbasis kegiatan *cyber* yang dirancang untuk menarik perhatian. *Phising* juga merupakan suatu kegiatan yang berpotensi membahayakan dan menjerat seseorang dengan cara memancing target untuk secara tidak langsung memberikan informasi data pribadi kepada penipu. Selain itu, tujuan dari *phishing* adalah mengirimkan tautan berbahaya, seringkali menyamar sebagai tautan *website* yang sah, melalui spam atau jejaring sosial, dengan maksud mendorong pengguna untuk mengunjungi tautan tersebut dan memberikan informasi pribadi mereka[7].

Phishing merujuk pada upaya memperoleh informasi rahasia seperti *username, password,* dan kartu kredit dengan menyamar sebagai entitas terpercaya melalui komunikasi elektronik resmi seperti surat elektronik atau pesan instan. Mengingat tingginya jumlah kasus penipuan yang dilaporkan, diperlukan metode tambahan atau perlindungan. Langkah-langkah perlindungan tersebut melibatkan pembuatan undang-undang, pelatihan pengguna, dan tindakan teknis. Proses *phishing* seringkali sulit dideteksi, terutama bagi mereka yang tidak memiliki latar belakang teknis[3].

Tindakan *phising* umumnya dilakukan dengan sengaja oleh pihak-pihak seperti orang dalam, *hacker*, atau penjahat internet yang berhasil meretas suatu situs *web* melalui celah keamanan yang ada pada situs tersebut. Setelah itu, mereka menempatkan halaman *phising* atau menciptakan halaman *phising* baru yang menyerupai. Tidak hanya itu, situs *phishing* juga kerap dijadikan sebagai sarana penyebaran *malware*

dan praktik penipuan oleh pelaku kejahatan internet, yang berusaha menyamar sebagai situs asli[8]. Tidak hanya itu, situs *phishing* juga kerap dijadikan sebagai sarana penyebaran *malware* dan praktik penipuan oleh pelaku kejahatan internet, yang berusaha menyamar sebagai situs asli.

Berdasarkan hasil laporan dari Anti *Phishing* Working Group (APWG), tercatat pada kuartal 2 tahun 2023 terdapat 1.286,208 serangan phishing di Indonesia[9]. Jumlah ini merupakan jumlah triwulan tertinggi ketiga yang pernah dicatat APWG selama melakukan pengamatan pada serangan *phishing*. Dalam laporan ini sektor keuangan termasuk perbankan menjadi sektor penyerangan tertinggi yaitu sebanyak 23,5% dari total laporan *phishing* yang ada. Sektor tertinggi kedua yaitu ada pada sektor media sosial dengan persentase 22,3%, dan persentase ketiga tertinggi ada pada sektor webmail atau SAAS sebesar 16,3%. Hingga saat ini sektor keuangan dan bank masih menjadi sektor tertinggi yang sangat rentan terhadap serangan phishing yang ada di Indonesia, termasuk serangan terhadap layanan pembayaran *online* merupakan 5,88% dari seluruh total serangan.



Gambar 2. 1 Laporan Phishing Berdasarkan APWG 2023

2.1.2 Deteksi *Phishing*

Deteksi *phishing* adalah cara untuk mengetahui apakah sebuah alamat situs *web (URL)* itu palsu atau asli. Ada beberapa cara untuk menilai seberapa aman sebuah *URL*, seperti menggunakan daftar hitam, daftar putih, statistik, atau teknologi pembelajaran mesin. Di antara semuanya, teknologi pembelajaran mesin lebih bagus karena lebih efisien dan akurat. Teknologi ini menggunakan algoritma khusus untuk memahami pola *URL* yang berbahaya dan mampu mengatakan jenis *URL*, apakah itu phishing atau situs yang aman, sesuai dengan kebutuhan kita.[4]

Pendeteksian *phishing* merupakan rangkaian strategi yang melibatkan deteksi keberadaan serangan phishing, penerapan pertahanan ofensif, tindakan korektif saat terjadinya, dan langkah-langkah pencegahan guna mengurangi kemungkinan serangan di masa depan. Namun, tidak ada solusi tunggal yang dapat secara sepenuhnya mengatasi atau meminimalisir segala kerentanan yang mungkin muncul. Karena serangan *phishing* memiliki variasi yang luas, dengan perkembangan taktik dan strategi yang terus berkembang, solusi terbaik seringkali memerlukan kombinasi pendekatan dan metode yang beragam untuk memberikan perlindungan yang lebih efektif.

Pendekatan deteksi *phishing* yang mengimplementasikan skema pendeteksian di sisi server terbukti lebih efektif dibandingkan dengan strategi pencegahan *phishing* dan sistem pelatihan pengguna. Sistem ini dapat diakses melalui *web* pada perangkat klien atau melalui perangkat lunak yang dihosting pada situs tertentu. Sistem ini mempresentasikan variasi pendekatan klasifikasi untuk mendeteksi *phishing*. Pendekatan yang berbasis heuristik dan *Machine Learning (ML)* diterapkan dengan menggunakan teknik pembelajaran yang bersifat diawasi maupun tidak diawasi. Untuk menerapkan pendekatan ini, fitur atau label diperlukan agar sistem dapat mempelajari lingkungannya dan membuat prediksi. Pendekatan proaktif dalam mendeteksi *URL* phishing mirip dengan

metode *ML*, namun, dalam hal ini, *URL* diproses untuk mendukung sistem dalam memprediksi apakah suatu *URL* dianggap sah atau berbahaya. Pertumbuhan eksponensial dalam domain *web* telah mengurangi kinerja dari metode-metode tradisional tersebut[10]

2.1.3 Jenis-jenis *Phishing*

Serangan *phishing* terdapat beberapa jenis yang ada[11], berikut merupakan beberapa jenis serangan *phishing* yang ada di Indonesia:

1. *Web phishing*: serangan *phishing* yang dilakukan dengan cara membuat *website* palsu suatu instansi atau organisasi untuk menipu korbannya. *Website* yang dibuat oleh pelaku sangat menyerupai *website* aslinya sehingga membuat korban percaya bahwa *website* tersebut merupakan *website* asli. Tentunya hal ini menjadi sangat sulit untuk individu dapat membedakan *website* asli dengan *website* palsu.
2. *Email phishing*: serangan yang diluncurkan oleh *spammer* dengan tujuan mengelabuhi korban dengan cara mengaku sebagai instansi/organisasi yang cukup terkenal dikalangan masyarakat supaya korban percaya dan memberikan informasi data pribadi mereka.
3. *Smishing*: *phishing* yang menggunakan pesan teks yang dikirim dalam bentuk *sms* ke nomor pribadi korban. Teks *sms* yang dikirim berupa teks pemberitahuan bahwa korban memenangkan undian atau mendapatkan hadiah dari *brand* tertentu, sehingga membuat korban penasaran dan mengklik *link* yang dikirimkan pelaku.
4. *Deceptive Phishing*: penipuan ini dilakukan melalui *email* maupun *whatsapp* dengan mengirimkan *link* yang diberikan pesan teks dengan menggunakan nama *brand* terkenal.
5. *Whaling*: *phishing* yang dilakukan kepada orang yang mempunyai kekuasaan tinggi di suatu instansi/perusahaan/organisasi.
6. *Spear Phishing*: penipuan yang telah memiliki target korban sejak lama dan memiliki suatu tujuan tertentu agar dapat melakukan penipuan terhadap korban sasarannya.

2.1.4 *Uniform Resource Locator (URL) Website*

URL adalah singkatan dari *Uniform Resource Locator*, yang merupakan rangkaian karakter tertentu, seperti angka, huruf, dan simbol, yang berdasarkan format standar tertentu. Fungsi *URL* adalah untuk menunjukkan alamat atau suatu sumber yang berada di internet, seperti *file*, dokumen, dan gambar yang ada di internet[12]. *URL* adalah alat yang sangat berguna dalam mengklasifikasikan situs *web*, baik yang berpotensi berbahaya maupun yang aman, dengan memanfaatkan fitur-fitur tertentu. Klasifikasi ini didasarkan pada analisis leksikal, informasi dari *host*, serta konten yang terdapat dalam *URL* itu sendiri. Pendekatan berbasis fitur ini memungkinkan pengidentifikasian yang lebih baik terhadap sifat situs *web*, membantu dalam memahami apakah suatu situs memiliki risiko atau bersifat aman bagi pengguna yang mengaksesnya[13].

URL yang dapat dianggap berbahaya merujuk pada situs *web* yang diciptakan dengan tujuan yang merugikan. Biasanya, situs-situs ini mengandung konten *spam*, serangan *phishing* yang berupaya mencuri informasi sensitif, serta aplikasi yang menyesatkan. Mereka sering dibentuk sedemikian rupa sehingga terlihat sangat mirip dengan situs *web* asli, hal ini dilakukan untuk mengecoh pengguna agar percaya dan terjebak dalam aktivitas atau transaksi yang merugikan. Tujuan utama dari *URL* berbahaya adalah untuk menipu pengguna, menyebarkan *malware*, atau mendapatkan akses tanpa izin ke informasi pribadi[14]. Dalam penelitianan ini *URL* merupakan salah satu faktor yang perlu diperhatikan, karena *URL* canggih memungkinkan penyerang untuk menuju ke situs *web* yang dirancang untuk *phishing*, oleh karena itu, penting untuk memahami struktur dan komponen URL serta menjaga keamanan saat mengakses situs *web* yang *suspekt*.

2.1.5 *Data Mining*

Data mining adalah teknologi yang memungkinkan konversi data menjadi pengetahuan yang berharga dengan mengungkap pola-pola

tersembunyi di dalamnya. Dengan kemampuannya dalam menemukan, mengklasifikasikan, mengelompokkan, serta merangkum hubungan antara berbagai dimensi data, data *mining* memainkan peran kunci dalam menerjemahkan informasi yang tersembunyi dalam kumpulan data yang besar dan kompleks[15]. Proses ini memungkinkan pengguna untuk mengeksplorasi data lebih dalam, mengidentifikasi hubungan yang signifikan, dan menghasilkan wawasan yang dapat digunakan untuk pengambilan keputusan yang lebih baik di berbagai bidang, mulai dari ilmu pengetahuan hingga bisnis.

Data mining merupakan proses ekstraksi informasi berharga dan pola-pola tersembunyi dari kumpulan data besar yang bersifat mentah. Melalui model ini, seringkali ditemukan wawasan yang tak terduga dan tersembunyi dalam basis data yang dimiliki[16]. Dengan memanfaatkan perangkat lunak khusus, data *mining* memungkinkan analisis yang mendalam terhadap data yang kompleks, memungkinkan pengguna untuk menggali dan memahami pola-pola, tren, serta hubungan yang ada di dalam data, yang nantinya dapat digunakan untuk pengambilan keputusan yang lebih baik.

Sebelum dilakukan data mining, ada serangkaian proses yang harus dilalui dalam implementasi *data mining*[12]. Langkah-langkah ini termasuk:

1. Penelitian Bisnis: Tahapan ini mengharuskan pemahaman menyeluruh mengenai tujuan organisasi, sumber daya yang tersedia, dan situasi saat ini yang sejalan dengan kebutuhan perusahaan.
2. Pemeriksaan Kualitas Data: Mengingat data dikumpulkan dari berbagai sumber, penting untuk memeriksa dan menyelaraskan data untuk memastikan kelancaran dalam integrasi data.
3. Pembersihan Data: Tahap ini melibatkan pemilihan, pembersihan, pengaturan format, dan pelabelan data agar siap untuk proses penambangan.

4. Transformasi Data: Transformasi data melibatkan beberapa langkah, seperti Penyusunan Data, Ringkasan Data, Generalisasi Data, Normalisasi Data, dan Konstruksi Atribut Data, yang bertujuan untuk meningkatkan pemahaman terhadap pola data.
5. Pemodelan Data: Dalam rangka mengidentifikasi pola data yang lebih efektif, berbagai model matematika diterapkan pada dataset berdasarkan kondisi tertentu. Langkah ini berfokus pada penggunaan model untuk menggali informasi yang signifikan dari data yang ada.

2.1.6 *Supervised*

Supervised learning, sebagai metode dalam data *mining*, adalah pendekatan yang digunakan untuk melakukan prediksi dengan bantuan *dataset* yang telah terlatih. Proses ini memerlukan kumpulan data pelatihan yang telah diberi nilai input yang berkaitan langsung dengan output yang menjadi target yang diinginkan[17]. Dengan menggunakan *dataset* ini, algoritma *machine learning* dapat "belajar" atau melatih dirinya sendiri untuk mengenali pola dan hubungan antara *input* dan *output*, sehingga nantinya dapat melakukan prediksi yang akurat terhadap data baru yang belum pernah dilihat sebelumnya. Metode *supervised learning* memainkan peran krusial dalam berbagai bidang, dari pengenalan pola hingga analisis prediktif, serta menjadi landasan penting dalam pengambilan keputusan yang diinformasikan.

Supervised learning melibatkan proses di mana algoritma *machine learning* membandingkan prediksi yang dihasilkannya dengan data sampel yang telah ditinjau atau diberi label oleh seorang ahli. Ahli ini memberikan jawaban atau label yang dianggap benar untuk data sampel yang digunakan dalam proses pembelajaran. Algoritma *supervised learning* kemudian menggunakan informasi ini untuk memperbaiki dan menyesuaikan prediksinya, dengan harapan dapat meningkatkan akurasi dan ketepatan hasil yang dihasilkan oleh algoritma tersebut[18]. Proses ini memungkinkan algoritma untuk belajar dari contoh-contoh yang telah

diawasi sehingga dapat menghasilkan prediksi yang lebih baik di masa mendatang.

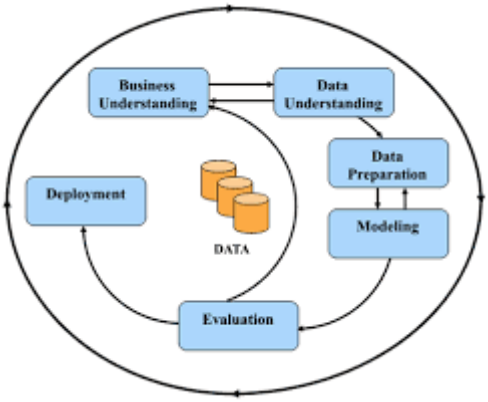
Setiap algoritma memiliki kelebihan, kelemahan, dan situasi yang cocok untuk penggunaannya tergantung pada sifat data dan masalah yang dihadapi dalam proses pembelajaran. Pada penelitian ini menggunakan tiga algoritma yaitu *Random Forest, Support Vector Machines,* dan *K-Nearest Neighbors.*

2.1.7 Klasifikasi

Klasifikasi merupakan sebuah metode yang melibatkan pengamatan terhadap perilaku dan atribut dari kelompok yang sudah didefinisikan sebelumnya. Teknik ini memungkinkan identifikasi klasifikasi pada data baru dengan memanipulasi data yang telah diklasifikasikan sebelumnya dan menggunakan hasilnya untuk menetapkan sejumlah aturan. Aturan-aturan ini kemudian diaplikasikan pada data baru untuk melakukan klasifikasi berdasarkan kesamaan atau pola yang telah teridentifikasi sebelumnya[19]. Dengan memanfaatkan informasi dari kumpulan data terklasifikasi, metode klasifikasi membantu dalam memprediksi kategori atau label yang tepat bagi data baru, memungkinkan identifikasi dan pengelompokan data dengan cara yang lebih terstruktur dan terarah.

2.2 *Framework*

CRISP-DM (Cross Industry Standard Process for Data Mining) adalah sebuah kerangka kerja yang digunakan secara luas dalam berbagai industri untuk menjalankan proses *data science*. Metodologi ini merinci tahapan dan tugas-tugas yang terlibat dalam proyek *data science* serta menjelaskan hubungan antara setiap tugasnya[20]. *CRISP-DM* dikenal sebagai salah satu pendekatan yang populer dan sering digunakan dalam praktik dan penelitian di bidang *data science.*



Gambar 2. 2 Alur CRISP-DM

Berikut adalah langkah-langkah dalam memproses data sesuai dengan kerangka kerja *CRISP-DM*[20]. Model ini terdiri dari lima tahap, yakni pemahaman bisnis, pemahaman data, persiapan data, pemodelan, dan evaluasi.

1. Tahap awal yaitu *Business Understanding*: tahapan ini melibatkan pengenalan proses dalam organisasi serta pemahaman terhadap sistem yang berjalan, serta kebutuhan-kebutuhan yang diperlukan untuk menyelesaikan masalah yang ada. Tahap ini mengubah pengetahuan yang ada menjadi definisi masalah *data mining* dan rencana awal yang dirancang untuk mencapai tujuan.
2. Tahap kedua adalah *Data Understanding*: pengertian atas data *(Data Understanding)* mencakup pengumpulan data dan pemahaman awal terhadap sifat data sebelum persiapan analisis data. Tahap ini dimulai dengan mengumpulkan data awal, langkah-langkah berikutnya melibatkan aktivitas yang memungkinkan pemahaman lebih lanjut terhadap data, mengidentifikasi masalah kualitas data, mengeksplorasi informasi yang terkandung dalam data, dan/atau mendeteksi *subset* data yang menarik. Tujuan dari langkah-langkah ini adalah membentuk hipotesis mengenai informasi yang mungkin tersembunyi dalam *dataset.*
3. Pada tahap ketiga yaitu *Data Preparation/ Data Preprocessing*: pada tahapan ini data yang telah terkumpul akan disortir dan diubah ke dalam bentuk yang sesuai dengan model yang akan digunakan dalam tahap selanjutnya. Tahap persiapan data melibatkan serangkaian kegiatan untuk

menyusun data akhir yang akan dimasukkan ke dalam alat pemodelan dari data mentah awal. Proses persiapan data mungkin berulang beberapa kali dan tidak memiliki urutan yang baku. Kegiatan ini termasuk pemilihan tabel, catatan, dan atribut yang relevan, serta transformasi dan pembersihan data agar siap digunakan dalam alat pemodelan.

4. Dalam tahap keempat yaitu Pemodelan (*Modelling*): analisis data dilakukan menggunakan algoritma atau metode yang telah ditentukan sesuai dengan kebutuhan untuk mewakili solusi terhadap masalah yang ada. Pada tahap ini, berbagai teknik pemodelan dipilih dan diterapkan, dengan penyesuaian parameter untuk mencapai nilai optimal. Dalam jenis masalah *data mining* yang sama, seringkali terdapat beberapa teknik yang dapat digunakan. Setiap teknik mungkin memiliki persyaratan khusus terkait tipe data yang digunakan, menjadikan tahap persiapan data sangat penting dalam proses ini.
5. Tahap kelima selanjutnya adalah proses *Evaluation*: pada tahap ini dilakukan evaluasi (*Evaluation*) terhadap model data untuk memeriksa kecocokannya terhadap kebutuhan dan kemampuannya dalam memberikan solusi atas masalah yang dihadapi. Sebelum melanjutkan ke langkah penerapan model akhir, evaluasi menyeluruh dan peninjauan langkah-langkah yang diambil untuk membuat model tersebut sangat penting. Tujuan utamanya adalah untuk memastikan bahwa model tersebut benar-benar mencapai tujuan bisnis yang telah ditetapkan. Proses evaluasi ini bertujuan untuk mengidentifikasi apakah ada aspek bisnis yang mungkin belum dipertimbangkan dengan memadai. Pada akhir tahap ini, keputusan penggunaan hasil dari data mining harus dicapai.
6. Tahap terakhir yaitu tahap Implementasi (*Deployment*): tahap ini melibatkan penerapan hasil dari model yang telah dievaluasi, menjadikannya dalam bentuk yang dapat diolah kembali atau diintegrasikan ke dalam sistem yang ada. Membuat model hanyalah bagian dari keseluruhan proyek. Tahap *deployment* dapat beragam, mulai dari

penyusunan laporan hingga penerapan proses pengumpulan data yang rumit di seluruh organisasi.

2.3 Algoritma

Algoritma merupakan serangkaian langkah-langkah yang dilakukan secara berurutan dan mandiri untuk menyelesaikan suatu masalah[21]. Ketika dirancang, algoritma mempertimbangkan beberapa faktor penting, seperti waktu yang dibutuhkan untuk menyelesaikan tugas, kemampuan untuk beradaptasi dengan perangkat komputer yang digunakan, serta tingkat kesederhanaannya. Tujuan utama dari algoritma adalah memberikan panduan yang jelas dan sistematis bagi komputer untuk menyelesaikan masalah dengan efisien dan tepat sesuai dengan kebutuhan yang ada[22]

2.3.1 *Random Forest*

Random Forest merupakan salah satu algoritma pembelajaran yang dikenal karena kemampuannya menghasilkan prediksi yang akurat, terutama dalam kerangka kerja dengan data berdimensi tinggi[23]. Kelebihan utama dari *Random Forest* adalah kemampuannya untuk mengatasi masalah *overfitting* yang sering terjadi dalam model yang kompleks, serta mampu berperforma baik tanpa memerlukan penyetelan parameter yang sangat spesifik[24]. Dengan menggunakan pendekatan *ensemble learning* yang memanfaatkan banyak pohon keputusan, algoritma ini mampu membuat prediksi yang andal dengan mempertimbangkan hasil dari sejumlah besar pohon keputusan yang beroperasi secara independen. Hal ini membuat *Random Forest* menjadi salah satu pilihan yang populer dalam analisis data, terutama dalam situasi di mana keakuratan prediksi yang tinggi dibutuhkan tanpa harus mengorbankan kompleksitas model atau memerlukan penyetelan parameter yang rumit[25].

2.3.2 *Support Vector Machine (SVM)*

SVM, yang merupakan singkatan dari *Support Vector Machine*, adalah algoritma pembelajaran mesin yang digunakan untuk membuat

model klasifikasi dan regresi. Tujuannya adalah menemukan garis pemisah terbaik di antara dua kelas data[26]. Algoritma ini menjadi salah satu yang sangat diperhitungkan dalam domain pengenalan pola dan memiliki berbagai aplikasi yang luas. Kemampuannya yang kuat dan tangguh dalam menemukan *hyperplane* terbaik yang memisahkan dua kelas dengan margin maksimum membuatnya menjadi alat yang penting dalam analisis data, pengenalan pola, dan aplikasi-aplikasi lainnya. *SVM* menjadi pilihan yang populer karena kemampuannya dalam menangani data dengan dimensi yang tinggi dan ketepatannya dalam menemukan garis pemisah yang optimal antara kelas-kelas yang berbeda. Algoritma ini memiliki parameter-parameter kunci seperti parameter *kernel* yang memungkinkan transformasi data ke ruang dimensi yang lebih tinggi untuk meningkatkan pemisahan kelas, serta parameter penalti yang mengatur *trade-off* antara kesalahan klasifikasi pada data latih dan margin pemisahan antara kelas[27].

2.3.3 *K-Nearest Neighbors (KNN)*

Algoritma *KNN (K-Nearest Neighbors)* adalah salah satu algoritma *data mining* yang populer dan telah banyak diimplementasikan dalam aplikasi analisis data di berbagai bidang penelitian ilmu komputer. Pendekatan *KNN* ini terkenal karena kemampuannya dalam mengklasifikasikan data berdasarkan kedekatan atau kesamaan dengan data latih yang ada. Metode ini sering digunakan untuk pemecahan masalah klasifikasi, di mana data akan ditempatkan dalam kategori berdasarkan mayoritas kategori dari tetangga terdekatnya dalam ruang fitur[28]. *KNN* menjadi alat yang penting dalam analisis data karena kemudahannya dalam implementasi, serta kemampuannya dalam menangani berbagai jenis data dan skenario dalam penelitian ilmu komputer.

Algoritma *KNN*, singkatan dari *K-Nearest Neighbors*, merupakan salah satu algoritma dalam pembelajaran mesin yang mengklasifikasikan

data berdasarkan kemiripan dengan data latih yang telah ada sebelumnya[29]. Pendekatan ini memungkinkan algoritma untuk memprediksi klasifikasi data baru dengan melihat kesamaannya dengan data yang telah terlatih sebelumnya. Metode *KNN* ini memanfaatkan prinsip bahwa data baru akan ditempatkan dalam kelas yang sama dengan mayoritas kelas dari tetangga terdekatnya dalam ruang fitur. *KNN* sering digunakan dalam berbagai aplikasi yang memerlukan prediksi klasifikasi, terutama karena sifatnya yang sederhana namun seringkali efektif dalam menangani berbagai jenis masalah klasifikasi[29].

2.4 *Tools*

2.4.1 Kaggle

Gambar 2. 3 Logo Kaggle

Kaggle adalah sebuah *platform* sains data yang menawarkan akses kepada deret waktu harian dan mingguan yang mencakup variabel eksogen serta informasi hirarki bisnis[30]. Di *platform* ini, *dataset* yang disebut sebagai *Kaggle dataset* tersedia untuk diakses oleh pengguna. Kaggle tidak hanya menyediakan akses terhadap *dataset*, tetapi juga menjadi wadah bagi pengembangan model, pelaksanaan kompetisi data, serta proyek-proyek ilmiah dan analisis data. *Dataset* yang tersedia di Kaggle meliputi berbagai topik, mulai dari ilmu pengetahuan dan teknologi hingga kesehatan, keuangan, dan banyak lagi, memberikan kesempatan bagi para pengguna untuk menjelajahi dan memanfaatkan data dari berbagai domain untuk keperluan pengembangan dan penelitian.

2.4.2 Phyton

Gambar 2. 4 Logo Python

Python, bahasa pemrograman yang diinisiasi pada tahun 1990 oleh Guido van Rossum di Belanda, telah menjadi salah satu bahasa yang sangat populer dalam proyek pengolahan data dan pengembangan aplikasi. Dikenal karena kemudahan penggunaannya, Python telah memperoleh popularitas yang besar baik di kalangan industri maupun lingkungan akademis. Keunggulan bahasa ini terletak pada kemampuannya untuk diinterpretasikan dan dieksekusi oleh komputer dengan mudah serta dapat diakses secara gratis[31]. Python menawarkan berbagai struktur data tingkat tinggi seperti *array*, *dynamic binding*, *class*, *exceptions*, *list*, dan fitur lainnya yang mempermudah pengembangan perangkat lunak. Logo yang menjadi representasi dari bahasa pemrograman Python juga dapat ditemukan dalam berbagai gambar yang menggambarkan keberadaannya dalam dunia pemrograman.

Python memiliki sejumlah keunggulan dan fitur yang membedakannya dari bahasa pemrograman lainnya[32]. Pertama, bahasa Python dikenal karena kesederhanaannya dengan baris kode yang cenderung lebih singkat dibandingkan dengan bahasa pemrograman lainnya. Struktur penulisan kode *Python* yang menyerupai Bahasa Inggris membuatnya lebih mudah dipahami oleh manusia. Selain itu, sebagai bahasa pemrograman *open-source*, Python mendapat dukungan luas dari berbagai *platform* dan sistem operasi seperti Windows, Linux,

dan Mac OS. Keunggulan lainnya terletak pada beragam *library* yang luas yang dimilikinya. Python menyediakan modul-modul untuk berbagai keperluan, mulai dari pemrosesan teks, jenis data, perhitungan matematika, hingga pengembangan perangkat lunak, dan bahkan model *machine learning*, yang menjadikannya pilihan yang sangat fleksibel untuk berbagai jenis pengembangan perangkat lunak dan kebutuhan pemrograman lainnya.

2.4.3 Google Colaboratory

Gambar 2. 5 Logo Google Colaboratory

Google Colab, singkatan dari Google Colaboratory, adalah layanan *cloud computing* gratis yang disediakan oleh Google. Ini merupakan lingkungan pengembangan berbasis *cloud* yang memungkinkan pengguna untuk menulis dan mengeksekusi kode Python melalui *browser web* tanpa memerlukan pengaturan tambahan atau pengaturan lingkungan lokal. Google Colab memanfaatkan infrastruktur Google Cloud dan menyediakan akses ke GPU dan TPU (Tensor Processing Unit) secara gratis untuk melatih model *machine learning* atau menganalisis data yang membutuhkan komputasi intensif. Ini sangat berguna bagi pengembang atau peneliti yang ingin berkolaborasi pada proyek dengan berbagi kode, data, dan hasil analisis secara real-time[20].

Google Colab ialah sebuah lingkungan pengembangan berbasis *cloud* yang memfasilitasi pengguna dalam melakukan pemrograman dan analisis data dengan menggunakan Python. *Platform* ini dilengkapi dengan beragam pustaka Python seperti Pandas, Matplotlib, dan Plotly yang dapat digunakan untuk memanipulasi data serta membuat

visualisasi. Kelebihan Google Colab terletak pada kemudahannya dalam membuat visualisasi data, karena tidak memerlukan instalasi perangkat lunak di komputer pengguna[33].

2.5 Penelitian Terdahulu

Tabel 2. 1 Penelitian Terdahulu

No	Judul	Penulis	Jurnal	Hasil
1	Komparasi *Machine Learning* Memprediksi *Phising* Dalam Keamanan *Website* [34]	Aswan Supriyadi Sunge	Sains dan Teknologi Vol.1 No.1 Tahun 2022	Hasil klasifikasi dari mendeteksi *phishing* dalam keamanan web dengan menggunakan 5 algoritma yaitu *decision tree, naïve bayes, NN, KNN,* dan *SVM* didapatkan hasil bahwa Neural Network lebih tinggi untuk mendeteksi *phishing* dengan akurasi 95,24% dan *naïve bayes* dengan akurasi terendah yaitu 72,56%.
2	Analisis Komparasi Algoritma Klasifikasi *Data Mining* Dalam Klasifikasi *Website Phishing* [2]	Nabila Bianca Putri, Arie Wahyu Wijayanto	Komputika: Jurnal Sistem Komputer Volume 11, Nomor 1, April 2022	Hasil dari klasifikasi algoritma *naïve bayes, decision tree, random forest,* dan *svm* nilai akurasi yang terbaik adalah algoritma *random forest* sebesar 90,77% dan terendah adalah *naïve bayes* dengan akurasi 82,31%.

No	Judul	Penulis	Jurnal	Hasil
3	*Website Phishing Detection Aplication Using Support Vector Machin* [35]	Diki Wahyudi, Muhammad Niswar, A. Ais Prayogi Alimuddin	*Journal of Information Technology and Its Utilization*, Volume 5, Issue 2, Juny-2022	Setelah dilakukan pengujian menggunakan algoritma *SVM, Decision Tree,* dan *KNN*, hasil terbaik yang diperoleh adalah *SVM kernel polynomial* dengan akurasi 85.71% dan algoritma *decision tree* dengan akurasi terendah yaitu 77.03%.
4	*Comparative Analysis of Phishing Website Prediction Classification Algorithm Using Logistic Regressio, Decision Tree, and Random Forest*[36]	Muhammad Fandru Al Rifqi, Mauli Dina, Anita, Marlince N.K.Nababa, Siti Aisyah	Jurnal Infokum, Volume 10, No.2, Juni2022	Hasil dari klasifikasi untuk prediksi situs *web phishing* menggunakan algoritma *logistic regression, decision tree*, dan *random forest*, hasil menunjukkan bahwa algoritma *random forest* adalah yang terbaik dalam melakukan klasifikasi dengan akurasi mencapai 97.10%, algoritma *decision tree* menduduki peringkat kedua dengan akurasi 94.57%, dan *logistic regression* yang memiliki tingkat akurasi terendah, yakni 92.76%.
5	Identifikasi *Website Phishing* dengan Perbandingan Algoritma Klasifikasi [37]	Agung Susilo Yuda Irawan, Nono Heryana, Hopi Siti Hopipah, Dyas Rahma Putri	Jurnal Informatika Vol. 10 No. 01, 2021 57-67	Penelitian tentang mengidentifikasi situs *phishing* membandingkan empat jenis algoritma: *Support Vector Machine, Decision Tree, Random Forest,* dan *Multilayer Perceptron*. Hasilnya menunjukkan bahwa kinerja algoritma tersebut cukup baik. Algoritma *Multilayer Perceptron* menjadi yang paling unggul dengan akurasi mencapai 93.15% dan nilai *AUC* sebesar 0.976.

Berdasarkan penelitian sebelumnya, yakni [2], [33] hingga [36] terdapat beberapa pendekatan yang berfokus pada pendeteksian *URL website* menggunakan metode klasifikasi seperti *naïve bayes, decision tree, random forest, KNN, SVM, logistic regression*. Perbedaan yang terdapat pada penelitian ini yaitu penelitian ini akan dilakukan menggunakan metode dari algoritma yang memiliki hasil terbaik berdasarkan penelitian sebelumnya, serta penelitian ini menggunakan *dataset* yang berbeda dari penelitian terdahulu, dimana *dataset* yang digunakan diambil dari *website* resmi Kaggle.com yang telah diperbarui 4 tahun lalu.