Penerapan Algoritma Naïve Bayes Classifier Untuk Meningkatkan Keamanan Data Dari Website Phising

Agus Fatkhurohman[1], Eli Pujastuti[2]
[1]*Sistem Informasi Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta*
[2]*Informatika Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta*
Jl. Ring Road Utara Condong Catur Depok Sleman Yogyakarta 55281 INDONESIA
[1]agusfatkhurohman@amikom.ac.id, [2]eli@amikom.ac.id

INTISARI

Perkembangan zaman di era modern ini sudah memasuki era revolusi industri 4.0, dimana segala aspek sudah bergantung pada sebuah teknologi. Dimana bentuk teknologi ini sudah bergerak secara otomatis dan saling terhubung dengan jaringan internet. Teknologi yang digunakan sudah bergerak otomatis dan sudah banyak yang menggunakan sebuah system cerdas. Berbicara mengenai jaringan internet dan system cerdas, maka keterkaitan dengan sebuah data yang digunakan akan semakin besar bahkan bisa sampai tergolong data dengan kategori big data. Yang mana tempat penyimpanan data juga harus disesuaikan dengan kebutuhan datanya. Selain dengan kebutuhan data yang semakin besar di era revolusi industry ini yang selalu dikaitkan dengan jaringan internet maka keamanan sebuah data juga perlu dipertanyakan. Secara tidak langsung keamanan data juga bisa menjadi masalah besar. Dengan jaringan internet setiap orang di setiap dunia bisa mengakses sebuah data yang terkoneksi dengan jaringan internet. Tidak asing lagi sekarang sudah banyak kasus pencurian data karena terhubung dengan jaringan internet. Kasus pencurian data ini sering memanfaatkan website sebagai sarana untuk mencuri datanya yang sering disebut dengan istilah website phising.

Phishing masih menjadi vektor serangan teratas yang memberi akses ke penyerang untuk membuka informasi pribadi seperti kredensial login dan nomor kartu kredit. Pada 2017 lalu, Indonesia menempati urutan ke-9 jumlah serangan phising yakni satu phishing insiden per 2.380 email. Laporan terbaru F5 yang bertajuk menyebutkan, insiden penipuan (fraud) pada Oktober, November, dan Desember melonjak 50% dari rata-rata tahunan. Berpura-pura menjadi seseorang atau entitas yang terkenal adalah taktik utama. Sebanyak 71% serangan phishing pada periode 1 september – 31 Oktober 2018 menggunakan modus mengaku dari perusahaan terkenal, khususnya di industri teknologi.

***Kata kunci** : jaringan, internet, sistem cerdas, website phising, serangan.*

ABSTRACT

The development of the era in this modern era has entered the era of industrial revolution 4.0, where all aspects have depended on a technology. Where this form of technology has moved automatically and interconnected with internet networks. The technology used has moved automatically and many have used an intelligent system. Talking about the internet network and intelligent systems, the linkages with the data used will be even greater and can even be classified as data with the big data category. Which is where the data storage must also be adjusted to the data requirements. In addition to the increasing data requirements in the industrial revolution era which are always associated with the internet network, the security of a data also needs to be questioned. Indirectly data security can also be a big problem. With the internet network everyone in every world can access a data that is connected to the internet network. No stranger now there are many cases of data theft because it is connected to the internet network. This data theft case often uses websites as a means to steal data which is often referred to as phishing websites.

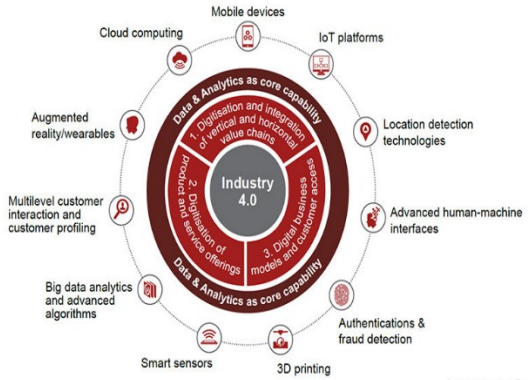
Phishing is still the top attack vector that gives attackers access to private information such as login credentials and credit card numbers. In 2017, Indonesia ranked 9th in the number of phishing attacks, one phishing incident per 2,380 emails. The latest F5 report entitled, fraud in October, November and December increased by 50% from the annual average. Pretending to be someone or a famous entity is the main tactic. As many as 71% of phishing attacks in the period 1 September - 31 October 2018 use the mode claimed to be from a well-known company, especially in the technology industry.

***Keywords:** network, internet, intelligents systems, website phishing, attack.*

I. PENDAHULUAN

Berbicara mengenai perkembangan teknologi seperti sekarang ini, bisa dilihat di negara Indonesia sangat berkembang dengan pesat dan cepat. Terlepas dari hal itu sekarang sudah banyak yang membicarakan tentang revolusi industry 4.0 yang makin marak menjadi topik utama perbicangan. Revolusi industry ini merupakan perubahan secara besar-besaran yang menyangkut dari segala aspek pokok dalam segala bidang, yang mana aspek tersebut menjadi ukuran berjalannya sebuah sistem. Mau tidak mau kita juga harus bisa mengikuti perkembangan revolusi industry agar tetap selalu bisa bersaing dan bertahan.

Revolusi industry ini berawal dari revolusi industry yang pertama kali atau yang sering disebut dengan revolusi industry 1.0, dimana pertama kali ditemukanya mesin uap kemudian disusul revolusi industry 2.0, 3.0 dan yang sekarang ini kita sudah memasuki era revolusi industry 4.0, dimana segala teknologi yang digunakan sudah berbasis computer, berbasis internet, berbasis jaringan dengan perangkat komputasi yang bekerja secara otomatis bahkan sampai menerapkan system cerdas. Segala bentuk teknologi yang digunakan sudah menganut system IoT(Internet of Things) dan untuk revolusi industry yang ke 4.0 bukan sekedar IoT biasa melainkan IoT(Industrial Internet of Things). Revolusi Industri 4.0 ini dapat diilustrasikan dengan gambar di bawah ini.



Gambar 1. Gambaran Umum Revolusi Industri 4.0*(dikutip dari : https://mobnasesemka.com/apa-itu-industri-4-0/)*

Dilihat dari ilustrasi gambar diatas mengenai revolusi industry 4.0 yang mana segala aspek sudah terubung satu sama lain secara otomatis, maka tidak mungkin kalau tidak melibatkan persoalan data dari segala aspek tersebut. Mengenai data tersebut otomatis juga akan banyak data yang digunakan sehingga dapat terjadi kemungkinan penumpukan data sampai melebihi kapasitas. Dengan kata lain persoalan sebuah data akan menjadi factor utama yang harus diperhatikan dengan baik, karena sebuah system dapat berjalan baik dan benar salah satunya unsur utamanya adalah data.

Maka dalam hal ini kemanan sebuah data menjadi sebuah permasalahan yang sangat penting. Bagaimana data tersebut dapat digunakan dengan baik, bagaimana cara mengatasi sebuah keamanan datanya dan bagaimana cara-cara menangani jika sampai terjadi pencurian sebuah data. Dari sebab itu untuk mengamankan sebuah data perlu dilakukan sebuah penelitian dengan metode naïve bayes classifier untuk menjaga sebuah keamanan data.

II. METODOLOGI PENELITIAN

A. Metode Eksperimental

Untuk menunjang sebuah penelitian ini agar berjalan sesuai dengan rencana dengan menghasilkan hasil yang tepat dan akurat maka perlu adanya beberapa strategi untuk penyelesiaanya, antara lain penggunaan metode eksperimental, yaitu metode penelitian yang digunakan untuk mencari pengaruh perlakuan tertentu terhadap yang lain dengan kondisi yang terkendalikan. Dengan beberapa langkah yang telah disusun berdasarkan metode eksperimental seperti yang dijelaskan dalam gambar di bawah ini.

Gambar 2. Bagan alur penelitian eksperimental

Gambar tersebut menjelaskan urutan langkah-langkah penelitian ekspeimental yang dilakukan dalam penelitian ini. Kemudian

untuk melakukan pengolahan datanya akna menggunakan algoritma naïve bayes, agar dapat diperoleh sebuah hasil kesimpulan yang terbaik.

B. Website Phising

Phising adalah sebuah tindakan criminal untuk mencuri informasi pribadi orang lain menggunakan entitas electronic, salah satunya adalah website[1]. Informasi ini dicuri dari website yang telah diakses yang mengandung phising atau dengan kata lain masuk ke dalam kategori website phising. Dikatakan website phising jika suatu website tersebut memenuhi kriteria atau karakteristik phising. Karakteristik phising tersebut digolongkan menjadi empat golongan utama yaitu, Address Bar based Feature, Abnormal based Feature, HTML and Javascript based Features dan Domain based Feature[2]. Penjelasan dari empat jenis karakteristik tersebut terdapat dalma penjelasan di bawah ini.

1. Address Bar Based

Pada Address Bar Based Feature ini terdapat 12 feature seperti yang dijelaskan di bawah ini :

1.1 Using the IP Address

Sebagai contoh untuk hal ini misal jika sebuah alamat IP Address yang digunakan sebagai alternative nama domain dalam URL seperti http://125.98.3.123/fake.html ini akan mengindikasi bahwa ada upaya seseorang untuk mengambil sebuah informasinya.

Rule:

IF

$$\begin{cases} \text{If The Domain Part has an IP Address} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

1.2 Long URL to Hide the Suspicious Part

Sebagai contoh untuk hal ini phising dapat menggunakan almat URL yang panjang untuk menyembunyikannya seperti *"http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html"* dari data alamat tersebut akan menghitung berapa panjang URL dan akan menghasilkan panjang rata-rata.

Rule:

IF

$$\begin{cases} URL\ length < 54 \rightarrow feature = \text{Legitimate} \\ else\ if\ URL\ length \geq 54\ and\ \leq 75 \rightarrow feature = Suspicious \\ otherwise \rightarrow feature = \text{Phishing} \end{cases}$$

1.3 Using URL Shorthening Service "Tiny URL"

Penyingkatan URL atau pemendekan URL adalah metode untuk membuat URL menjadi lebih kecil panjangnya akan tetapi tetap mengarah ke alamat yang dituju, misalnya http://portal.hud.ac.uk/ dapat disingkat menjadi "bit.ly/19DXSk4".

$$\underline{Rule}: \text{IF} \begin{cases} \text{TinyURL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

1.4 URL's having "@" Symbol

Penggunaan symbol @ di dalam URL membuat browser mengabaikan segala sesuatu yang mendahului symbol @ dan alamat aslinya sering mengikuti symbol @

Rule:

$$\text{IF} \begin{cases} \text{Url Having @ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

1.5 Redirect using"//"

Penggunaan tanda"//" dalam URL berarti pengguna nantinya akan diarahkan ke situs lain.

Rule:

IF

{(ThePosition of the Last Occurrence of ""//" " in the URL > 7→ Phishing@Otherwise→ Legitimate)}

1.6 Adding Prefix or Suffix Separated by(-) to the Domain

Simbol tanda hubung (-) jarang digunakan untuk oenamaan sebuah URL. Untuk tujuan Pisher cenderung akan menambahkan awlan atau sufiks yang dipisahkan oleh (-) ke nama domain sehingga pengguna merasa sudah masuk ke alamat yang benar. Misalnya http://www.Confirme-paypal.com/.

Rule: IF

$$\begin{cases} \text{Domain Name Part Includes } (-)\text{Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

1.7 Sub Domain and Multi Sub Domain

Terkadang penamaan suatu alamat website ada yang menggunakan sub domain ataupun muti domain. Jadi teknik untuk phising website ini juga memanfaatkan nama sub domain ataupun multi domain sehingga seolah-olah terlihat seperti turunan dari website aslinya.

Rule:IF

$$\begin{cases} \text{Dots In Domain Part} = 1 \rightarrow \text{Legitimate} \\ \text{Dots In Domain Part} = 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$$

1.8. HTTPS(Hyper Text Transfer Protocol with Secure Sockets Layer)

Penggunaan HTTPS sangatlah penting dalam URL, dengan begitu legalitas sebuah situs web akan semakin diakui keasliannya.

Rule: IF

{(Use https and Issuer Is Trusted &and Age of Certificate ≥ 1 Years → Legitimate@ Using https and Issuer Is Not Trusted → Suspicious@Otherwise→ Phishing)

1.9 Domain Registration Length

Berapa lama usia domain suatu website ini juga berpengaruh dalam mendeteksi apakah suatu website termasuk website phising atau tidak. Sejauh ini umur website sebagai pelaku phising tidak lebih dari satu tahun.

Rule: IF $\begin{cases} \text{Domains Expires on} \leq 1 \text{ years} \\ \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

1.10 Favicon

Favicon adalah gambar grafik (ikon) yang dikaitkan dengan halaman web tertentu. Banyak agen pengguna yang ada seperti browser grafis dan pembaca berita menunjukkan favicon sebagai pengingat visual dari identitas situs web di bilah alamat.

Rule: IF

$\begin{cases} \text{Favicon Loaded From External Domain} \\ \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

1.11 Using Non-Standart Port

Fitur ini berguna dalam memvalidasi jika layanan tertentu naik atau turun di server tertentu. Untuk mengendalikan intrusi, lebih baik membuka port yang Anda butuhkan saja. Beberapa firewall, server Proxy dan Network Address Translation (NAT) akan, secara default, memblokir semua atau sebagian besar port dan hanya membuka yang dipilih.

Rule:

IF $\begin{cases} \text{Port \# is of the Preffered Status} \\ \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

1.12 The Existance of "HTTPS" Token in the Domain Part of thr URL

Phisher dapat menambahkan token "HTTPS" ke bagian domain dari URL untuk mengelabui pengguna. Sebagai contoh, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/

Rule: IF

$\begin{cases} \text{Using HTTP Token in Domain Part} \\ \text{of The URL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

2. Abnormal Based Features

Pada Based Features ini ada 6 features yang dijelaskan di bawah ini :

2.1 Request URL

URL Permintaan memeriksa apakah objek eksternal yang terkandung dalam halaman web seperti gambar, video, dan suara dimuat dari domain lain.

Rule: IF

$\begin{cases} \text{\% of Request URL} < 22\% \rightarrow \text{Legitimate} \\ \text{\%of Request URL} \geq 22\% \text{ and } 61\% \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{feature} = \text{Phishing} \end{cases}$

2.2 URL of Anchor

Fitur ini diperlakukan persis seperti "URL Permintaan". Dengan begitu ada beberapa hal yang harus diperhatikan antara lain :

Rule:

IF $\begin{cases} \text{\% of URL Of Anchor} < 31\% \rightarrow \textit{Legitimate} \\ \text{\% of URL Of Anchor} \geq 31\% \text{ And} \leq 67\% \\ \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$

2.3 Links in <Meta>, <Script> and <Link> tags

Web yang sah menggunakan tag <Meta> untuk menawarkan metadata tentang dokumen HTML; Tag <Script> untuk membuat skrip sisi klien; dan tag <Link> untuk mengambil sumber daya web lainnya

Rule: IF

{(% of Links in "<Meta>","<Script>" and "<"Link>\""<17% → Legitimate@% of Links in <Meta>","<Script>" and "<"Link>\"" ≥17% And≤ 81% → Suspicious @Otherwise→ Phishing)}

2.4 Server Form Handler (SFH)

SFH yang berisi string kosong atau "about: blank" dianggap meragukan karena tindakan harus diambil atas informasi yang disampaikan.

Rule: IF

$\begin{cases} \text{SFH is "about: blank" Or Is Empty} \rightarrow \text{Phishing} \\ \text{SFH Refers To A Different Domain} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$

2.5 Submitting Information to Email

Formulir web memungkinkan pengguna untuk mengirimkan informasi pribadinya yang diarahkan ke server untuk diproses. Phisher mungkin mengarahkan informasi pengguna ke email pribadinya.

Rule: IF

{(Using ""mail()\" or \"mailto:\" Function to Submit User Information" → Phishing@Otherwise → Legitimate)}

2.6 Abnormal URL

Fitur ini dapat diekstraksi dari database WHOIS. Untuk situs web yang sah, identitas biasanya merupakan bagian dari URL-nya

Rule: IF

$$\begin{cases}\text{The Host Name Is Not Included In URL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

3. HTML and JavaScript based Features

3.1 Website Forwarding

Garis tipis yang membedakan situs web phishing dari yang sah adalah berapa kali situs web telah dialihkan

Rule: IF

{(#ofRedirect Page $\leqslant 1 \rightarrow$ Legitimate@#of Redirect Page $\geqslant 2$ &And<4 $\rightarrow$ Suspicious@Otherwise $\rightarrow$ Phishing)}

3.2 Status Bar Customization

Phisher dapat menggunakan JavaScript untuk menampilkan URL palsu di bilah status kepada pengguna.

Rule:

$$\text{IF}\begin{cases}\text{onMouseOver Changes Status Bar} \rightarrow \text{Phishing} \\ \text{It Does}'\text{t Change Status Bar} \rightarrow \text{Legitimate}\end{cases}$$

3.3 Disabling Right Click

Phisher menggunakan JavaScript untuk menonaktifkan fungsi klik kanan, sehingga pengguna tidak dapat melihat dan menyimpan kode sumber halaman web.

Rule: IF

$$\begin{cases}\text{Right Click Disabled} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

3.4 Using Pop-up Window

Tidak lazim menemukan situs web yang sah meminta pengguna untuk mengirimkan informasi pribadi mereka melalui jendela sembulan.

Rule: IF

$$\begin{cases}\text{Popoup Window Contains Text Fields} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

3.5 IFrame Redirection

IFrame adalah tag HTML yang digunakan untuk menampilkan halaman web tambahan menjadi yang saat ini ditampilkan. Phisher dapat menggunakan tag "iframe" dan membuatnya tidak terlihat, yaitu tanpa bingkai batas.

Rule: IF $\begin{cases}\text{Using iframe} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$

4. Domain based Features

4.1 Age of Domain

Fitur ini dapat diekstraksi dari database WHOIS (Whois 2005). Sebagian besar situs web phishing hidup dalam waktu singkat.

Rule:

$$\text{IF}\begin{cases}\text{Age Of Domain} \geq 6 \text{ months} \rightarrow \text{Legitimate} \\ \text{Otherwise} \rightarrow \text{Phishing}\end{cases}$$

4.2 DNS Record

Untuk situs web phishing, baik identitas yang diklaim tidak diakui oleh database WHOIS (Whois 2005) atau tidak ada catatan yang ditemukan untuk nama host (Pan dan Ding 2006)

Rule: IF

$$\begin{cases}\text{no DNS Record For The Domain} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

4.3 Website Traffic

Fitur ini mengukur popularitas situs web dengan menentukan jumlah pengunjung dan jumlah halaman yang mereka kunjungi.

Rule:

$$\text{IF}\begin{cases}\text{Website Rank} < 100{,}000 \rightarrow \text{Legitimate} \\ \text{Website Rank} > 100{,}000 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phish}\end{cases}$$

4.4 Page Rank

PageRank adalah nilai mulai dari "0" hingga "1". PageRank bertujuan untuk mengukur seberapa penting suatu halaman web di Internet. Semakin besar nilai PageRank, semakin penting halaman web.

Rule:

$$\text{IF}\begin{cases}\text{PageRank} < 0.2 \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

4.5 Google Index

Fitur ini memeriksa apakah suatu situs web ada dalam indeks Google atau tidak.

Rule:

$$\text{IF}\begin{cases}\text{Webpage Indexed by Google} \rightarrow \text{Legitimate} \\ \text{Otherwise} \rightarrow \text{Phishing}\end{cases}$$

4.6 Number of Links Pointing to Page

Jumlah tautan yang menunjuk ke halaman web menunjukkan tingkat legitimasinya, bahkan jika beberapa tautan memiliki domain yang sama (Dean, 2014)

Rule:

$$\text{IF}\begin{cases}\text{Of Link Pointing to The Webpage} = 0 \rightarrow \text{Phishing} \\ \text{Of Link Pointing to The Webpage} > 0 \text{ and} \leq 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

4.7 Statistical-Reports Based Feature

Rule: IF

$$\begin{cases} \text{Host Belongs to Top Phishing IPs or} \\ \text{Top Phishing Domains} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

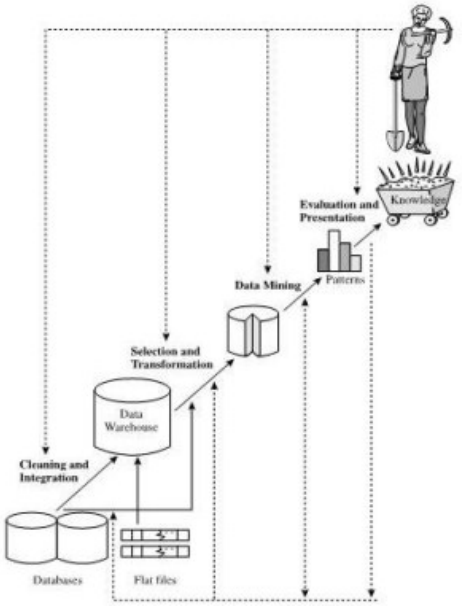
C. Internet Of Things (IoT)

Internet of Things merupakan perkembangan keilmuan yang sangat menjanjikan untuk mengoptimalkan kehidupan berdasarkan sensor cerdas dan peralatan pintar yang bekerjasama melalui jaringan internet[3]. Sejak mulai dikenalnya internet pada tahun 1989, mulai banyak hal kegiatan melalui internet, dan pada tahun 1990 John Romkey menciptakan 'perangkat' pemanggang roti yang bisa dinyalakan dan dimatikan melalui Internet. Kemudian WearCam diciptakan pada tahun 1994 oleh Steve Mann dan ada tahun 1997 Paul Saffo memberikan penjelasan singkat pertama tentang sensor dan masa depan[4]. Selain itu juga masih banyak lagi kegunaan internet yang berkaitan dengan IoT.

D. Data Mining

Data Mining adalah proses penemuan keteraturan pola, dan hubungan dalam set data berukuran besar. Data yang dapat dianalisa dengan data mining bisa dari database, data warehouse, web, repositori informasi atau data yang dapat dialirkan ke dalam sistem secara dinamis[5]. Hasil dari pengolahan data dengan metode data mining ini dapat digunakan untuk mengambil keputusan di masa depan. Data mining ini juga dikenal dengan istilah pattern recognition[10]. Seorang pakar menyebutkan bahwa KDD atau Knowledge Discovery from Data, merupakan proses terstruktur, yaitu sebagai berikut[11]:

1. Data Cleaning adalah Proses membersihkan data dari data noise dan tidak konsisten.
2. Data Integration adalah Proses untuk menggabungkan data dari beberapa sumber yang berbeda.
3. Data Selection adalah Proses untuk memilih data dari database yang sesuai dengan tujuan analisis.
4. Data Transformation adalah Proses mengubah bentuk data menjadi data yang sesuai untuk proses Mining.
5. Data Mining adalah Proses penting yang menggunakan sebuah metode tertentu untuk memperoleh sebuah pola dari data.
6. Pattern Evaluation adalah Proses mengidentifikasi pola.
7. Knowledge Presentation adalah yang dapat merepresentasikan informasi yang dibutuhkan, proses dimana informasi yang telah didapatkan kemudian digunakan oleh pemilik data.



Gambar 3. Data Mining sebagai dari proses *knowledge discovery*[12]

Gambar 3 menunjukkan proses penjelajahan pengetahuan dimulai dari beberapa database dilakukan proses cleaning dan integration sehingga menghasilkan data warehouse. Dilakukan proses selection dan transformation yang kemudian disebut sebagai data mining hingga menemukan pola dan memperoleh pengetahuan dari data (knowledge)[12].

E. Naïve Bayyes Classifier

Naive Bayes classifier (NBC) merupakan salah satu metoda pembelajaran mesin yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Teori Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifierlainnya. Hal ini dibuktikan oleh Xhemali, Hinde dan Stone dalam jurnalnya "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages" mengatakan bahwa "Naïve Bayes Classifier memiliki tingkat akurasi yang lebih baik disbanding model classifier lainnya.

Metode naïve bayes classifier akan mencari beberapa kemungkinan aman atau tidaknya situs web yang kita akses tersebut. Dengan begitu kita bisa menjaga keamanan data kita dengan tidak mengakses web yang sudah terdeteksi sebagai web phising. Metode naïve

bayes ini dipilih karena dirasa paling tepat untuk menyelesaikan masalah deteksi website phising. Metode ini mampu menyeleseksi data dengan cara mengklasifikasikan sekumpulan data dengan memanfaatkan probabilitas dan statistic. Dimana probabilitas yang digunakan yaitu dengan menggunakan prediksi probabilitas masa depan dengan dasar-dasar masa sebelumnya. Klasifikasi Naive Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Persamaan dari teorema Bayes adalah :
Persamaan dari teorema Bayes adalah[6] :

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad \text{..... (1)}$$

Dimana :
X : Data dengan class yang belum diketahui
H : Hipotesis data merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X(posteriori probabilitas)
P(H) : Probabilitas hipotesis H(prior probabilitas)
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X

Untuk menjelaskan metode naïve bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok untuk sampel yang sedang dianalisis tersebut. Oleh karena itu metode naïve bayes diatas disesuaikan swbagai berikut :

$$P(C|F1\ldots Fn) = \frac{P(C)P(F1\ldots.Fn|C)}{P(F1\ldots.Fn)} \quad \text{.....(2)}$$

Dimana variable C merepresentasikan kelas, smenetara variable F1…Fn merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumurs tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C(Posterior) adalah peluang munculnya kelas C(sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C(disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global(disebut juga evidence). Karena itu, rumus diatas dapat pula ditulis secara sederhana sebagai berikut :

$$\text{Posterior} = \frac{\text{Prior x likelihood}}{\text{Evidence}} \quad \text{.....(3)}$$

Nilai evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dai posterior tersebut nantinya akan dibandingkan dengan nilai-nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus bayes tersebut dilakukan dengan menjabarkan (C|F1,..Fn) menggunakan aturan perkalian sebagai berikut :

$$\begin{aligned} & P(C|F1, \ldots Fn = P(C)P(F1\ldots.Fn|C) \\ = \; & P(C)P(F1|C)P(F2,\ldots Fn|C,F1) \\ = \; & P(C)P(F1|C)P(F2|C,F1)P(F3,\ldots Fn|C,F1,F2) \\ = \; & (C)P(F1|C)P(F2|C,F1)P(F3|C,F1,F2) \\ & P(F4,\ldots Fn|C,F1,F2,F3) \\ = \; & P(C)P(F1|C)P(F2|C,F1)P(F3|C,F1,F2)\ldots \\ & P(Fn|C,F1,F2,F3\ldots Fn\text{-}1) \quad \text{.....(4)} \end{aligned}$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya factor-faktor syarat yang mempengaruhi nilai probabilitas, yang hamper mustahil untuk dianalisa satu persatu. Akibatnya perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi(naif), bahwa masing-masing petunjuk (F1,F2..Fn) saling bebas(independent) satu sama lain. Dengan asumsi tersebut maka berlaku satu kesamaan sebaai berikut :

$$P(F_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i) \qquad (5)$$

Untuk i≠j , sehingga

$$P(F_i|C,F_j) = P(F_i|C) \qquad (6)$$

Gambar 4. Persamaan ke 5 dan 6 untuk pehitungan naïve bayes

Persamaan diatas merupakan model dari teorema naïve bayes yang selanjutnya akan digunakan dalam proses klasifikasi. Untuk klasifikasi dengan data kontinyu digunakan rumus Densitas Gauss :

$$P(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2_{ij}}} \qquad (7)$$

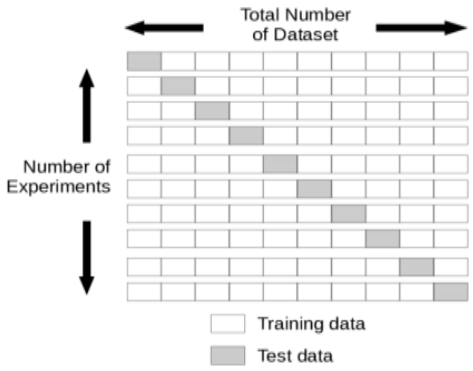
Gambar 4. Rumus Densitas Gauss

Dimana :
P : Peluang
Xi : Atribut ke i
xi : nilai atribut ke i
Y : kelas yang dicari
yi : Sub kelas Y yang dicari

u : mean, menyatakan rata-rata seluruh atribut
q : Deviasi standar, menyatakan varian dari seluruh atribut

F. 10-Folds Cross Validation

Setelah data telah dibagi menjadi 80% data training dan 20% data testing, maka akan dilakukan 10-fold cross validation pada data training. Cross Validation adalah teknik untuk mengevaluasi model dengan cara mempartisi sampel asli ke dalam training set untuk melatih model, dan test set untuk mengevaluasi model. Dalam k-fold cross validation, sampel asli secara acak dipartisi dalam k equal size subsample. Dari subsample k, satu subsample akan digunakan sebagai testing data dan sisanya akan menjadi training data. Proses cross validation akan diulang sebanyak k kali (kelipatan), dengan masing – masing dari subsample k digunakan sekali sebagai validation data[7].



Gambar 5. 10 Folds Cross Validation

K-fold cross validation merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. K fold cross validation diawali dengan membagi data sejumlah n-fold yang diinginkan. Dalam proses cross validation data akan dibagi dalam n buah partisi dengan ukuran yang sama D1, D2, D3.. Dn selanjutnya proses uji dan latih dilakukan sebanyak n kali. Dalam iterasi ke-i partisi. Di akan menjadi data uji dan sisanya akan menjadi data latih. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan 10-fold cross validation dalam model[8]. Skenario pengujian merupakan tahap penentuan pengujian yang dilakukan. Pengujian dilakukan menggunakan metode k-cross validation dengan nilai k sebanyak 10 fold, pengujian ini bertujuan untuk mengetahui akurasi metode naïve bayes classifier yang diterapkan pada analisis dan diuji dengan data training dan data testing yang berbeda . Penggunaan 10 fold ini dianjurkan karena merupakan jumlah fold terbaik untuk uji validitas[9].

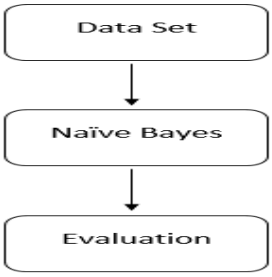
Fold	Data	*Subset*
Fold 1	Training Testing	$S_2, S_3, S_4, S_5, S_6, S_7 S_8, S_9, S_{10}$ S_1
Fold 2	Training Testing	$S_1, S_3, S_4, S_5, S_6, S_7 S_8, S_9, S_{10}$ S_2
Fold 3	Training Testing	$S_1 S_2, S_4, S_5, S_6, S_7 S_8, S_9, S_{10}$ S_3
Fold 4	Training Testing	$S_1 S_2, S_3, S_5, S_6, S_7 S_8, S_9, S_{10}$ S_4
Fold 5	Training Testing	$S_1 S_2, S_3, S_4, S_6, S_7 S_8, S_9, S_{10}$ S_5
Fold 6	Training Testing	$S_1 S_2, S_3, S_4, S_5, S_7, S_8, S_9, S_{10}$ S_6
Fold 7	Training Testing	$S_1 S_2, S_3, S_4, S_5, S_6, S_8, S_9, S_{10}$ S_7
Fold 8	Training Testing	$S_1 S_2, S_3, S_4, S_5, S_6, S_7 S_9, S_{10}$ S_8
Fold 9	Training Testing	$S_1 S_2, S_3, S_4, S_5, S_6, S_7, S_{10}$ S_9
Fold 10	Training Testing	$S_1 S_2, S_3, S_4, S_5, S_6, S_7 S_8, S_9$ S_{10}

Gambar 4. Gambar tabel scenario stabilitas uji validitas 10-cross validation[9]

III. HASIL DAN PEMBAHASAN

Peranan jaringan internet dalam era revolusi industri 4.0 ini sangat besar, karena hampir semua sistem yang digunakan sudah melibatkan jaringan internet. Dari hal tersebut maka tingkat keamanan data yang digunakan juga sangat berpotensi menjadi sasaran untuk tindak kejahatan. Kejahatan dengan menggunakan internet ini biasanya menggunakan perantara website atau yang sering disebut dengan website phising. Dengan cara seperti itu pelaku kejahatan bisa dengan bebas mencuri data yang menjadi incaran. Dari permasalahan tesebut penelitian ini akan membahas penerapan sebuah algoritma naïve bayes calssifier untuk menanggulangi kejahatan website phising agar kemanan data yang kita gunakan terjaga dengan baik.

Penelitian ini adalah penelitian eksperimen dimana penelitian dilakukan dengan menerapkan k-fold cross-validation pada dataset website phising. Data set diambil dari repository UCI Machine Learning. Selanjutnya dari hasil feature reduction, dataset diterapkan pada algoritma machine learning yang popular yaitu naïve bayes untuk diukur tingkat akurasinya. Software yang digunakan pada penlitian ini adalah WEKA, dengan pemodelan alur peneletian seperti di bawah ini :



Gambar 5. Alur Penelitian

A. Analisis Data

Dalam tahapan ini peneliti mengumpulkan beberapa data yang akan diolah dengan sumber data dari UCI Machine Learning. Dengan menggunakan algoritma naïve bayes data tersebut akan diolah untuk mencari hasil dengan tingkat akurasi yang terbaik. Algoritma Naïve Bayes Classifier ini nantinya akan mengidentifikasi sebuah sebuah alamat website yang dicurigai sebagai alamat website phising. Dengan begitu dari sisi pengguna jaringan intenet akan mengetahui bahwa alamat website tersebut aman atau tidak. Dengan menggunakan dataset yang mempunyai jumlah atribut sebanyak 30 atribut yang diambil dari karakteristik phising yang digolongkan dari empat golongan utama yaitu, Address Bar based Feature, Abnormal based Feature, HTML and Javascript based Features dan Domain based Feature. Untuk pengujian Naïve Bayes menggunakan metode k-fold cross-validation untuk mengetahui kinerja algoritma tersebut. Tabel kategorikal atribut dapat dilihat dalam tabel di bawah ini :

TABEL I.
ATRIBUT DAN LABEL

Nilai	Atribut
1 = valid 0 = mencurigakan -1 = phising	having_IP_Address, URL_Length, Shortining_Service, having_At_Symbol, double_slash_redirecting, Prefix_Suffix, having_Sub_Domain, SSLfinal_State, Domain_registeration_length, *Favicon, port,* HTTPS_token, Request_URL, URL_of_Anchor, Links_in_tags, SFH, Submitting_to_email, Abnormal_URL *Redirect,* on_mouseover, RightClick, popUpWidnow, *Iframe,* age_of_domain, DNSRecord, web_traffic, Page_Rank, Google_Index, Links_pointing_to_page, Statistical_report, *Result* (*Label*)

Penggunaan k-fold cross-validation ini nantinya akan menggunakan 10 folds dari total dataset yang ada. Dataset yang dipakai disini ada sekitar 11055 data atribut, yang berarti akan terbagi sekitar 1100 data dalam setiap folds. Dari perhitungan tersebut akan menghasilkan detail akurasi yang diambil sampai dengan F-Measure yang digambarkan hasilnya dengan tabel dibawah ini.

TABEL II.
HASIL PERFORMANCE EVALUATION CROSS VALIDATION

Hasil 10 Folds Cross Validation							
	1	**2**	**3**	**4**	**5**	**6**	**7**
Weight average	0.93 0	0.07 6	0.93 0	0.93 0	0.93 0	92.9 8	7.01

Keterangan Tabel :

Hasil tersebut diatas adalah dalam hitungan persen %

1. TP Rate
2. FP Rate
3. Precision
4. Recall
5. F-Measure
6. Akurasi
7. Error

Dari perhitungan detail akurasi tersebut akan menemukan hasil cross validation dengan tingkat akurasi yang bisa dibilang sangat akurat karena akan memperoleh nilai sebesar 92.98% dan nilai toleransi error yang kecil sebesar 7.01%.

IV. KESIMPULAN

Algoritma Naïve bayes sangat tepat untuk memperhitungkan klasifikasi website phising. Dari dataset yang telah di dapat ditarik kesimpulan dengan hasil sebagai berikut, Hasil dari pengujian algoritma Naive Bayes diperoleh nilai rata-rata akurasi sebesar 92.98% dengan TP Rate yang diperoleh sebesar 0.930%, FP Rate sebesar 0.076%, Precision sebesar 0.930%, Recall sebesar 0.930% dan F-measure sebesar 0.930%. Dengan demikian hasil penerapan algoritma naive bayes tersebut untuk melindungi data dari website phising dikatakan sangat baik, dan penggunaan algoritma tersebut sudah tepat jika digunakan untuk pencegahan pencurian data dari sebuah ancaman website phising.

UCAPAN TERIMA KASIH

Di akhir kata kami selaku peneliti mengucapkan terima kasih sebesar-besarnya kepada seluruh akademisi yang mendukung penelitian kami, baik dari pihak internal akademisi sendiri maupun pihak akademisi yang menerbitkan publikasi jurnal ini khususnya Universitas Respati Yogyakarta selaku penerbit jurnal ini. Tidak lupa kami juga mengucap syukur kepada Tuhan Yang Maha Esa yang telah melimpahkan segala rahmat dan hidayahnya kepada kami sehigga kami dapat menyelesaikan penelitian ini dengan baik, walaupun dari penelitian ini masih ada beberapa kekurangan atau bisa dibilang belum sempurna, maka untuk hasil ke depan yang lebih baik lagi kami mohon kepada semua peneliti agar bisa melanjutkan penelitian kami agar kami dapat mengetahui hasil yang lebih baik lagi. Semoga penelitian ini dapat bermanfaat buat kami sendiri, institusi kami maupun untuk orang lain..

REFERENSI

[1] Susanto Bekti Maryuni, 2016, *"Identifikasi Website Phising Dengan Seleksi Atribut Berbasis Korelasi"*, Seminar Nasional Teknologi dan Komunikasi(SENTIKA), 18-19 Maret 2016

[2] Mohammad, R., McCluskey, T., & Thabtah, F. A. 2012. "An Assesment of Features Related to Phishing Websites using an Automated Technique. International Conference For Internet Technology And Secure Transaction. Ss 492-497. London:ICITST 2012

[3] Keoh, S. L., Kumar, S., & Tschofenig, H. (2014). Securing the Internet of Things: A Standardization Perspective. *IEEE Internet of Things Journal*, *1*(3), 1–1

[4] Junaidi Apri, 2015, *"Internet Of Things, Sejarah, Teknologi, dan Penerapannya : Review"*. Jurnal Ilmiah Teknologi Informasi Terapan (JITTER), Vol 1 No 3, Agustus, 2015

[5] Salim Tomy, Giap Yo Ceng, 2017, *"Data Mining Identifikasi Website Phising Menggunakan Algoritma C4.5"*, Jurnal TAM(Technology Acceptance Model) Volume 8, Desember 2017, hal. 130-135

[6] Saleh Alfa, 2015, *"Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga",* Citec Journal, Vol. 2 No. 3 Mei 2015

[7] Sandag Green Arther, Leopod Jonathan, Ong Vinky Fransiscus, 2018, *"Klasifikasi Malicious Website Menggunakan Algoritma K-NN Berdasarkan Application Layers dan Network Characteristics"*, Cogito Smart Journal, Vol.4 No.1 June

[8] R. Kohavi, *"A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,"*1995.[Online]. Available:http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf [Accessed 1 Februari 2019]

[9] Pitria Pipit, 2016, *"Analisis Sentimen Pengguna Twitter Pada Akun Resmi Samsung Indonesia Dengan Menggunakan Naïve Bayes",* [Online], https://elib.unikom.ac.id/files/disk1/714/jbptunikompp-gdl-pipitpitri-35651-7-unikom_p-a.pdf[akses pada 2 Februari 2019]

[10] Sulastri Heni, Gufroni Acep Irham, 2017, "Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia", Jurnal Nasional Teknologi da Sistem Informasi, Vol. 03 No. 02

[11] Han, J. Kamber M&Jian, Pei, 2011, "Data Mining : Concepts and Techniques, Thrid Edition, America : Morgan Kauffman,San Francisco.

[12] Haryati Siska, Sudarsono Aji, Suryana Eko, 2015, *"Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu),* Jurnal Media Informatika, Vol. 11 No. 2 September