Prediksi Website Pemancing Informasi Penting (Phishing) Menggunakan Algoritma *Support Vector Machine*, *Naïve Bayes, dan Decision Tree*

T E S I S

Diajukan Sebagai Salah Satu Syarat Untuk Menyelesaikan

Program Strata Dua (S2) Magister Komputer

OLEH :

Zuhri Halim

045141221027

PROGRAM STUDI TEKNIK INFORMATIKA

PROGRAM PASCA SARJANA (S2) MAGISTER KOMPUTER

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER

ERESHA JAKARTA

2016

LEMBAR PERSETUJUAN TESIS

Nama	: Zuhri Halim
NPM	: 045141221027
Konsentrasi	: Rekayasa Bisnis
Judul tesis	: Prediksi Website Pemancing Informasi Penting (Phishing) Menggunakan Algoritma *Support Vector Machine, Naïve Bayes, dan Decision Tree.*

Telah disetujui untuk disidangkan pada Sidang Tesis pada Program Pasca Sarjana (S2) Magister Komputer, Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer ERESHA.

Jakarta, 9 Mei 2016

Pembimbing Utama	Pembimbing Pendamping
(Dr. Abu Khalid Rivai, M.Eng)	(Agus Suharto, M.Kom)

Mengetahui :

Ketua	Ketua Program Studi Pasca Sarjana
(Dr. Rasmadi M.Pd)	(Dr. Makhsun, M.Si)

LEMBAR PENGESAHAN TESIS

Nama	: Zuhri Halim
NPM	: 045141221027
Konsentrasi	: Rekayasa Bisnis
Judul tesis	: Prediksi Website Pemancing Informasi Penting (Phishing) Menggunakan Algoritma *Support Vector Machine, Naïve Bayes,* dan *Decision Tree.*

Telah disidangkan dan dinyatakan Lulus Sidang Tesis pada Program Pasca Sarjana (S2) Magister Komputer, Program Studi Teknik Informatika Sekolah Tinggi Manajemen Informatika dan Komputer Eresha pada tanggal 12 Mei 2016.

Nama Penguji	Tanda Tangan
Dr. I Putu Susila, M.Eng.	1. .....
Joko Trianto, M.Kom.	2. .....

Mengetahui :
Ketua Program Studi
Pasca Sarjana

(Dr. Makhsun, M.Si)

LEMBAR PERNYATAAN KEASLIAN TESIS

Nama	: Zuhri Halim
NPM	: 045141221027
Konsentrasi	: Rekayasa Bisnis
Judul tesis	: Prediksi Website Pemancing Informasi Penting (Phishing) Menggunakan Algoritma *Support Vector Machine, Naïve Bayes, dan Decision Tree*

Dengan ini saya menyatakan bahwa dalam Tesis ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar Magister di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Serpong, 12 Mei 2016

Zuhri Halim

Zuhri Halim, 045141221027

Judul tesis : Prediksi Website Pemancing Informasi Penting (Phishing) Menggunakan Algoritma Prediksi Website Pemancing Informasi Penting (Phishing) Menggunakan Algoritma *Support Vector Machine, Naïve Bayes, dan Decision Tree*; di bawah bimbingan Bapak Dr. Abu Khalid Rivai, M.Eng dan Bapak Agus Suharto, M.Kom

ABSTRAK

Perkembangan teknologi informasi dan komunikasi khususnya internet berdampak pada semua sektor kehidupan manusia tidak terkecuali dengan sektor perbankan dan keuangan. Selain memberikan dampak positif dengan dipermudahnya pelanggan dalam proses transaksi yang dapat dilakukan kapanpun dan di manapun tanpa dibatasi oleh ruang dan waktu menggunakan media internet, juga membawa potensi besar terhadap pihak-pihak yang tak bertanggungjawab untuk melakukan pencurian data dan informasi penting, salah satunya dengan teknik *phishing,* sehingga metode untuk mendeteksi serangan situs *phishing* memerlukan perhatian serius. Dalam penelitian ini penulis telah melakukan memberikan gambaran metode yang paling akurat untuk mendeteksi website phishing dengan membandingkan tiga metode antara lain *Support Vector Machine, Naïve Bayes*, dan *Decision Tree* menggunakan dataset publik dari UCI *Machine Learning Repository* (www.uci.edu) yang dioptimasi dengan *feature selection* dan diolah menggunakan program *RapidMiner. Hasil penelitian* menunjukan bahwa metode *Decision Tree* mempunyai tingkat akurasi sebesar 91,84%, metode *Naïve Bayes* sebesar 74,07% dan *Support Vector Machine* sebesar 92,34%. Hal ini menunjukan bahwa metode *Support Vector Machine* mempunyai tingkat akurasi yang paling tinggi.

Kata Kunci: *Phishing, Support Vector Machine, Naïve Bayes, Decision Tree*

Zuhri Halim, 045141221027

Judul tesis : Prediksi Website Pemancing Informasi Penting (Phishing) Menggunakan Algoritma *Support Vector Machine, Naïve Bayes, dan Decision Tree* ; di bawah bimbingan Bapak Dr. Abu Khalid Rivai, M.Eng dan Bapak Agus Suharto, M.Kom

ABSTRACT

The development of information and communication technologies, especially the Internet, have an impact in all sectors of human life with exception in the banking and financial sectors in addition to a positive impact to make essier customer in the transaction process that can do anytime and anywhere without being limited by space and time using the internet, it also brings great potential against parties not responsible for the theft of critical data and information, one of them with phishing techniques, so the method for detecting a phishing site requires serious attention. In this study the authors try to give an overview of the most accurate methods to detect phishing websites to compare three methods such as Support Vector Machine, Naïve Bayes, and Decision Tree using public datasets from the UCI Machine Learning Repository (www.uci.edu) *optimized with feature selection and processed* using RapidMiner program that showed *Decision Tree* has a accuracy rate of 91.84%, Naïve Bayes method amounted to 74.07% and Support Vector Machine by 92.34%. Hereby declare that the method of Support Vector Machine has the highest degree of accuracy.

Keyword: *Phishing, Support Vector Machine, Naïve Bayes, Decision Tree.*

KATA PENGANTAR

Puji syukur kepada Allah Yang Maha dari Segala Maha Sempurna karena hanya dengan berkat dan kemurahan-Nya, tesis yang merupakan syarat kelulusan program pasca sarjana dengan judul Prediksi *Website* Pemancing Informasi Penting (Phishing) Menggunakan Algoritma *Support Vector Machine, Naïve Bayes, dan Decision Tree* dapat diselesaikan dengan baik dan tepat pada waktunya.

Penulisan tesis ini tidak akan berjalan lancar tanpa dukungan dari keluarga dan para sahabat yang senantiasa mendampingi dan mendoakan kelancaran usaha penulis. Tiada ucapan yang dapat mewakili rasa terima kasih penulis kepada ayah, ibu, keluarga dan para sahabat selain sedikit penghargaan dalam tulisan ini.

Penulis juga menyadari bahwa seluruh kegiatan perkuliahan pasca sarjana hingga penulisan tesis ini tidak akan terwujud tanpa dukungan dan bantuan dari berbagai pihak, maka dengan kegembiraan dan kerendahan hati penulis menghaturkan banyak terima kasih kepada:

1. Keluarga, khususnya kedua orang tua dan adik.
2. Bapak Dr. Rasmadi M.Pd selaku Ketua Sekolah Tinggi Manajemen Informatika dan Komputer Eresha.
3. Bapak Dr. Makhsun, M.Si, selaku Ketua Program Studi Teknik Informatika Pasca Sarjana Sekolah Tinggi Manajemen Informatika dan Komputer Eresha.
4. Bapak Dr. Abu Khalid Rivai, M.Eng selaku Pembimbing Utama dan Bapak Agus Suharto, M.Kom sebagai Pembimbing Pendamping. Ketelitian dan kesabaran beliau dalam membimbing, memberi arah dan memudahkan dalam menyelesaikan tesis ini dengan tepat waktu sehingga tesis ini menjadi berarti.
5. Segenap Dosen Penguji, terutama Bapak Dr. I Putu Susila, M.Eng dan Bapak Joko Trianto, M.Kom yang telah memberikan saran dan masukan demi terwujudnya Tesis yang lebih baik.
6. Bapak Adi Wijaya, M.Kom yang telah memotivasi penulis untuk memulai pemilihan judul dan penulisan tesis jauh-jauh hari.

7. Seluruh Karyawan, Dosen dan Pimpinan STMIK Eresha yang telah memberi ilmu pengetahuan berharga serta membantu kelancaran studi selama masa pendidikan.
8. Semua rekan-rekan mahasiswa pasca sarjana STMIK Eresha khususnya Angkatan 45 yang selalu memberikan semangat dan dukungan dalam menempuh ilmu.
9. Semua rekan-rekan alumni pasca sarjana STMIK Eresha yang telah membantu penulis dalam diskusi dan *sharing* sehingga tesis ini dapat selesai.

Saya menyadari, penelitian ini masih jauh dari sempurna oleh karena itu memerlukan banyak masukan dari bapak/ibu sekalian yang membacanya. Semoga tesis ini dapat memberikan wawasan baru bagi pembaca, civitas akademika STMIK Eresha.

Serpong, Mei 2016

Penulis,

DAFTAR ISI

DAFTAR TABEL

Hal

DAFTAR GAMBAR

Hal

BAB I
PENDAHULUAN

1.1. Latar Belakang Penulisan

Perkembangan Ilmu Pengetahuan dan Teknologi (IPTEK), terutama Teknologi Informasi (*Information Technology*) seperti internet sangat menunjang setiap orang mencapai tujuan hidupnya dalam waktu singkat, baik legal maupun illegal dengan menghalalkan segala cara karena ingin memperoleh keuntungan materi atau pun non-materi.

Kemajuan Teknologi Informasi yang serba digital membawa orang berminat ke dalam dunia bisnis yang revolusioner karena dirasakan lebih mudah, praktis, dan dinamis berkomunikasi dan memperoleh informasi. Di sisi lain, berkembangnya Teknologi Informasi menimbulkan pula sisi rawan yang gelap sampai tahap mencemaskan dengan kekhawatiran pada perkembangan tindak pidana dibidang Teknologi Informasi yang berhubungan dengan "*cybercrime*" dan "*cyberlaw*" atau kejahatan dunia maya.

Phishing pertama kali terkenal pada tahun 1996 ketika salah seorang phisher mencuri American Online (AOL) account dengan metode-metode yang dikenal dengan serangan phishing, kata "*phishing*" sendiri berasal dari rentang waktu 1990-an Istilah ini diciptakan berdasarkan analogi yang digunakan untuk menipu seperti kail untuk "phish" *username*, *password* dan informasi sensitif lainnya. Penggunaan huruf "ph" diyakini berasal dari kata "*phreaking*" menurut Antonio San Martino, X. P. (2010).

Berbicara mengenai *phishing* maka akan dikaitkan juga dengan *social engineering* menurut buku dengan judul *No-tech hacking* oleh Johnny Long mengatakan senjata paling penting dalam dunia "*hacker*" adalah *social engineering*, menurut Johnny Long setiap orang harus merubah "*mind set*"nya mengenai social engineering yang merupakan alat bantu untuk mengenali kelemahan dari komunikasi data, yang jika dikaitkan dengan serangan phishing menurut Aboli Bhanji, S. B. (2013), bahkan menurut Lance James and Joe Stewart, Syngress *Publishing* (2005), dalam bukunya yang berjudul *phishing exposed* menjelaskan bahwa serangan *phishing* meledak pada tahun 2005 dan

phishing merupakan cara untuk memikat orang agar mudah jatuh dalam perangkap penipuan seperti halnya memancing, menunggu korban untuk "menggigit" umpan yang telah disediakan dan phishing juga merupakan kombinasi dari *social engineering*, dengan mencari kelemahan di dalam web site dan kelemahan di dalam e-mail, pada dasarnya *phishing* menggunakan hampir semua teknik peretasan yang digunakan untuk membuat umpan.

Phishing (memancing informasi penting) adalah suatu bentuk penipuan yang dicirikan dengan percobaan untuk mendapatkan informasi rahasia, seperti kata sandi dan kartu kredit, dengan menyamar sebagai orang atau bisnis yang tepercaya dalam sebuah komunikasi elektronik resmi, seperti surat elektronik atau pesan instan. Istilah *phishing* dalam bahasa Inggris berasal dari kata *fishing* ('memancing'), dalam hal ini berarti memancing informasi keuangan dan kata sandi pengguna. Dengan banyaknya kasus pengelabuan yang dilaporkan, metode tambahan atau perlindungan sangat dibutuhkan. Upaya-upaya itu termasuk pembuatan undang-undang, pelatihan pengguna, dan langkah-langkah teknis.

Selanjutnya penelitian yang dilakukan oleh He Chunjian dan Zhang Cuilian Zhao Yan dengan judul *A New SVM Merged into Data Information*, dengan metode kernel fungsi dimana beberapa kernel dilatih dan kernel terbaik tampil di set validasi kemudian dipilih untuk pengujian dan kinerjanya dievaluasi pada set tes dan menunjukkan bahwa pendekatan secara efektif dapat meningkatkan klasifikasi akurasi (Chunjiang & Yan, 2009).

Neural Network mempunyai kelebihan dalam hal kemampuan generalisasi tergantung pada seberapa baik *Neural Network* meminimalkan resiko empiris namun *Neural Network* mempunyai kelemahan dimana menggunakan data pelatihan cukup besar (Vapnik, 1999). *Decison tree* dan ID3 mempunyai kelebihan untuk keputusan pengklasifikasi memiliki akurasi yang baik namun memiliki kelemahan karena perlu mengumpulkan lebih banyak data (Han, Rodriguze, & Beheshti, 2008). *Support Vector Machine* adalah kasus khusus dari keluarga algoritma yang kita sebut sebagai *regularized* metode klasifikasi linier dan metode yang kuat untuk minimalisasi resiko (Sholom M. Weiss, Indurkhya, & Zhang, 2010). Dan kelebihan *Support*

Vector Machine lainnya adalah dapat meminimalkan kesalahan melalui memaksimalkan margin dengan misahkan antara *hyper-lane* dan satu set data bahkan dengan jumlah sample yang kecil (Chunjiang & Yan, 2009).

Namun demikian masalah aplikasi tertentu, tidak semua fitur ini sama-sama penting dan kinerja yang lebih baik dapat dicapai dengan membuang beberapa fitur dengan begitu fitur dalam *Support Vector Machine* memiliki pengaruh penting dalam akurasi klasifikasi (Zhao, Fu, Ji, Tang, & Zhou, 2011). Dataset yang tidak penting, fitur yang banyak atau sangat berhubungan secara signifikan akan mengurangi tingkat akurasi klasifikasi dengan menghapus fitur ini, dengan begitu tingkat akurasi efisiensi dan klasifikasi dapat diperoleh (Lin a, Shiue b, & Chen, 2009).

Seleksi fitur adalah terkait erat dengan masalah pengurangan dimensi dimana tujuannya adalah untuk mengidentifikasi fitur dalam kumpulan data-sama pentingnya, dan membuang fitur lain seperti informasi yang tidak relevan dan berlebihan dan akurasi dari seleksinya pada masa depan dapat ditingkatkan (Maimon, 2010). Seleksi fitur adalah salah satu faktor yang paling penting yang dapat mempengaruhi tingkat akurasi klasifikasi karena jika dataset berisi sejumlah fitur, dimensi ruang akan menjadi besar dan non-bersih, merendahkan tingkat akurasi klasifikasi (Liu, Wang, Chen, Dong, Zhu, & Wang, 2011). Masalah dalam seleksi adalah pengurangan dimensi, dimana awalnya semua atribut diperlukan untuk memperoleh akurasi yang maksimal. Empat alasan utama untuk melakukan pengurangan dimensi (Maimon, 2010)

Dalam penelitian ini penulis mencoba memberikan gambaran kinerja prediksi terhadap website *phishing* menggunakan metode *Support Vector Machine* kemudian membandingkannya dengan metode *Naïve Bayes* dan *Decision Tree,* dari perbandingan tersebut diharapkan penelitian ini dapat memberikan gambaran metode yang paling efisien dan akurat dalam memprediksi *website phishing.*

1.2 Identifikasi Masalah

Support Vector Machine dapat menyelesaikan masalah *decision tree* khususnya sampel data yang kecil, tetapi *Support Vector Machine* memiliki kelemahan pada sulitnya pemilihan fitur yang sesuai dan optimal pada bobot atribut yang digunakan sehingga menyebabkan tingkat akurasi prediksi menjadi rendah.

1.3 Rumusan masalah

Rumusan masalah yang diangkat pada penelitian ini adalah seberapa besar akurasi metode *Support Vector Machine, Naïve Bayes, dan Decision Tree* yang ditingkatkan dengan cara melakukan seleksi atribut untuk prediksi *website phishing*?

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah menerapkan seleksi atribut yang optimal dan membobot atribut dari *dataset* pada metode *Support Vector Machine* yang dibandingkan dengan *Naïve Bayes dan Decision Tree* guna meningkatkan akurasi prediksi *phishing* website.

1.5 Manfaat penelitian

1. Manfaat praktis dari hasil penelitian ini adalah dapat digunakan oleh pihak *Designer Website* untuk penentuan *website* yang lebih baik.
2. Manfaat kebijakan dari hasil penelitian ini adalah dapat digunakan sebagai bahan pertimbangan dalam penentuan Situs *phishing*.
3. Manfaat teoritis dari penelitian ini yaitu diharapkan dapat memberikan sumbangan untuk pengembangan teori yang berkaitan dengan penerapan pada *Support Vector Machine, Naïve Bayes dan Decision Tree* untuk meningkatkan akurasi prediksi w*ebsite phishing*.

1.6 Ruang Lingkup Penelitian

Ruang lingkup pembahasan dalam penelitian ini dibatasi pada perbandingan Metode *Decision Tree,* Metode *Naïve Bayes* dan Metode *Support Vector Machine* dengan cara menganalisis atribut menjadi bobot atribut *(attribute weighting)* dalam prediksi *phishing* halaman Website.

1.7 Sistematika Penulisan

Disajikan dalam lima bab dan masing-masing bab terdiri dari beberapa sub bab yaitu sebagai berikut:

Bab I Pendahuluan

Bab ini membahas tentang latar belakang penulisan, identifikasi permasalahan, rumusan masalah, tujuan penelitian, ruang lingkup penelitian dan hipotesis.

Bab II Landasan Teori

Bab ini membahas tentang landasan teori yang melandasi penelitian.

Bab III Metode Penelitian

Bab ini berisi tentang metode penelitian yang membahas tentang perancangan penelitian, tahap *computing approach* dan pengembangan sistem.

Bab IV Hasil dan Pembahasan

Bab ini berisi tentang hasil dan pembahasan yang menguraikan tentang implementasi sistem, pengukuran serta implikasi penelitian.

Bab V Kesimpulan dan Saran

Bab ini membahas kesimpulan dari penelitian dan saran untuk penelitian selanjutnya.

BAB II
LANDASAN TEORI

2.1 Tinjuan Pustaka

Dalam penulisan tesis ini penulis melakukan tinjauan dengan menggunakan buku dan jurnal yang berhubungan dengan tema yang dipilih. Secara lebih detail tinjauan dalam penulisan tesis ini dapat dijelaskan sebagai berikut :

Literatur mengenai pembahasan prediksi *phishing* website telah banyak dilakukan dengan beberapa metode. Berikut metode-metode yang pernah digunakan untuk menyelesaikan masalah prediksi pengelabuan website :

1. Penelitian yang dilakukan oleh V. Ramakanth, Neela Megha Shyam Desai dan T.Shyam Prasad pada tahun 2014 tentang Survey Serangan Dan Pertahanan Mekanisme Dalam Phishing (ISSN (online): 2349-0020 http://ijraonline.com, 2014).
2. Penelitan yang dilakukan oleh He Chunjiang dan Zhang Cuilian Zhao Yan pada tahun 2009. *Support Vector Machine* adalah mesin belajar yang terbaru yang telah banyak digunakan sebagai alat untuk klasifikasi data, regresi fungsi, dan lain-lain. Metode memodifikasi kernel untuk meningkatkan kinerja berdasarkan pada pendekatan diferensial, dengan memodifikasi kernel untuk meningkatkan kinerja klasifikasi *Support Vector Machine* oleh fungsi kernel dan metode yang diusulkan dapat diterapkan ke kernel fungsi yang sudah ada. Hasil penelitian menunjukkan bahwa pendekatan secara efektif dapat meningkatkan klasifikasi akurasi (Chunjiang & Yan, 2009).
3. Penelitian yang dilakukan James Luke berjudul "*Data mining of automatically promotion tweet for products and services using Naïve Bayes algorithm to increase twitter engagement followers at PT. Bobobobo*" pada tahun 2015, menjelaskan bahwa data mining adalah pemilihan atau mencari ilmu dari data dalam jumlah besar. *Naïve Bayes* Klasifikasi (NBC) adalah salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikasi. *Naïve Bayes* melakukan keputusan dari masing-masing

kelas, menghitung probabilitas dengan syarat bahwa kelas dianggap sebagai keputusan yang tepat, dengan koleksi objek informasi. Algoritma ini diasumsikan bahwa atribut dari sebuah objek yang independen atau bebas

4. Penelitian Dharm Singh *Analysis of Data Mining Classification with Decision Tree Technique,* pada tahun 2013, melakukan penelitian dengan *Decision Tree* yang memainkan peran penting dalam proses data mining dan analisis data. pembelajaran *Decision Tree* menggunakan satu set data *training* untuk menghasilkan *Decision Tree* yang benar mengklasifikasikan data *training* itu sendiri. Jika proses *machine learning*, *Decision Tree* akan mengklasifikasikan input data baru dengan benar. pohon keputusan berbeda sepanjang beberapa dimensi seperti kriteria membelah, menghentikan aturan, kondisi cabang (*univariat, multivariat*), gaya operasi cabang, jenis pohon akhir
5. Penelitian Isredza Rahmi A Hamid yang berjudul "*Using Feature Selection and Classification Scheme for Automating Phishing Email Detection*" pata tahun 2013, menyajikan sebuah pendekatan untuk mendeteksi email phishing menggunakan fitur *hybrid* yang menggabungkan *content-based* and *behavior-based*. Tujuan utama dari makalah ini adalah untuk mengidentifikasi fitur perilaku berbasis di *email phishing* yang tidak dapat disamarkan oleh penyerang. Dengan menganalisis pola penyerang, teramati bahwa email *phishing* yang memiliki kecenderungan untuk berasal dari lebih dari satu domain bisa menunjukkan aktivitas abnormal

Tabel 2.1 *State of the art research* prediksi Website Phishing menggunakan *Support Vector Machine, Naïve Bayes dan Decision Tree*

No	Penelitian	Tahun	Masalah	Metode	Hasil
1	V. Ramakanth , Neela Megha Shyam Desai dan T.Shyam Prasad	2014	Survei Serangan Dan Pertahanan Mekanisme Dalam *Phishing*	Pertahanan Mekanisme Dalam *Phishing*	Dapat mengetahui melalui survei tentang mempertahankan data dari serangan *Phiser*

2	He Chunjiang dan Zhang Cuilian Zhao Yan	2009	fungsi kernel dan metode yang diusulkan dapat diterapkan ke kernel fungsi yang sudah ada	*Support Vector Machine*	Hasil penelitian menunjukkan bahwa pendekatan secara efektif dapat meningkatkan klasifikasi akurasi
3	James Luke	2015	"*Data mining of automatically promotion tweet for products and services using Naïve Bayes algorithm to increase twitter engagement followers at PT. Bobobobo*"	*Naïve Bayes*	Algoritma ini diasumsikan bahwa atribut dari sebuah objek yang independen atau bebas
4	Dharm Singh	2013	*Analysis of Data Mining Classification with Decision Tree Technique*	*Decision Tree*	Jika proses *machine learning*, *Decision Tree* akan mengklasifikasikan input data baru dengan benar
5	Isredza Rahmi A Hamid	2013	"*Using Feature Selection and Classification Scheme for Automating Phishing Email Detection*"	*Feature Selection and Classification*	menganalisis pola penyerang, teramati bahwa email *phishing* yang memiliki kecenderungan untuk berasal dari lebih dari satu domain bisa

Berdasarkan tinjauan pustaka di atas dapat disimpulkan bahwa ada beberapa peneliti yang sudah menggunakan metode *Support Vector Machine* namun tidak ada penelitian sebelumnya yang menggunakan optimasi dalam hal pemilihan fitur yang sesuai. Dalam penelitian ini akan menggunakan algoritma *Support Vector Machine, Naïve Bayes dan Decision Tree,* yang dibandingkan untuk menentukan K lasifikasi *Phishing* yang terbaik pada bobot atribut yang sesuai dan optimal sehingga hasil prediksi lebih akurat

2.2 Landasan Teori

2.2.1 *Phishing*

Situs *Phishing* adalah pengaruh yang besar pada perdagangan keuangan dan online, mendeteksi dan mencegah serangan ini merupakan langkah penting untuk melindungi terhadap situs phishing serangan, ada beberapa pendekatan untuk mendeteksi serangan ini. Pada bagian ini, kita meninjau ada solusi *phishing* anti dan daftar karya-karya terkait.

Salah satu pendekatan adalah cerdas *Phishing Situs Detection System* menggunakan Teknik *Fuzzy*. Hal ini didasarkan pada logika fuzzy dan menghasilkan enam kriteria "s" dari situs phishing serangan. Ada banyak karakteristik dan faktor-faktor yang dapat membedakan situs yang sah asli dari situs *phishing* yang ditempa dipalsukan seperti kesalahan ejaan, alamat URL yang panjang dan catatan DNS abnormal. Situs phishing tingkat deteksi dilakukan berdasarkan enam kriteria dan ada nomor yang berbeda dari komponen untuk setiap kriteria, kriteria tersebut adalah:

1. URL dan Domain Identitas IP
    a. Menggunakan alamat
    b. URL permintaan abnormal.
    c. Abnormal URL jangkar.
    d. Catatan DNS abnormal.
    e. Abnormal URL.
2. Keamanan dan Enkripsi
    a. Menggunakan sertifikat SSL.
    b. Otoritas sertifikasi.

c. Cookie abnormal.
d. *Distinguished* Nama Sertifikat (DNC).

3. Kode Sumber dan Java Script.
   a. *Redirect* halaman.
   b. Tak terpengaruh serangan.
   c. Serangan *Pharming*.
   d. Menggunakan *on Mouse Over* untuk menyembunyikan link tersebut.
   e. Formulir *Server Handler* (SFH).
4. Gaya Halaman dan Isi.
   a. Kesalahan *Spelling.*
   b. Situs Menyalin.
   c. Menggunakan bentuk dengan tombol "Kirim".
   d. Menggunakan jendela *Popup*.
   e. Menonaktifkan klik kanan.
5. *Web Address Bar*.
   a. Alamat URL panjang.
   b. Mengganti karakter yang sama untuk URL.
   c. Menambahkan awalan atau akhiran.
   d. Menggunakan simbol @ membingungkan.
   e. Menggunakan kode karakter heksadesimal.
6. Sosial Faktor Manusia.
   a. Banyak penekanan pada keamanan dan respon.
   b. Salam generik Umum.
   c. Membeli Waktu untuk *Access Account*.

Aturan dasar memiliki parameter input (kriteria) dan satu output yang berisi semua "*IF-THEN*" aturan sistem. Output untuk masing-masing kriteria adalah salah satu dari berikut: *Genuine*, Diragukan atau Penipuan. *Output* dari situs *phishing* akhir adalah salah satu hasil akhir (Sangat sah, sah, Mencurigakan, *phishy* atau Sangat *phishy*) yang mewakili tingkat situs phishing akhir.

Pendekatan kedua adalah pertahanan sisi klien terhadap pencurian berbasis web identitas. Ini mengusulkan kerangka kerja untuk pertahanan sisi klien: *browser plug-in* yang disebut *SpoofGuard* yang meneliti halaman web dan memperingatkan pengguna ketika permintaan data dapat menjadi bagian dari serangan spoof, itu menghitung indeks *spoof* (ukuran kemungkinan tertentu Halaman adalah bagian dari serangan *spoof*), dan memperingatkan pengguna jika indeks melebihi tingkat yang dipilih oleh pengguna. *SpoofGuard* menggunakan kombinasi evaluasi dan pemeriksaan data pasca keluar halaman untuk menghitung indeks *spoof*.

Ketika pengguna memasukkan *username* dan *password* pada situs *spoof* yang berisi beberapa kombinasi dari URL yang mencurigakan, menyesatkan nama domain, gambar dari sebuah situs yang jujur, dan *username* dan *password* yang sebelumnya telah digunakan pada situs yang jujur, *Spoof Guard* akan mencegat pos dan memperingatkan pengguna dengan popup yang menggagalkan serangan. Makalah ini menjelaskan sifat umum dari sepuluh situs *spoof* baru ini menemukan, mereka Logos, URL Mencurigakan, masukan Pengguna, pendek tinggal, Salinan, kecerobohan atau kurangnya keakraban dengan bahasa Inggris dan HTTPS. *Browser plug-in* berlaku tes untuk semua halaman *download* dan menggabungkan hasil menggunakan mekanisme penilaian. Total Indeks *spoof* halaman menentukan apakah peringatan plug-in pengguna dan menentukan keparahan dan jenis peringatan. Sejak *popup* peringatan yang mengganggu dan menjengkelkan, ia mencoba untuk memperingatkan pengguna melalui indikator toolbar pasif dalam kebanyakan situasi.

Dalam rangka untuk menerapkan gambar dan URL cek, *SpoofGuard plug-in* disertakan dengan *database* tetap gambar dan domain terkait. Ketika *browser download* halaman login semua gambar pada halaman dibandingkan dengan gambar dalam *database SpoofGuard*. *Spoof-nilai* untuk halaman meningkat jika kecocokan ditemukan tetapi halaman "s domain bukanlah domain yang valid untuk gambar. *File history browser* dan sejarah tambahan disimpan oleh *SpoofGuard* digunakan untuk mengevaluasi halaman pengarah. Ketika pengguna mengisi data formulir, penyadapan *SpoofGuard* dan

memeriksa data pasca HTML, memungkinkan pos yang sebenarnya untuk melanjutkan hanya jika indeks *spoof* bawah pengguna ambang batas tertentu untuk posting.

Pendekatan ketiga adalah Anomali Berbasis Web *Phishing* Halaman Deteksi. Itu menguji anomali di halaman web, khususnya, perbedaan antara website "s identitas dan fitur struktural dan transaksi HTTP. Sebuah halaman web terstruktur terdiri dari benda-benda W3C DOM. Di antara mereka, itu daftar lima kategori, berdasarkan relevansinya dengan identitas web. Mereka Kata Kunci / Keterangan (KD), URL *Request* (RURL), URL dari *Anchor* (AURL), Formulir *Server Handler* (SFH) dan Badan Utama (MB) kategori .Ini adalah sumber utama yang identitas dan fitur yang berasal dari dan daftar karakteristik *phishing* seperti *Abnormal* URL, *Abnormal record* DNS, Jangkar *Abnormal, Abnormal Form Server Handler*, URL Permintaan *Abnormal*, *Abnormal* kue dan *sertifikat abnormal* pada SSL.

Hal ekstrak objek web terkait dari halaman web dan mengubahnya menjadi sebuah fitur *vektor* berdasarkan karakteristik analisis *phishing*. Halaman *classifier* mengambil vektor fitur sebagai masukan dan menentukan apakah halaman tersebut palsu atau tidak. Detektor phishing yang diusulkan terdiri dari dua komponen Identitas Extractor dan Halaman *Classifier"s.; Identity Extractor* unik mengidentifikasi kepemilikan website identitas merupakan singkatan dari organisasi "s nama lengkap dan / atau string unik muncul dalam nyanama domain.

Halaman *classifier* mengacu pada benda-benda / properti sebagai fitur struktural. Salah satu sumber fitur struktural yang terkait identitas objek W3C DOM dalam halaman web, misalnya URI domain dari sebuah jangkar. Sumber lain dari fitur struktural adalah transaksi HTTP. Halaman *classifier* mempekerjakan *Support Vector Machine*, algoritma terkenal untuk klasifikasi. Ini output label 1 yang menunjukkan halaman *phishing* atau label -1 yang menunjukkan otentik satu.

Untuk memudahkan klasifikasi *Support Vector Machine* berdasarkan, mereka mengukur fitur tersebut ke dalam vektor. *Output* dari pelaksanaan extractor identitas halaman web adalah karakter string dari kata identitas

diekstrak. Inisialisasi vektor fitur dari halaman web yang alamat URL, data DNS, URL jangkar, permintaan URL, bentuk *server handler*, domain dalam *cookie* dan sertifikat SSL di. Diberi identitas dan satu set fitur, tugas menentukan keaslian halaman web dijalankan oleh *Support Vector Machine*, yang merupakan *classifier* terkenal dan telah banyak digunakan dalam pengenalan pola.

Dari pendekatan kita menyimpulkan bahwa, ada banyak karakteristik dan anomali dapat ditemukan di halaman web dan kita dapat mendeteksi serangan yang mungkin didasarkan pada karakteristik ini, *phisher* menggunakan karakteristik ini di *phishing* halaman web untuk mendapatkan informasi sensitif dari pengguna.

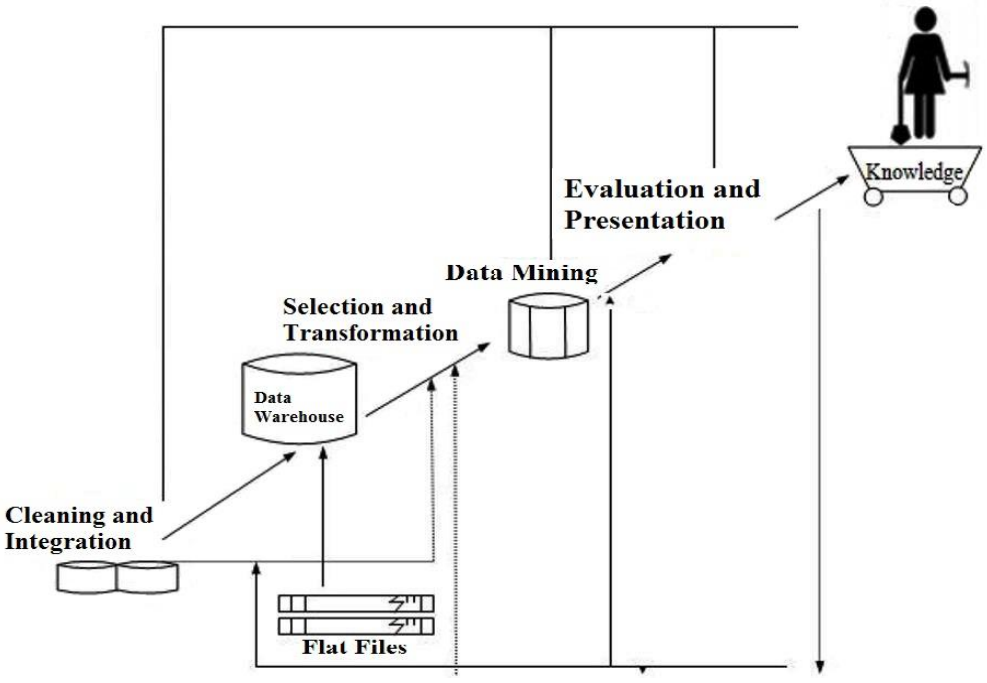
Dalam pendekatan kami didasarkan pada memeriksa karakteristik *phishing* di file kode sumber halaman web, kita mengekstrak karakteristik ini dari standar W3C untuk mengevaluasi keamanan situs dan membuat persentase keamanan berdasarkan berat akhir untuk memutuskan apakah halaman web aman atau tidak.

2.2.2 *Data Mining*

Data Mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Beberapa aplikasi data mining fokus pada prediksi, mereka meramalkan apa yang akan terjadi dalam situasi baru dari data yang menggambarkan apa yang terjadi di masa lalu (Witten, Frank, & Hall, 2011). Kakas data mining meramalkan tren dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting. Analisis yang diotomatisasi yang dilakukan oleh data mining melebihi yang dilakukan oleh sistem pendukung keputusan tradisional yang sudah banyak digunakan (Veronika S Moertini, 2002). Secara khusus koleksi metode yang dikenal sebagai 'data mining' menawarkan metodologi dan solusi teknis untuk mengatasi analisis data medis dan konstruksi prediksi model (Bellazzi & Zupanb, 2008).

Secara umum, tugas data mining dapat diklasifikasikan menjadi dua kategori: deskriptif dan prediktif. Tugas pertambangan deskriptif mengkarakterisasi sifat umum data dalam *database* pertambangan prediktif tugas data pada saat ini untuk membuat prediksi (Han & Kamber, 2007).

Penerapan metode *machine learning* untuk *database* besar disebut data mining, dalam data mining volume besar data diproses untuk membangun sebuah model sederhana, misalnya memiliki akurasi prediksi yang tinggi. Area aplikasi data mining berlimpah, selain retail, bank menganalisi data masa lalu untuk membangun model aplikasi kredit, deteksi penipuan dan pasar saham. Dalam manufaktur, model pembelajaran digunakan untuk optimasi, kontrol, dan pemecahan masalah. Dalam pengobatan, program pembelajaran adalah digunakan untuk diagnosis medis. Dalam telekomunikasi, pola panggilan dianalisis untuk optimasi jaringan dan memaksimalkan kualitas layanan. Dalam ilmu pengetahuan, sejumlah besar data dalam fisika, astronomi, dan biologi dapat hanya dianalisis cukup cepat oleh komputer (Alpaydın, 2010). Model data mining memberikan contoh penerapannya pada berbagai algoritma dan pada data set yang besar (Larose, 2007).



Gambar 2.1 Tahapan data mining (Han & Kamber, 2007)

Tahapan data mining dalam proses penemuan pengetahuan (Han & Kamber, 2007):

1. Pembersihan data (untuk menghilangkan noise dan data tidak konsisten)
2. Integrasi data (dimana beberapa sumber data dapat dikombinasikan)
3. Data seleksi (dimana data yang relevan dengan tugas analisis basis data yang akan diambil)
4. Data transformasi (dimana data diubah atau dikonsolidasikan ke dalam bentuk yang sesuai untuk pertambangan dengan melakukan operasi ringkasan atau agregasi)
5. *Data Mining* (proses esensial dimana metode cerdas diaplikasikan untuk mengekstrak pola data)
6. Pola evaluasi (untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan didasarkan pada beberapa langkah-langkah interestingness)
7. Pengetahuan presentasi (dimana visualisasi dan teknik representasi pengetahuan digunakan untuk menyajikan pengetahuan ditambang kepada pengguna)

Terdapat empat pelompokan dalam data mining yaitu klasifikasi, asosiasi, *clustering* dan prediksi (Witten, Frank, & Hall, 2011):

1. Klasifikasi

   Proses klasifikasi didasarkan Kelas variabel dependen dari model yang merupakan variabel kategori mewakili yang 'label' memakai objek setelah klasifikasi, contohnya loyalitas pelanggan, kelas bintang (galaksi), kelas gempa bumi (badai) (Gorunescu, 2011).

   a. Prediksi variabel bebas dari model diwakili oleh karakteristik (atribut) dari data yang harus diklasifikasikan dan berdasarkan klasifikasi yang dibuat. Contoh prediktor tersebut adalah: merokok, konsumsi alkohol, tekanan darah, frekuensi pembelian, status perkawinan, karakteristik (satelit) gambar, dan kecepatan arah angin (Gorunescu, 2011).
   b. Pelatihan dataset yang merupakan sekumpulan data yang berisi nilai untuk dua sebelumnya komponen, dan digunakan untuk

'pelatihan' model untuk mengenali sesuai kelas, berdasarkan prediksi tersedia. Contoh set tersebut adalah: kelompok pasien diuji pada serangan jantung, kelompok pelanggan dari supermarket (diselidiki oleh internal polling), database yang berisi gambar untuk pemantauan dan pelacakan objek teleskopik astronomi (Gorunescu, 2011).

c. Pengujian dataset yang berisi data baru yang akan diklasifikasikan oleh (*classifier*) model dibangun di atas, dan akurasi klasifikasi (kinerja model) sehingga dapat dievaluasi (Gorunescu, 2011). Model klasifikasi yang populer adalah *Decision/Classification Trees, Bayesian* C*lassifiers / Naïve Bayes classifiers, Neural networks,* Algoritma Genetika, *Support Vector Machines*

2. Asosiasi

Setiap asosiasi antara fitur-fitur yang dicari, bukan hanya satu yang memprediksi nilai kelas tertentu (Witten, Frank, & Hall, 2011). Pada prinsipnya, penemuan aturan asosiasi/asosiasi mempelajari aturan bagaiman kita memahami proses mengidentifikasi aturan antara ketergantungan yang berbeda dari fenomena kelompok. Dengan demikian, mari kita perkirakan kumpulan set yang kita punya masing-masing berisi sejumlah objek/benda-benda. Jadi tujuan kita untuk mencari peraturan yang menghubungkan (asosiasi), obyek ini berdasarkan peraturan ini, untuk dapat memprediksi terjadinya objek/item, berdasarkan kejadian lain (Gorunescu, 2011).

3. Clustering

Cluster adalah menemukan kelompok (kelompok) objek, berdasarkan kemiripan (semacam kemiripan), sehingga dalam setiap kelompok ada kemiripan yang besar, sementara kelompok cukup berbeda dari satu sama lain (Gorunescu, 2011).

4. Prediksi

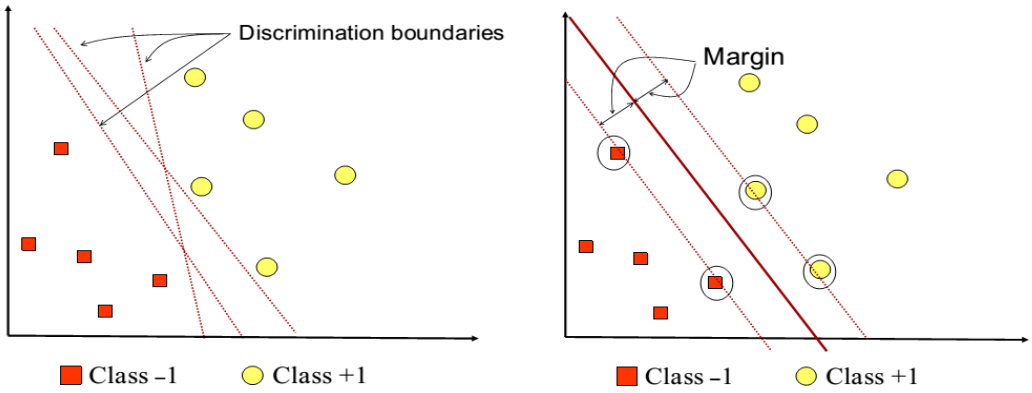
Prediksi/perkiraan model yang berkaitan dengan kemampuan untuk memprediksi tanggapan terbaik (output), yang paling dekat kenyataan, berdasarkan input data. Dengan demikian, semakin kecil perbedaan antara apa yang diharapkan terjadi (hasil yang diharapkan) dan apa yang sebenarnya terjadi (diamati hasil), semakin baik prediksi, contohnya prediksi ramalan cuaca (misalnya, untuk 24 atau 48 jam) atau diagnosis untuk penyakit tertentu yang diberikan kepada pasien tertentu, yang didasarkan pada data medis (Gorunescu, 2011).

Konsep data mining, menemukan pola berharga dalam data, adalah respon yang jelas untuk pengumpulan dan penyimpanan volume data yang besar (Sholom M. Weiss, Indurkhya, & Zhang, 2010). Secara khusus, koleksi metode yang dikenal sebagai 'data mining' menawarkan metodologi dan solusi teknis untuk mengatasi analisis data medis dan konstruksi dari prediksi (Bellazzi & Zupanb, 2008). Untuk semua aplikasi data mining, akurasi prediksi tergantung pada kualitas prediksi atribut (Weiss, Indurkhya, & Zhang, 2010).

2.2.3. Support Vector Machine

Support Vector Machine adalah sebuah metode seleksi yang membandingkan parameter standar seperangkat nilai diskrit yang disebut kandidat set, dan mengambil salah satu yang memiliki akurasi klasifikasi terbaik (Dong, Xia, Tu, & Xing, 2007). *Support Vector Machine* adalah salah satu alat yang paling berpengaruh dan kuat untuk memecahkan klasifikasi (Burges, 1998). *Support Vector Machine* adalah seperangkat metode yang terkait untuk suatu metode pembelajaran, untuk kedua masalah klasifikasi dan regresi (Maimon, 2010). Dengan berorientasi pada tugas, kuat, sifat komputasi mudah dikerjakan, *Support Vector Machine* telah mencapai sukses besar dan dianggap sebagai *state-of-the-art classifier* saat ini (Huang, Yang, King, & Lyu, 2008).

Dua kelas data yang digambarkan sebagai lingkaran dan padat titik-titik yang disajikan di angka ini. Secara intuitif diamati, ada banyak keputusan hyperplanes yang dapat digunakan untuk memisahkan kedua kelompok data. Namun, yang digambarkan dengan angka ini dipilih sebagai yang menguntungkan memisahkan bidang, karena mengandung maksimal margin antara dua kelas. Karena itu, dalam tujuan fungsi *Support Vector Machine*, sebuah istilah *regularization* mewakili margin muncul. Apalagi seperti yang terlihat di angka ini, hanya mereka yang penuh poin disebut mendukung vektor terutama menentukan memisahkan bidang, sementara poin lain tidak memberi kontribusi untuk margin di semua. Dalam kata lain, hanya sejumlah titik penting untuk klasifikasi tujuan dalam kerangka *Support Vector Machine* dan dengan demikian harus diambil (Huang, Yang, King, & Lyu, 2008). Konsep *Support Vector Machine* dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input space. Untuk n-dimensional *space*, input data x_i (i=1. . .k), dimana milik kelas 1 atau kelas 2 dan label yang terkait menjadi -1 untuk kelas 1 dan +1 untuk kelas 2. Gambar 1a memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah class: positif (dinotasikan dengan +1) dan negatif (dinotasikan dengan –1). Pattern yang tergabung pada class negatif disimbolkan dengan kotak, sedangkan pattern pada class positif, disimbolkan dengan lingkaran. Jika data input dapat dipisahkan secara linear, pemisahan *hyper plane* dapat diberikan dalam proses pembelajaran dalam problem klasifikasi diterjemahkan sebagai upaya menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut. Berbagai alternative garis pemisah (*discrimination boundaries*) ditunjukkan pada Gambar 2.2 (Nugroho, 2008).



Gambar 2. 2 *Support Vector Machine* Berusaha Menemukan *Hyperplane* Terbaik Yang Memisahkan Kedua *Class Negatif* dan Positif 2 (Nugroho, 2008)

Hyperplane pemisah terbaik antara kedua class dapat ditemukan dengan mengukur margin hyperplane tersebut. dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut dengan data terdekat dari masing-masing class. Subset data training set yang paling dekat ini disebut sebagai *support vector*. Garis solid pada Gambar 2.2 menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua *class*, sedangkan titik kotak dan lingkaran yang berada dalam lingkaran hitam adalah *support vector*. Upaya mencari lokasi hyperplane optimal ini merupakan inti dari proses pembelajaran pada *Support Vector Machine* (Nugroho, 2008)

2.2.3.1 Contoh Penerapan *Support Vector Machine*

Untuk ilustrasi bagaimana *Support Vector Machine* bekerja, mari kita ikuti dua contoh berikut. Satu adalah contoh dimana data yang ada bisa dipisahkan secara linier. Untuk contoh ini kita gunakan *problem AND*. Contoh yang kedua adalah contoh untuk problem yang tidak bisa dipisahkan secara linier. Untuk contoh ini kita gunakan *problem Exclusive OR* (XOR). *Problem AND* adalah klasifikasi dua kelas dengan empat data (lihat Tabel 2.2). Karena ini problem linier, kernelisasi tidak diperlukan.

Tabel 2.2 AND Problem (Gorunescu, 2011)

X1	X2	y
1	1	1
-1	1	-1
1	-1	-1
-1	-1	-1

dapatkan formulasi masalah optimisasi sebagai berikut:

$$min\frac{1}{2}\left(w\frac{2}{1}+\ w\frac{2}{2}\right)+\ C(t1=t2+t3+t4)$$

Subject to

w1 + w2 + b + t1 ≥ 1 w1 − w2 − b + t2 ≥ 1 −w1 + w2 − b + t3 ≥ 1 w1 + w2 − b + t4 ≥ 1 t1, t2, t3, t4 ≥ 0

Karena fungsi AND adalah kasus klasifikasi linier, maka bisa dipastikan nilai variable slack ti=0. Jadi Kita bisa masukkan nilai C=0. Setelah menyelesaikan problem optimasi di atas didapat solusi

w1 = 1,w2 = 1, b = −1

Persamaan fungsi pemisahnya adalah

f(x) = x1 + x2 − 1.

Untuk menentukan *output* atau label dari setiap titik data/obyek kita gunakan fungsi g(x) = *sign*(x). Dengan fungsi *sign* ini semua nilai f(x)<0 diberi label −1 dan lainnya diberi label +1.

2.2.4. Naïve Bayes

Klasifikasi Bayesian adalah klasifikasi statistik yang bisa memprediksi probabilitas sebuah *class*. Klasifikasi Bayesian ini dihitung berdasarkan Teorema Bayes berikut ini :

$$\mathrm{P}(H|X) \quad = \frac{P(X|H)P(H)}{P(X)} \qquad (2.1)$$

Berdasarkan rumus di atas kejadian *H* merepresentasikan sebuah kelas dan *X* merepresentasikan sebuah atribut. *P(H)* disebut *prior probability H*, contoh dalam kasus ini adalah probabilitas kelas yang mendeklarasikan normal. *P(X*) merupakan *prior probability X*, contoh untuk probabilitas sebuah atribut *protocol_type. P(H|X)* adalah *posterior probability* yang merefleksikan probabilitas munculnya kelas normal terhadap data atribut *protocol_type*. *P(X|H)* menunjukkan kemungkinan munculnya prediktor *X* (*protocol_type*) pada kelas normal. Dan begitu juga seterusnya untuk proses menghitung probabilitas ke-empat kelas lainnya.

2.2.5. Decision Tree

Decision tree adalah algoritma yang paling banyak digunakan untuk masalah pengklasifikasian. Sebuah *Decision Tree* terdiri dari beberapa simpul yaitu *tree's roo, internal nod* dan *leafs.* Konsep entropi digunakan untuk penentuan pada atribut mana sebuah pohon akan terbagi (*split*). Semakin tinggi *entropy* sebuah sampel, semakin tidak murni sampel tersebut. Rumus yang digunakan untuk menghitung *entropy* sampel S adalah sebagai berikut :

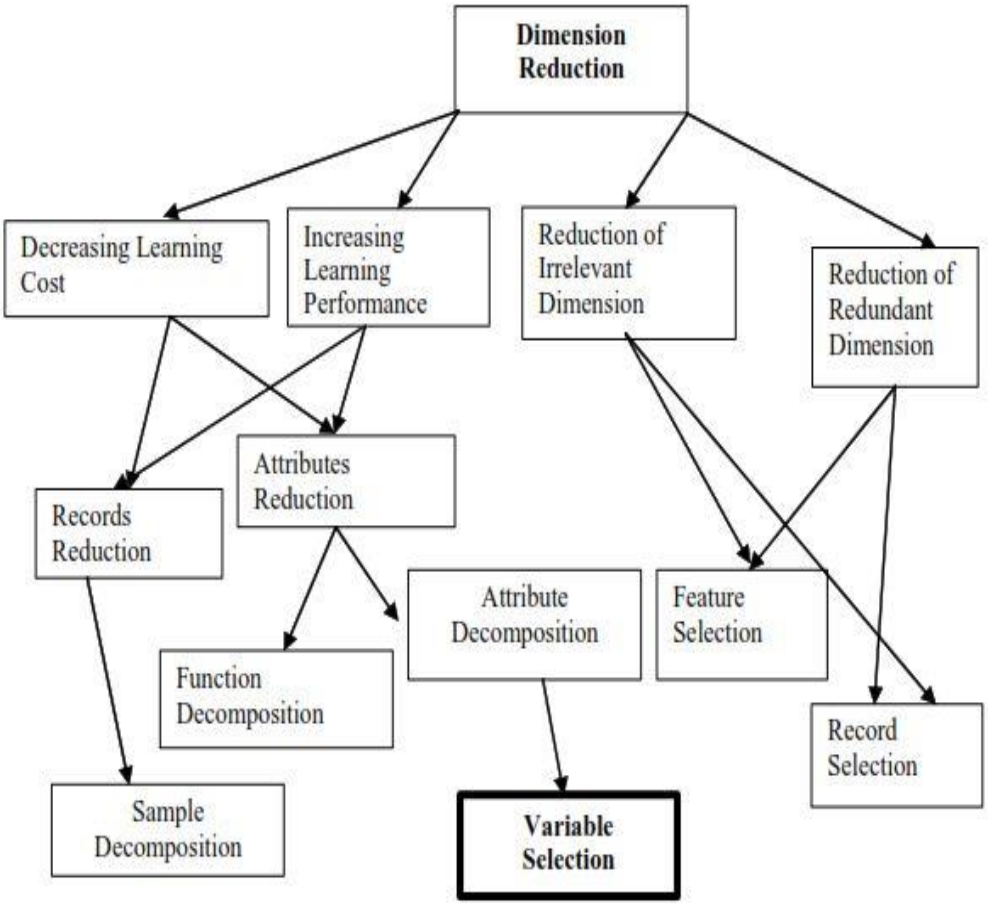
$$Entropy(S) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \quad (2.2)$$

Dimana *p*1, *p*2, ..., *p*n masing-masing menyatakan proposi kelas 1, kelas 2, ..., kelas *n* dalam output.

2.2.6. ***Feature Selection***

Seleksi fitur adalah terkait erat dengan masalah pengurangan dimensi dimana tujuannya adalah untuk mengidentifikasi fitur dalam kumpulan data-sama pentingnya, dan membuang fitur lain seperti informasi yang tidak relevan dan berlebihan dan akurasi dari seleksinya pada masa depan dapt ditingkatkan (Maimon, 2010). Seleksi fitur adalah salah satu faktor yang paling penting yang dapat mempengaruhi tingkat akurasi klasifikasi karena jika dataset berisi sejumlah fitur, dimensi ruang akan menjadi besar dan non-bersih,

merendahkan tingkat akurasi klasifikasi (Liu, Wang, Chen, Dong, Zhu, & Wang, 2011). Masalah dalam seleksi adalah pengurangan dimensi, dimana awalanya semua atribut diperlukan untuk memperoleh akurasi yang maksimal. Empat alasan utama untuk melakukan pengurangan dimensi (Maimon, 2010):



Gambar 2.3 Taksonomi masalah pengurangan dimensi (Maimon, 2010)

1. *Decreasing the learning* (model) cost (penurunan pembelajaran (model) biaya
2. *Increasing the learning* (model) performance (meningkatkan pembelajaran (model) kinerja)
3. *Reducing irrelevant* dimensions (mengurangi dimensi relevan)
4. *Reducing redundant* dimensions (mengurangi dimensi berlebihan)

Tujuan seleksi fitur adalah untuk pengurangan fitur, untuk menghilangkan dari dataset subset dari variabel yang tidak dianggap relevan untuk tujuan dari kegiatan data mining dan fitur metode seleksi dapat diklasifikasikan ke dalam tiga kategori utama (Vercellis, 2009):

1. Metode *filter*

   Metode *Filter* adalah memilih atribut yang relevan sebelum pindah ke tahap pembelajaran berikutnya, atribut yang dianggap paling penting yang dipilih untuk pembelajar, sedangkan sisanya dikecualikan.

2. Metode *wrapper*

   Metode *wrapper* menilai sekelompok variabel dengan menggunakan klasifikasi yang sama atau algoritma regresi digunakan untuk memprediksi nilai dari variabel target.

3. Metode *embedded*.

   Untuk metode *embedded*, proses seleksi atribut terletak di dalam algoritma pembelajaran, sehingga pemilihan set optimal atribut secara langsung dibuat selama fase generasi model.

2.2.7. RapidMiner

RapidMiner sebelumnya dikenal sebagai YALE *(Yet Another Learning Environment)* dikembangkan mulai tahun 2001 oleh Ralf Klinkenberg, Ingo mierswa, dan Simon Fischer di *Unit Artificial Intelligence* dari *Technical University of Dortmund.* Mulai tahun 2006, perkembangannya adalah didorong oleh Rapid-I, perusahaan yang didirikan oleh Ingo Mierswa dan Ralf Klinkenberg pada tahun yang sama. Pada tahun 2007, nama *software* tersebut berubah dari YALE ke *RapidMiner*. Pada tahun 2013 perusahaan yang bernama Rapid-I berubah nama menjadi *RapidMiner*.

RapidMiner merupakan aplikasi data mining yang tidak perlu dipertanyakan lagi dan berbasis sistem *open-source* dunia yang terkemuka dan ternama. Tersedia sebagai aplikasi yang berdiri sendiri

untuk analisis data dan sebagai mesin data mining untuk integrasi ke dalam produk sendiri. Ribuan aplikasi *RapidMiner* di lebih dari 40 negara memberikan pengguna mereka keunggulan yang kompetitif. Solusi yang di usung antara lain :Integrasi data, Analitis ETL, Data Analisis, dan Pelaporan dalam satu suite tunggal. *Powerfull* tapi memiliki antarmuka pengguna grafis yang intuitif untuk desain analisis proses. Repositori untuk proses, data dan penanganan metadata Hanya solusi dengan transformasi meta data: lupakan trial and error dan memeriksa hasil yang telah di inspeksi selama desain. Hanya solusi yang mendukung *on-the-fly* kesalahan dan dapat melakukan perbaikan dengan cepat, lengkap dan fleksibel. Ratusan loading data, transformasi data, pemodelan data, dan metode visualisasi data *RapidMiner* menyediakan prosedur *data mining* dan *machine learning* termasuk: ETL (*extraction, transformation, loading*), data *preprocessing*, visualisasi, *modelling* dan evaluasi. Proses data mining tersusun atas operator-operator yang *nestable*, dideskripsikan dengan XML, dan dibuat dengan GUI. Ditulis dalam bahasa pemrograman Java. Mengintegrasikan proyek data mining Weka dan statistika R.

Terminologi dasar atribut dan atribut target atribut: karakteristik atau fitur dari data yang menggambarkan sebuah proses atau situasi. ID, atribut biasa. Atribut target: atribut yang menjadi tujuan untuk diisi oleh proses data mining. *Label, cluster, weight*. Peran atribut *(attribute role) Label, cluster, weight*, ID, biasa Terminologi Dasar Tipe nilai (*value type)* nominal: nilai secara kategori numeric: nilai numerik secara umum integer: bilangan bulat *rea*l: bilangan nyata *text*: teks bebas tanpa struktur binominal: nominal dua nilai polynominal: nominal lebih dari dua nilai *date_time*: tanggal dan waktu date: hanya tanggal time: hanya waktu Terminologi Dasar Data dan metadata Data menyebutkan obyek-obyek dari sebuah konsep. Ditunjukkan sebagai baris dari tabel. Metadata menggambarkan karakteristik dari konsep tersebut. Ditunjukkan sebagai kolom dari tabel. Modelling Penggunaan metode data mining terhadap

data. Hasilnya disebut model. Fungsi menu Process control Untuk mengontrol aliran proses, seperti loop atau conditional branch. *Utility* Untuk mengelompokkan *subprocess*, juga *macro* dan *logger*. *Repository access* untuk membaca dan menulis repository. *Import* untuk membaca dari berbagai format eksternal. *Export* untuk menulis data ke berbagai format eksternal. Data transformation untuk transformasi data dan metadata. Modelling untuk proses data mining yang sesungguhnya. Seperti klasifikasi, regresi, *clustering*, aturan asosiasi dll. *Evaluation* untuk menghitung kualitas dari modeling.

1. Keunggulan dan Kelemahan *RapidMiner*
   a. Keunggulan *RapidMiner*

      Sudah tidak diragukan lagi *RapidMiner* memiliki keunggulan tersendiri *RapidMiner* adalah aplikasi data mining yang tidak perlu dipertanyakan lagi dan berbasis sistem *open-source* dunia yang terkemuka dan ternama. Tersedia sebagai aplikasi yang berdiri sendiri untuk analisis data dan sebagai mesin data mining untuk integrasi ke dalam produk sendiri. Ribuan aplikasi *RapidMiner* di lebih dari 40 negara memberikan pengguna mereka keunggulan yang kompetitif. Solusi yang diberikan antara lain :

      1. Integrasi data
      2. Analitis ETL
      3. Data Analisis

      Pelaporan dalam satu suite tunggal. *Powerfull* tapi memiliki antarmuka pengguna grafis yang intuitif untuk desain analisis proses. Repositori untuk proses, data dan penanganan meta data hanya solusi dengan transformasi meta data: lupakan *trial* and *error* dan memeriksa hasil yang telah di inspeksi selama desain. Hanya solusi yang mendukung *on-the-fly* kesalahan dan dapat melakukan perbaikan dengan cepat Lengkap

dan fleksibel : ratusan *loading* data, transformasi data, pemodelan data, dan metode visualisasi data.

b. Kelemahan *RapidMiner*

Walaupun tidak tercantum dalam aplikasi *requirement hardware* akan lebih baik didukung dengan *high performance memory* maupun *processor* yang digunakan, artinya membutuhkan perangkat keras yang cukup memadai dalam penggunaan *RapidMiner.*

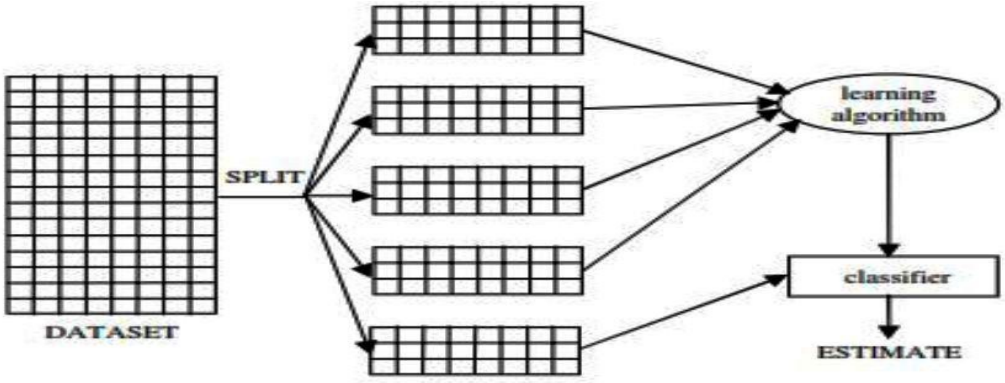
2. Target/Dataset Menggunakan *RapidMiner*

RapidMiner telah menjadi salah satu perangkat lunak analisis *open source* atau bahkan analisis dianggap sebagai kata kunci di fashion. Perangkat lunak RapidMiner telah menjadi perintis di banyak daerah (seperti membangun pasar untuk *RapidMiner* Ekstensi). *RapidMiner* sebagai cara untuk merancang terstruktur, proses berulang dan kemampuan untuk mengoptimalkan parameter pelajar dengan cara yang sistematis. Hal ini dalam data set besar pun berjalan dengan baik pada 32- bit *Windows*, terutama ketika versi 5.0 dirilis, semakin lebih baik dalam analisis *open source.*

2.2.8. Pengujian Evaluasi dan Validasi Metode Klasifikasi Data Mining

2.2.8.1. Pengujian K-*fold Cross-Validation*

Pendekatan alternatif untuk '*training* dan *testing*' yang sering diadopsi ketika sejumlah contoh kecil (dan yang banyak yang memilih menggunakan terlepas dari ukuran) dikenal sebagai *K-fold Cross-validation*. Jika dataset terdiri kasus N, ini dibagi menjadi bagian-bagian k sama, k biasanya menjadi sejumlah kecil seperti 5 atau 10. (Jika N tidak tepat habis dibagi oleh k, bagian akhir akan memiliki kasus lebih sedikit dari k lain - 1 bagian serangkaian berjalan k kini dilakukan : Setiap bagian k pada gilirannya digunakan sebagai ujian menetapkan dan k lainnya -1 bagian digunakan sebagai *training set* (Bramer, 2007) seperti pada gambar 2.4 :



Gambar 2.4 K-*fold Cross-validation* (Bramer, 2007)

2.2.8.2.Confusion matrix

Confusion matrix merupakan dataset hanya memiliki dua kelas, kelas yang satu sebagai positif dan kelas yang lain sebagai negatif (Bramer, 2007). Metode ini menggunakan tabel matrix seperti pada tabel 2.3.

Tabel 2.3 Model *Confusion Matrix* (Vercellis, 2009)

		Predictions		Total
		-1(negative)	+(positf)	
	-1 (negatif)	P	q	p + q
Examples	+1 (positif)	U	v	u + v
	Total	p + u	q + v	m

Keterangan:

a. p adalah jumlah prediksi yang tepat bahwa *instance* bersifat negatif.
b. q adalah jumlah prediksi yang salah bahwa *instance* bersifat positif.
c. u adalah jumlah prediksi yang salah bahwa *instance* bersifat negatif.
d. v adalah jumlah prediksi yang tepat bahwa *instance* bersifat positif.

Berikut adalah persamaan model *confusion matrix*:

a. Nilai akurasi (acc) adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan menggunakan persamaan:

$$\text{acc} = \frac{p+v}{(p+q+u+v)} = \frac{p+v}{m} \quad (2.3)$$

b. Tingkat negatif benar (tn) didefinisikan sebagai proporsi kasus negatif yang diklasifikasikan dengan benar, yang dihitung dengan menggunakan persamaan:

$$\text{tn} = \frac{p}{p+q} \quad (2.4)$$

c. Tingkat negatif palsu (fn) adalah proposi kasus positif yang salah diklasifikasikan sebagai negatif, yang dihitng dengan menggunakan persamaan:

$$\text{fn} = \frac{u}{u+v} \quad (2.5)$$

d. Tingkat negatif palsu (fp) adalah proporsi kasus negatif yang salah diklasifikasikan sebagai positif, yang dihitung dengan menggunakan persamaan:

$$\text{fp} = \frac{q}{p+q} \quad (2.6)$$

e. Penarikan kembali (*recall*) atau tingkat positif benar (tp) adalah proporsi kasus positif yang diklasifikasikan dengan benar, yang dihitung dengan menggunakan persamaan:

$$\text{tp} = \frac{u}{u+v} \quad (2.7)$$

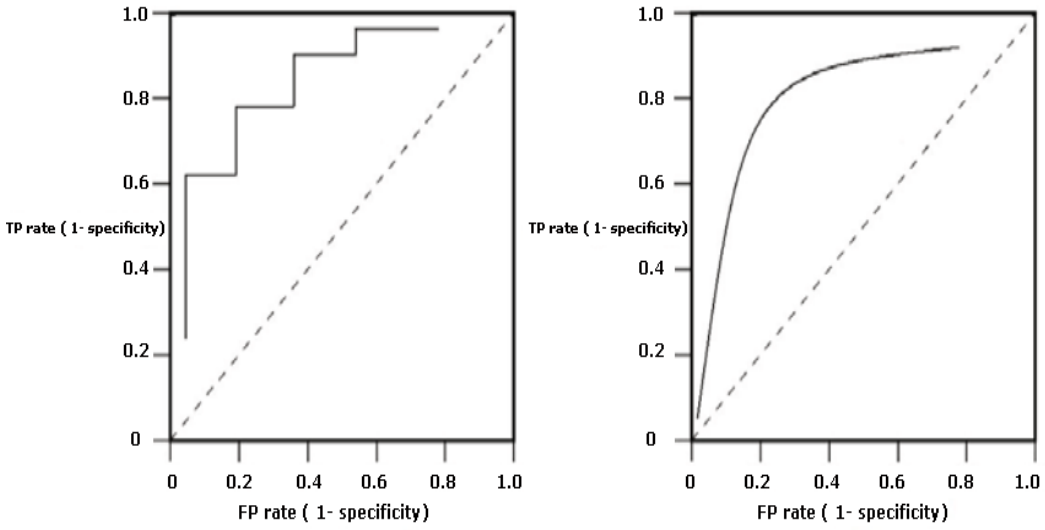
f. Presisi (p) adalah proporsi prediksi kasus positif yang benar, yang dihitung dengan menggunakan persamaan:

$$\text{prc} = \frac{v}{q+v} \quad (2.8)$$

2.2.8.3. Kurva ROC (Receiver Operating Characteristic)

Grafik kurva ROC (*Receiver* operasi karakteristik) digunakan untuk mengevaluasi akurasi classifier dan untuk membandingkan klasifikasi yang berbeda model (Vercellis, 2009). Sebuah grafik ROC adalah grafik dua dimensi dengan

proporsi negatif pada sumbu horisontal dan proporsi positif yang benar di sumbu vertikal (Vercellis, 2009). Kegunaan kurva ROC adalah untuk radar selama Perang Dunia II untuk mendeteksi benda-benda musuh di medan perang, teori deteksi sinyal, dalam psikologi ke rekening untuk deteksi sinyal persepsi, penelitian medis dan dalam mesin pembelajaran dan penelitian data mining, serta masalah klasifikasi. Dalam masalah klasifikasi menggunakan kelas keputusan dua (klasifikasi biner), masing-masing objek dikelompokkan dalam (P, N), yaitu positif atau negatif. Sementara model klasifikasi beberapa (misalnya, pohon keputusan) menghasilkan label kelas diskrit (menunjukkan hanya kelas diprediksi objek), pengklasifikasi lainnya (misalnya, Naive Bayes, jaringan saraf) menghasilkan output yang berkesinambungan, yang ambang batas yang berbeda mungkin diterapkan untuk memprediksi keanggotaan kelas, secara teknis, ROC kurva, juga dikenal sebagai grafik ROC, dua-dimensi grafik di mana tingkat TP diplot pada sumbu Y-dan tingkat FP diplot pada X-sumbu (Gorunescu, 2011).



Gambar 2.5 Grafik ROC (*discrete/continuous case*)(Gorunescu, 2011)

Pada Gambar 2.5 garis diagonal membagi ruang ROC, yaitu:

1. Poin di atas garis diagonal merupakan hasil klasifikasi yang baik.
2. Point di bawah garis diagonal merupakan hasil klasifikasi yang buruk.

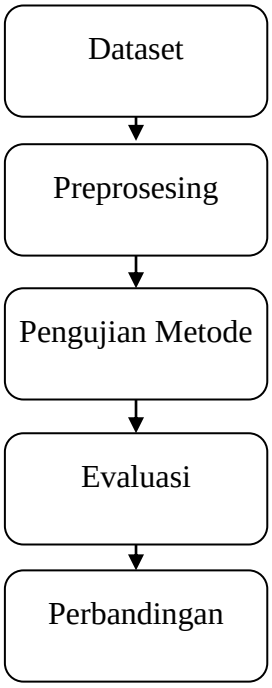
Dapat disimpulkan bahwa, satu point pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik.

Untuk keakuransian nilai AUC dalam klasifikasi *data mining* dibagi menjadi lima kelompok (Gorunescu, 2011), yaitu:

a. 0,90 – 1,00 = klasifikasi sangat baik (*excellent classification*)
b. 0,80 – 0,90 = klasifikasi baik (*good classification*)
c. 0,70 – 0,80 = klasifikasi cukup (*fair classification*)
d. 0,60 – 0,70 = klasifikasi buruk (*poor classification*)
e. 0,50 – 0,60 = klasifikasi salah (*failure*)

2.3. Kerangka Pemikiran

Model kerangka pemikiran yang digunakan adalah *method improvement* (perbaikan metode), yang sering digunakan pada penelitian di bidang sains dan teknik, termasuk bidang computing didalamnya. Komponen dari model kerangka pemikiran perbaikan metode (*metode improvement*) adalah ***Indicators***, ***Proposed Method***, ***Objectives***, dan ***Measurements*** (Polancic, 2010). Kerangka pemikiran pada penelitian ini dimulai dari prediksi hasil pemilihan umum. Maka dengan ini penulis mencoba membuat sebuah *soft computing* dengan mengunakan *Support Vector Machine*, *Naïve Bayes* dan *Decision Trees*, dengan kerangka pemikiran seperti yang ditampilkan pada Gambar 2.6 di bawah ini :



Gambar 2.6 Kerangka Pemikiran

BAB III
METODOLOGI PENELITIAN

3.1 Desain Penelitian

Pengertian penelitian dalam akademik yaitu digunakan untuk mengacu pada aktivitas yang rajin dan penyelidikan sistematis atau investigasi di suatu daerah, dengan tujuan menemukan atau merevisi fakta, teori, aplikasi dan tujuannya adalah untuk menemukan dan menyebarkan pengetahuan baru (Berndtsson, Olsson, & Lundell, 2008).

Menurut (Dawson, 2009) ada empat metode penelitian yang umum digunakan yaitu tindakan penelitian, eksperimen, studi kasus dan survey. Dalam konteks penelitian, metode yang dilakukan mengacu kepada pemecahan masala yang meliputi mengumpulkan data, merumuskan hipotesis atau proposisi, pengujian hipotesis, menafsirkan hasil, dan kesimpulan (Berndtssom, Hansson, Olsson, & Lundell, 2008).

Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian sebagai berikut :

1. Pengumpulan data

   Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses.

2. Pengolahan data awal

   Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model.

3. Metode yang diusulkan

   Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (*training data*) dan data uji (*testing data*) juga diperlukan untuk pembuatan model.

4. Eksperimen dan pengujian metode

   Pada tahap ini model yang diusulkan akan diuji untuk melihat hasil berupa *rule* yang akan dimanfaatkan dalam pengambilan keputusan.

5. Evaluasi dan validasi

   Pada tahap ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

3.2. Pengumpulan Data

Teknik pengumpulan data ialah teknik atau cara-cara yang dapat digunakan untuk menggunakan data. Dalam pengumpulan data terdapat sumber data, sumber data yang terhimpun langsung oleh peneliti disebut dengan sumber primer, sedangkan apabila melalui tangan kedua disebut sumber sekunder. Data yang diperoleh adalah data sekunder karena diperoleh dari Rami Mustafa A Mohammad (*University of Huddersfield*) database dalam UCI. Masalah yang harus dipecahkan di sini adalah Memprediksi Situs *Phishing* dengan menggunakan dataset public dari UCI kami menjelaskan fitur penting yang telah terbukti bagus dan efektif dalam memprediksi situs *phishing*. Selain itu, kami menetapkan beberapa *Attribut* yang digunakan pada penelitian ini seperti pada tabel 3.1.

3.3 Pengolahan data awal

Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 11056 data, namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (*preparation data*)**.** Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan sebagai berikut (vecellis, 2009):

1. Data *validation,* untuk mengidentifikasikan dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*).
2. Data *integration and transformation,* untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan kedalam *software RapidMiner*.

Tabel kategorikal atribut terlihat pada tabel 3.1.

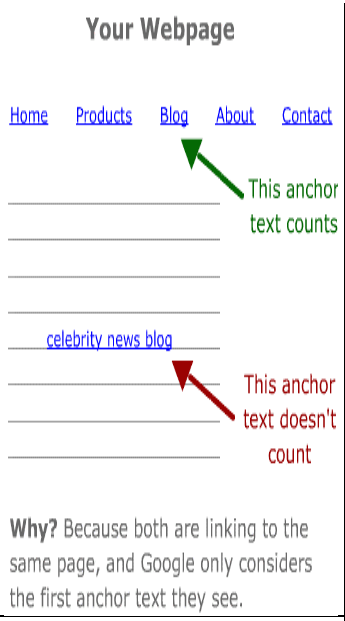
3. *Data size reduction and discritization*, untuk memperoleh data set dengan jumlah atribut dan *record* yang lebih sedikit tetapi bersifat informative.

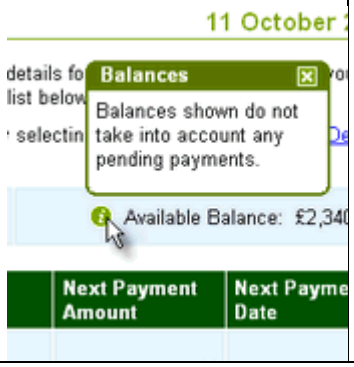
Tabel 3.1 Tabel *Atribut* yang digunakan (UCI *Repository*)

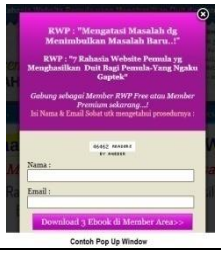
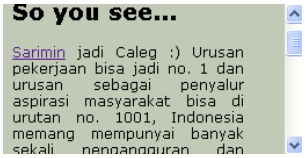
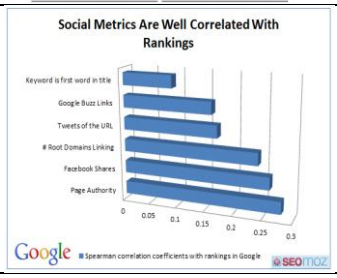
Attribut	Contoh	Keterangan
Having IP Address	192.168.100.1 , 10.57.38.223	Sebuah alamat pada komputer agar komputer bisa saling terhubung dengan komputer lain, IP Address terdiri dari 4 Blok, setiap Blok di isi oleh angka 0 - 255
URL Length	http://support.microsoft.com/kb/208427	Panjang suatu alamat di internet yang mempunyai maksimum panjang 2.000 character
Shortining Service	"http://en.wikipedia.org/wiki/URL_shortening" dapat disingkat menjadi "http://tinyurl.com/urlwiki"	URL shortening (Shortener URL) adalah teknik di World Wide Web di mana Uniform Resource Locator (URL) dapat dibuat secara substansial lebih pendek dan masih mengarahkan ke halaman yang dibutuhkan
Having At Symbol	"www.zuhri@yahoo.co.id maka halaman akan menuju ke https://id.yahoo.com/	Penggunaan simbol "@" di URL mengarah browser untuk mengabaikan segala sesuatu yang mendahului "@" simbol dan alamat sebenarnya sering mengikuti simbol "@".
double slash redirecting	http://www.legitimate.com//http://www.phishing.com	Keberadaan "//" dalam jalur URL berarti bahwa pengguna akan diarahkan ke situs lain. Kami menemukan bahwa jika URL dimulai dengan "HTTP", yang berarti "//" akan muncul di posisi keenam. Namun, jika URL mempekerjakan "HTTPS" maka "//" akan muncul di posisi ketujuh.
Prefix Suffix	Misalnya http://www.Confirme-paypal.com/	Simbol dasbor jarang digunakan dalam URL yang sah. Phisher cenderung menambah awalan atau akhiran dipisahkan oleh (-) untuk nama domain sehingga pengguna merasa bahwa mereka berhadapan dengan halaman web yang sah

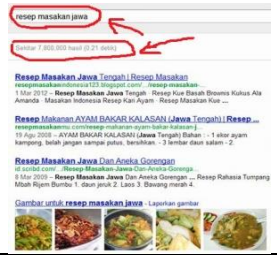
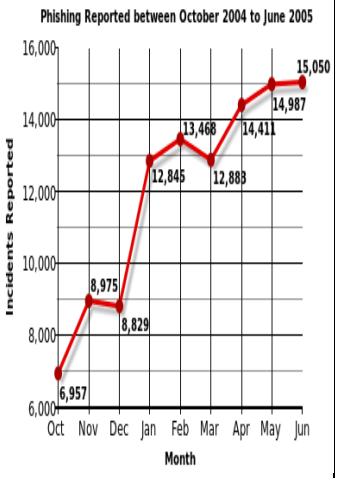
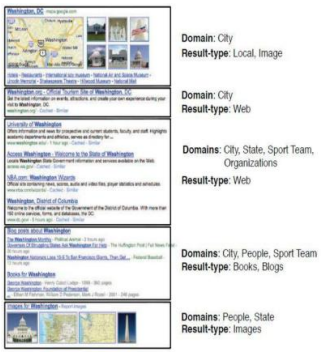
Having Sub Domain	http://www.hud.ac.uk/students/	Mari kita asumsikan kita memiliki link berikut: http://www.hud.ac.uk/students/. Sebuah nama domain mungkin termasuk negara-kode domain tingkat atas (ccTLD), yang pada contoh kita adalah "uk". The "ac" bagian adalah singkatan untuk "akademis", gabungan "ac.uk" disebut domain tingkat kedua (SLD) dan "hud" adalah nama sebenarnya dari domain. Untuk menghasilkan aturan untuk mengekstraksi fitur ini, kita terlebih dahulu harus menghilangkan (www.) Dari URL yang sebenarnya sub domain sendiri. Kemudian, kita harus menghapus (ccTLD) jika ada. Akhirnya, kita menghitung titik-titik yang tersisa. Jika jumlah titik lebih besar dari satu, maka URL tersebut diklasifikasikan sebagai "Mencurigakan" karena memiliki satu sub domain. Namun, jika titik-titik yang lebih besar dari dua, itu diklasifikasikan sebagai "Phishing" karena akan memiliki beberapa sub domain. Jika tidak, jika URL tidak memiliki sub domain, kami akan menetapkan "sah" untuk fitur tersebut.
SSLfinal State		SSL atau Secure Sockets Layer adalah sebuah protokol keamanan data yang digunakan untuk menjaga pengiriman data web server dan pengguna situs web tersebut. Jenis SSL yang paling aman dapat dilihat dari tingkat keamanan SSL, yang terletak pada kekuatan enkripsi yang didukungnya (misalnya 256 bit). Semakin besar tingkat enkripsi semakin susah untuk dibobol. Secara teknis, semua SSL dengan tingkat enkripsi yang sama, mempunyai tingkat keamanan yang sama. Untuk mengetahui apabila transaksi diamankan oleh SSL adalah sebuah icon berlambangkan

		gembok yang terkunci akan muncul di browser yang telah diamankan dengan SSL. Dengan meng-klik icon tersebut akan diketahui otoritas sertifikasi dari sertifikat SSL tersebut.
Domain Registration Length		Domain Registration Length adalah panjang nama domain yang bisa didaftarkan pada penyedia jasa pendaftaran domain
Favicon		sebuah ikon yang diasosiasikan dengan logo sebuah situs
port	20 TCPUDP ftp-data File Tran, 21 TCP,UDP ftp File Tran (control, 23 TCP,UDP telnet telnet, 25 TCP,UDP smtp Simplemail, 53 TCP,UDP domain Domain, 67 TCP,UDP bootpc DHCP/, 68 TCP,UDP bootpc DHCP/, 69 TCP,UDP tftp Trivial Fi, 80 TCP,UDP www World W, 110 TCP,UDP pop3 PostOff, 123 TCP,UDP ntp Network	Mekanisme yang mengizinkan sebuah komputer untuk mendukung beberapa sesi koneksi dengan komputer lainnya dan program di dalam jaringan
HTTPS token	https://www.google.co.id/	bentuk protokol domain valid dan aman
Request URL		
URL of Anchor		URL of Anchor : Area berupa node-node di antara konten yang merupakan source atau tujuan dari sebuah link. Dengan meng-klik mouse pada anchor area, maka pada window akan terbuka link atau source yang dituju. Jadi anchor area ini merupakan semacam highlight. Anchor area ini juga dikenal sebagai span, region, button, atau extent. Berguna agar text dan graphic dapat di-link pada suatu tempat dalam satu dokumen yang sama. Link ini membutuhkan 2 bagian yaitu : Anchor, yang

		bertujuan untuk menandai suatu text/grafik. Link,bertujuan untuk mengantar ke tempat yang telah di tandai tadi.
Links in tags	Alamat subdomain ditulis dengan lengkap, seperti **http://www.mrizqiariadi.wordpress.com**, atau jika kita merujuk/ menambahkan halaman/links kepada halaman tertentu, menjadi **http://www.mrizqiariadi.wordpress.com/nama_halaman.html**.	Links in tags adalah membuat sebuah text jika di-klik akan pindah ke halaman lainnya
SFH	Misalnya, bentuk dapat digunakan untuk memasukkan pengiriman atau data kartu kredit untuk memesan produk, atau dapat digunakan untuk mengambil hasil pencarian dari mesin pencari.	Server Form Handler (SFH) adalah Sebuah formulir web, bentuk web atau HTML formulir di halaman web memungkinkan pengguna untuk memasukkan data yang dikirim ke server untuk diproses. Formulir dapat menyerupai kertas atau database bentuk karena pengguna web mengisi formulir menggunakan checkbox, tombol radio, atau bidang teks.
Submitting to email		Bentuk web memungkinkan pengguna untuk mengirimkan informasi pribadi yang diarahkan ke server untuk diproses. Sebuah phisher mungkin mengarahkan informasi pengguna ke email pribadinya. Untuk itu, bahasa script

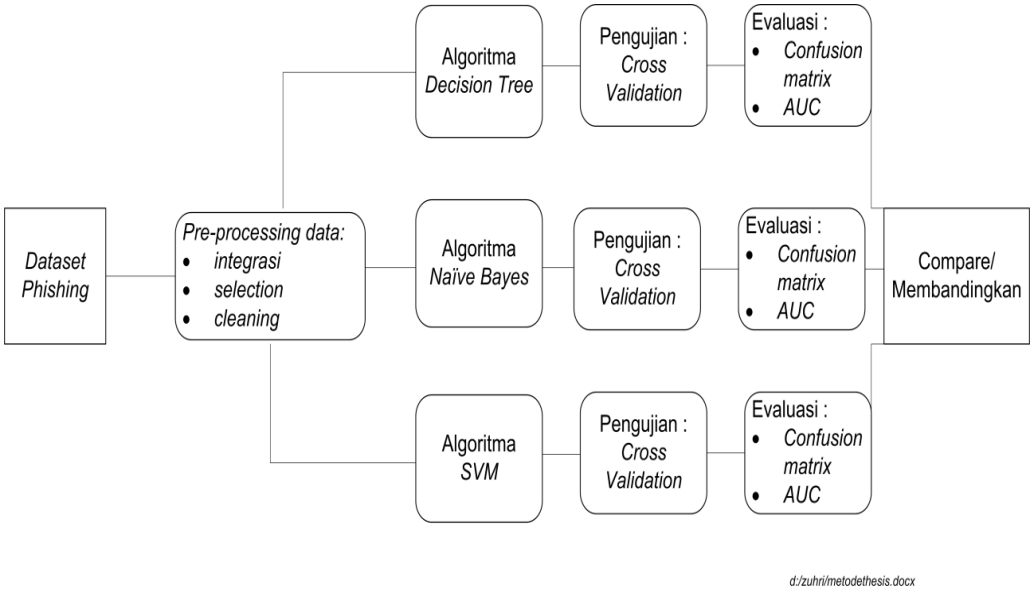
		sisi server dapat digunakan seperti "mail ()" fungsi dalam PHP. Satu lagi fungsi client-side yang dapat digunakan untuk tujuan ini adalah "mailto:" fungsi.
Abnormal URL	http://twitpics.videos.mittelfingerstempel.com/login-scure.html	Abnormal URL singkatan dari "Abnormal Uniform Resource Locator" adalah rangkaian karakter dengan format tertentu yang digunakan untuk merepresentasikan alamat atau sumber dokumen di internet. atau suatu sarana yang digunakan untuk menentukan lokasi informasi pada suatu Web server tidak normal, Abnormal URL dapat diibaratkan sebagai suatu alamat palsu, dimana alamat tersebut terdiri dari Protokol yang digunakan oleh suatu browser untuk mengambil informasi, Nama dari komputer di mana informasi tersebut berada, dan Jalur serta nama file dari suatu informasi.
Redirect	www.ur13zuhry.blogspot.com redirect to www.mr-file.co.nr	*Redirect* adalah mengarahkan kembali (re = kembali, direct = arah) tujuan utama dari sebuah domain atau subdomain.
on mouseover		Event yang akan aktif saat user menaruh kursor di posisi tertentu

Right Click		perintah klik kanan pada mouse
pop Up Window		Jendela yang biasanya muncul tiba-tiba ketika mengunjungi halaman web
Iframe		Salah satu cara untuk menciptakan sebuah 'jendela' dalam halaman web yang mampu menampilkan dokumen terpisah di dalam jendela yang tanpa reload seluruh halaman
age of domain	www.detik.com	sebuah *domain* yg berusia minimal sudah berusia 1 tahun
DNS Record	contoh jika kita ingin memetakan domain bestariwebhost.com ke alamat IP 119.81.75.214 maka kita bisa memberikan catatan pada A Record dengan format : bestariwebhost.com 14400 in A 119.81.75.214 dimana angka 14400 adalah TTL (Time To Live, dalam satuan detik)	Data yang menyajikan pemetaan dan pengalamatan sebuah*nama domain internet*
web traffic		lalu lintas pengunjung (visitor) yang membuka (loading) halaman suatu website
Page Rank		sebuah algoritma yang telah dipatenkan yang berfungsi menentukan situs web mana yang lebih penting/popular

Google Index		kondisi suatu web/blog ter-**index**(masuk dalam) database **google**
Number Links pointing to page		Jumlah link yang menunjuk ke halaman web menunjukkan tingkat legitimasi, bahkan jika beberapa link yang dari domain yang sama (Dean, 2014). Dalam dataset kami menemukan bahwa 98% dari item dataset phishing tidak memiliki link yang menunjuk kepada halaman web tersebut. Di sisi lain, situs yang sah memiliki minimal 2 link eksternal menunjuk kepada halaman web tersebut.
Statistical report		Beberapa pihak seperti PhishTank (PhishTank Statistik, 2010-2012), dan StopBadware (StopBadware, 2010-2012) merumuskan berbagai laporan statistik dari situs phishing pada setiap periode waktu tertentu; beberapa bulanan dan lain kuartalan. Dalam penelitian kami, kami menggunakan 2 bentuk sepuluh statistik dari PhishTank: "Top 10 Domain" dan "Top 10 IP" menurut statistik-laporan yang diterbitkan dalam tiga tahun terakhir, mulai di January2010 untuk November 2012. Sedangkan untuk "StopBadware ", kami menggunakan" Top 50 "alamat IP.
Result		Result adalah hasil lokasi domain

3.4 Metode yang diusulkan

Pada tahap modeling ini dilakukan pemrosesan data training sehingga akan membahas metode algoritma yang diuji dengan memasukan data Website Phishing kemudian dianalisa dan dikomparasi. Berikut ini bentuk gambaran metode algoritma yang akan diuji seperti pada gambar 3.1 di bawah ini.



Gambar 3.1 Metode yang diusulkan

3.5 Eksperimen dan Pengujian Metode

Tahap modeling untuk menyelesaikan prediksi situs phishing dengan menggunakan dua metode yaitu algoritma *Support Vector Machine, Naïve Bayes* dan *Decision Tree*

1. ***Support Vector Machine*** yaitu suatu metode sebuah metode seleksi fitur, dan mengambil salah satu yang memiliki akurasi klasifikasi terbaik.
2. ***Naïve Bayes Classifier*** merupakan sebuah metode klasifikasi yang berakar pada teorema *Bayes*. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik
3. ***Decision Tree***/Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki

Pada penelitian kali ini yang digunakan adalah penelitian *Experiment*.

Penelitian eksperimen melibatkan penyelidikan hubungan kausal menggunakan tes dikendalikan oleh si peneliti itu sendiri. Dalam penelitian eksperimen digunakan spesifikasi software dan hardware sebagai alat bantu dalam penelitian pada Tabel 3.2:

Tabel 3.2 Spesifikasi *hardware* dan *software*

Software	Hardware
Sistem Operasi: Windows 7 or Higher	CPU: Intel Pentium Dual Core or Higher
Data Mining: *RapidMiner* versi 7.0	RAM : 2 GB or Higher
	Hardisk : Minimum 2 GB free disk space

3.6 Evaluasi dan Validasi Hasil

Model yang diusulkan pada penelitian tentang prediksi Situs *Phishing* adalah dengan menerapkan *Support Vector Machine*, *Decision Tree* dan *Naïve Bayes*. Penerapan algoritma *Support Vector Machine, Decision Tree* dan *Naïve Bayes* dengan menentukan nilai *weight* terlebih dahulu. Setelah didapatkan nilai akurasi dan AUC terbesar, nilai *weight* tersebut akan dijadikan nilai yang akan digunakan untuk mencari nilai akurasi dan AUC tertinggi. Setelah ditemukan nilai akurasi yang paling ideal dari parameter tersebut langkah selanjutnya adalah menentukan nilai *weight*. sehingga terbentuk struktur algoritma yang ideal untuk pemecahan masalah tersebut.

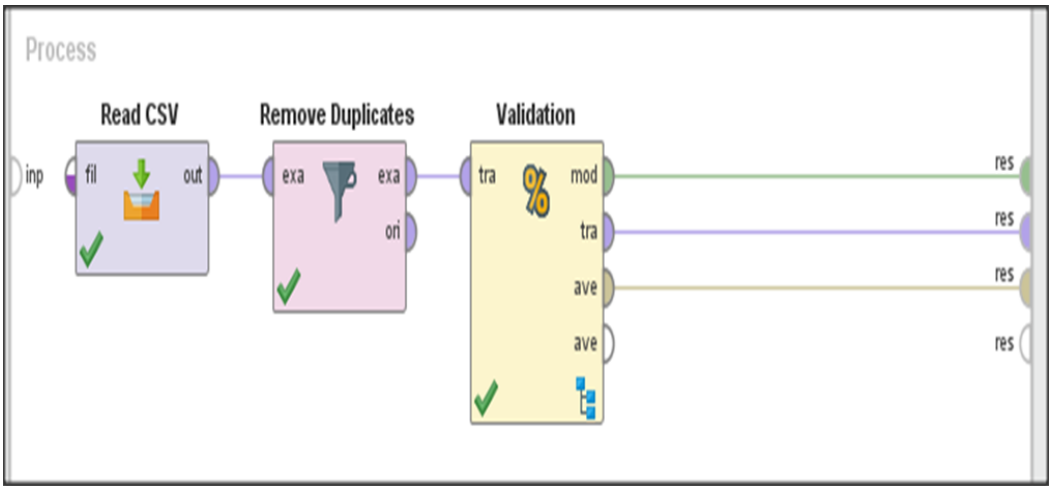
BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Eksperimen dan Pengujian Metode

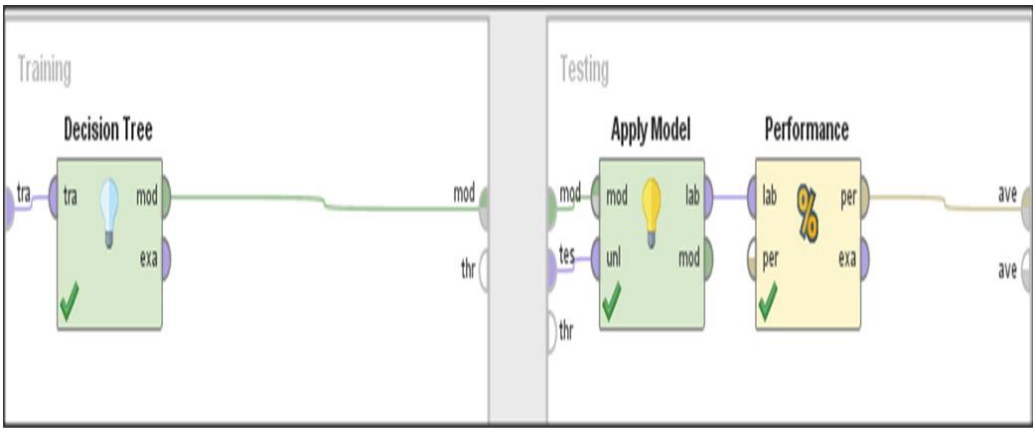
Pengujian dilakukan pada software *RapidMiner* versi 7.0 dengan 3 metode yang akan dibandingkan performanya yaitu *Support Vector Machine, Naïve Bayes dan Decision Trees*, dari pengujian tiga metode tersebut akan menjadi acuan bagi penulis menentukan metode yang paling efektif digunakan untuk mendeteksi kinerja web *phishing* dari dataset yang digunakan, hasil dari pengujian akan dijelaskan pada sub bab di bawah ini

4.1.1 Pengujian model *Decision Tree*

Pada penelitian penentuan hasil *phishing* pada website menggunakan algoritma *Decision Tree* berbasis pada framework *RapidMiner*, Data set di import ke *RapidMiner* dengan type data CSV, lalu diberikan *tools* "*Remove Duplicates*" untuk menyeleksi data yang ganda atau duplikat sehingga analasis yang dihasilkan lebih efisien ditampilkan oleh gambar 4.1, selanjutnya dataset yang sudah diseleksi dilakukan *Cross Validation* untuk menemukan performa dari pengujian seperti pada gambar 4.2



Gambar 4.1. Pengujian *Decision Tree Pada Rapidminer*



Gambar 4.2. Model pengujian validasi *Decision Tree*

Nilai *accuracy, precision*, dan *recall* dari *data training* dapat dihitung dengan menggunakan RapidMiner. Hasil pengujian dengan menggunakan model *Decision Tree* didapatkan hasil pada gambar 4.3

PerformanceVector
PerformanceVector:
accuracy: 91,84% +/- 1,32% (*mikro* 91,84%)
ConfusionMatrix:

True	*no phising*	*Phising*
no phising	2689	177
Phising	295	2624

precision: 89,95% +/- 1,80% (*mikro* 89,89%) (*positive class phising*)
ConfusionMatrix

True	*no phising*	*Phising*
no phising	2689	177
Phising	295	2624

recall: 93,68% +/- 2,91% (*mikro* 93,68%) (*positive class phising*)
ConfusionMatrix

True	*no phising*	*Phising*
no phising	2689	177
Phising	295	2624

AUC (*optimistic*): 0,991 +/- 0,003 (*mikro* 0,991) (*positive class phising*)
AUC: 0,928 +/- 0,020 (*mikro* 0,928) (*positive class phising*)
AUC (*pessimistic*): 0,865 +/- 0,038 (*mikro* 0,865) (*positive class phising*)

Gambar 4.3. Nilai *accuracy, precision*, dan *recall* Pengujian *Decision Tree*

1. *Confusion Matrix*

   Tabel 4.1. menunjukkan hasil dari confusion matrix metode *Decission Tree*.

Tabel 4.1 Hasil *Confusion Matrix* untuk Metode *Decission Tree*

accuracy: 91,84% +/- 1,32% (mikro 91,84%)

	true no phising	*true phising*	*class precision*
pred. no phising	2689	177	93,82%
pred. phising	295	2624	89,89%
class recall	90,11%	93,68%	

Jumlah *True Positive* (TP) adalah 2689 diklasifikasikan sebagai 1 sesuai dengan prediksi yang dilakukan dengan metode *Decission Tree*., lalu *False Negative* (FN) sebanyak 177 data diprediksi sebagai 1 tetapi ternyata -1, kemudian *True Negative* (TN) sebanyak 2624 data sebagai -1 sesuai dengan prediksi, dan *False Positive* (FP) sebanyak 295 data diprediksi -1 ternyata 1. Tingkat akurasi yang dihasilkan dengan menggunakan algoritma *Decission Tree*. adalah sebesar 91,84 % dan dapat dihitung untuk mencari nilai *accuracy, sensitivity, specificity*, ppv, dan npv pada persamaan di bawah ini:

$$Accuracy = \frac{\mathrm{TP+TN}}{TP+TN+FP+FN} = \frac{2689+2624}{2689+2624+295+177} = 0{,}9184$$

$$Sensitivity = \frac{TP}{TP+FN} = \frac{2689}{2689+177} = 0{,}9382$$

$$Specificity = \frac{TN}{TN+FP} = \frac{2624}{2624+295} = 0{,}8989$$

$$\mathrm{PPV} = \frac{TP}{TP+FP} = \frac{2689}{2689+295} = 0{,}9011$$

$$\mathrm{NPV} = \frac{TN}{TN+FN} = \frac{2624}{2624+177} = 0{,}9368$$

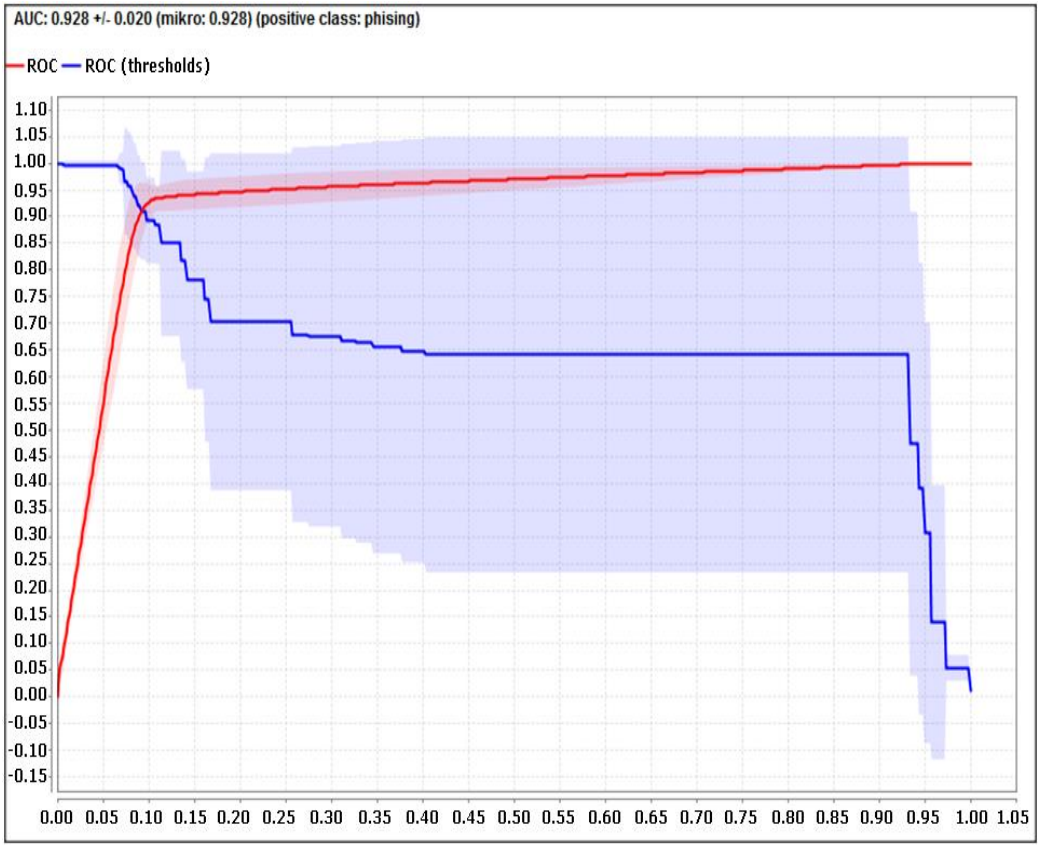
Hasil perhitungan terlihat pada tabel 4.2 di bawah ini:

Tabel 4.2 Nilai *accuracy, sensitivity, specificity, ppv* dan *npv* metode *Decission Tree*

	Nilai (%)
Accuracy	0,9184
Sensitivity	0,9382
Specificity	0,8989
PPV	0,9011
NPV	0,9368

2. Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua *class* bisa dilihat pada Gambar 4.4 yang merupakan kurva ROC untuk algoritma *Decission Tree*. Kurva ROC pada gambar 4.4 mengekspresikan *confusion matrix* dari table 4.4. Garis horizontal adalah *false positives* dan garis vertikal *true positives*. Menghasilkan nilai AUC (*Area Under Curve*) sebesar 0,928 dengan nilai akurasi klasifikasi sangat baik (*excellent classification*)



Gambar 4.4. Kurva ROC dengan Metode *Decission Tree*

4.1.2. Pengujian model *Naïve Bayes*

Nilai *accuracy*, *precision*, dan *recall* dari *data training* dapat dihitung dengan menggunakan *RapidMiner*. Hasil pengujian dengan menggunakan model *Naïve Bayes* didapatkan hasil pada gambar 4.5

PerformanceVector
PerformanceVector:
accuracy: 74,07% +/- 1,99% (mikro 74,07%)
ConfusionMatrix:

True	*no phising*	*Phising*
no phising	*2975*	*1491*
Phising	*9*	*1310*

precision: 99,30% +/- 0,74% (mikro: 99,32%) (positive class: phising)
ConfusionMatrix:

True	*no phising*	*Phising*
no phising	*2975*	*1491*
Phising	*9*	*1310*

recall: 46,77% +/- 4,02% (mikro: 46,77%) (positive class: phising)
ConfusionMatrix:

True	*no phising*	*Phising*
no phising	*2975*	*1491*
Phising	*9*	*1310*

AUC *(optimistic): 0,969 +/- 0,011 (mikro: 0,969) (positive class: phising)*
AUC*: 0,969 +/- 0,011 (mikro: 0,969) (positive class: phising)*
AUC *(pessimistic): 0,969 +/- 0,011 (mikro: 0,969) (positive class: phising)*

Gambar 4.5. Nilai *accuracy*, *precision*, dan *recall* Pengujian *Naïve Bayes*

1. *Confusion Matrix*

   Tabel 4.3. menunjukkan hasil dari *confusion matrix* metode *Naïve Bayes*.

Tabel 4.3 Hasil *Confusion Matrix* untuk Metode *Naïve Bayes*

accuracy 74,07% +/- 1,99% (mikro 74,07%)

	true no phising	*true phising*	*class precision*
pred. no phising	2975	1491	66,61%
pred. phising	9	1310	99,32%
class recall	99,70%	46,77%	

Jumlah *True Positive* (TP) adalah 2975 diklasifikasikan sebagai 1 sesuai dengan prediksi yang dilakukan dengan metode *Naïve Bayes*, lalu *False Negative* (FN) sebanyak 1491 data diprediksi sebagai 1, kemudian *True Negative* (TN) sebanyak 1310 data sebagai -1 sesuai dengan prediksi, dan *False Positive* (FP) sebanyak 9 data diprediksi -1 ternyata 1. Tingkat akurasi yang dihasilkan dengan menggunakan algoritma *Naïve Bayes* adalah sebesar 66,61% dan dapat dihitung untuk mencari nilai *accuracy, sensitivity, specificity,* ppv, dan npv pada persamaan di bawah ini:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2975+1310}{2975+1310+9+1491} = 0{,}7407$$

$$Sensitivity = \frac{TP}{TP+FN} = \frac{2975}{2975+1491} = 0{,}6661$$

$$Specificity = \frac{TN}{TN+FP} = \frac{1310}{1310+9} = 0{,}9931$$

$$PPV = \frac{TP}{TP+FP} = \frac{2975}{2975+9} = 0{,}9969$$

$$NPV = \frac{TN}{TN+FN} = \frac{1310}{1310+1491} = 0{,}4676$$

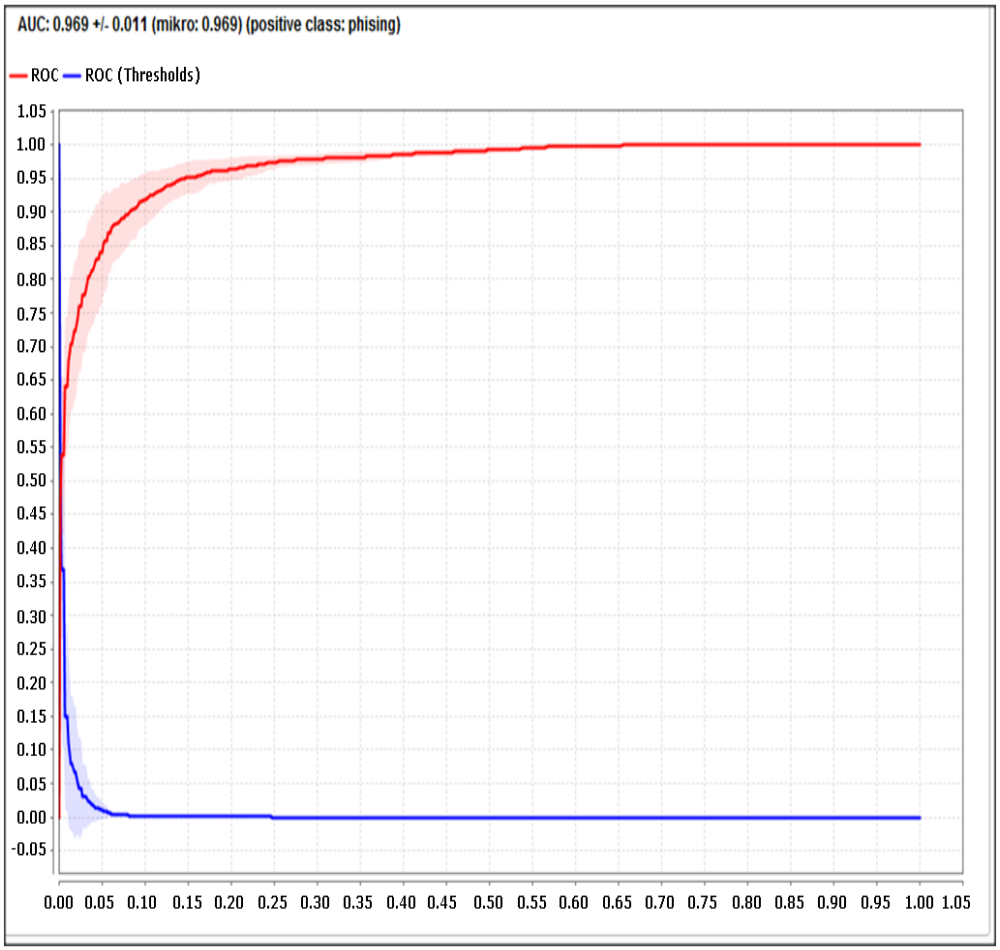
Hasil perhitungan terlihat pada tabel 4.4 di bawah ini:

Tabel 4.4 Nilai *accuracy, sensitivity, specificity*, *ppv* dan *npv* metode *Decission Tree*

	Nilai (%)
Accuracy	0,7407
Sensitivity	0,6661
Specificity	0,9931
PPV	0,9969
NPV	0,4676

2. Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua *class* bisa dilihat pada Gambar 4.6 yang merupakan kurva ROC untuk algoritma *Naïve Bayes*. Kurva ROC pada gambar 4.6 mengekspresikan confusion matrix dari tabel 4.4. Garis horizontal adalah *false positives* dan garis vertikal *true positives*. Menghasilkan nilai AUC (*Area Under Curve*) sebesar 0,928 dengan nilai akurasi klasifikasi sangat baik (*excellent classification*).



Gambar 4.6. Kurva ROC dengan Metode *Naïve Bayes*

4.1.3. Pengujian model *Support Vector Machine*

Nilai *accuracy*, *precision*, dan *recall* dari *data training* dapat dihitung dengan menggunakan RapidMiner. Hasil pengujian dengan menggunakan model *Naïve Bayes* didapatkan hasil pada gambar 4.7.

PerformanceVector
PerformanceVector:
accuracy: 92,34% +/- 1,02% (mikro: 92,34%)
ConfusionMatrix:

True	*no phising*	*phising*
no phising	2682	141
phising	302	2660

precision: 89,80% +/- 0,79% (mikro: 89,80%) (positive class: phising)
ConfusionMatrix:

True	*no phising*	*phising*
no phising:	2682	141
phising:	302	2660

recall: 94,97% +/- 1,63% (mikro: 94,97%) (positive class: phising)
ConfusionMatrix:

True	*no phising*	*phising*
no phising	2682	141
phising	302	2660

AUC (optimistic): 0,977 +/- 0,007 (mikro: 0,977) (positive class: phising)
AUC: 0,977 +/- 0,007 (mikro: 0,977) (positive class: phising)
AUC (pessimistic): 0,977 +/- 0,007 (mikro: 0,977) (positive class: phising)

Gambar 4.7. Nilai *accuracy, precision* dan *recall* Pengujian *Support Vector Machine*

1. *Confusion Matrix*

   Tabel 4.5. menunjukkan hasil dari *confusion matrix* metode *Support Vector Machine*

Tabel 4.5 Hasil *Confusion Matrix* untuk Metode *Support Vector Machine*

accuracy: 92,34% +/- 1,02% (mikro 92,34%)

	true no phising	*true phising*	*class precision*
pred. no phising	2682	141	95,01%
pred. phising	302	2660	89,80%
class recall	89,88%	94,97%	

Jumlah *True Positive* (TP) adalah 2682 diklasifikasikan sebagai 1 sesuai dengan prediksi yang dilakukan dengan metode *Support Vector Machine*, lalu *False Negative* (FN) sebanyak 141 data diprediksi sebagai 1

tetapi ternyata -1, kemudian *True Negative* (TN) sebanyak 2660 data sebagai -1 sesuai dengan prediksi, dan *False Positive* (FP) sebanyak 302 data diprediksi -1 ternyata 1. Tingkat akurasi yang dihasilkan dengan menggunakan algoritma *Support Vector Machine* adalah sebesar 95,01% dan dapat dihitung untuk mencari nilai *accuracy*, *sensitivity*, *specificity*, ppv, dan npv pada persamaan di bawah ini:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2682+2660}{2682+2660+302+141} = 0{,}9234$$

$$Sensitivity = \frac{TP}{TP+FN} = \frac{2682}{2682+141} = 0{,}9500$$

$$Specificity = \frac{TN}{TN+FP} = \frac{2660}{2660+302} = 0{,}8988$$

$$PPV = \frac{TP}{TP+FP} = \frac{2682}{2682+302} = 0{,}8987$$

$$NPV = \frac{TN}{TN+FN} = \frac{2660}{2660+141} = 0{,}9496$$

Hasil perhitungan terlihat pada tabel 4.6 di bawah ini:

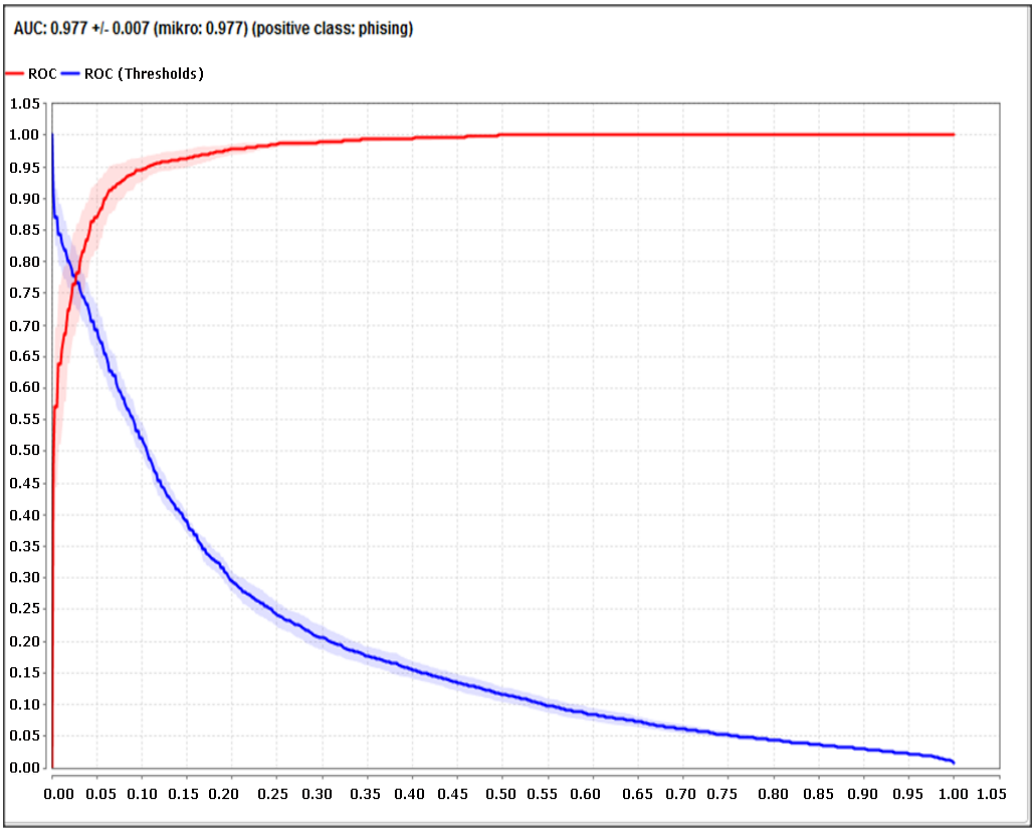
Tabel 4.6 Nilai *accuracy, sensitivity, specificity, ppv* dan *npv* metode *Decission Tree*

	Nilai (%)
Accuracy	0,9234
Sensitivity	0,9500
Specificity	0,8987
PPV	0,8371
NPV	0,9496

2. Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC. Perbandingan kedua class bisa dilihat pada Gambar 4.8 yang merupakan kurva ROC untuk algoritma *Support Vector Machine*. Kurva ROC pada gambar 4.6

mengekspresikan *confusion matrix* dari tabel 4.6. Garis horizontal adalah *false positives* dan garis vertikal *true positives*. Menghasilkan nilai AUC (*Area Under Curve*) sebesar 0,928 dengan nilai akurasi klasifikasi sangat baik (*excellent classification*).

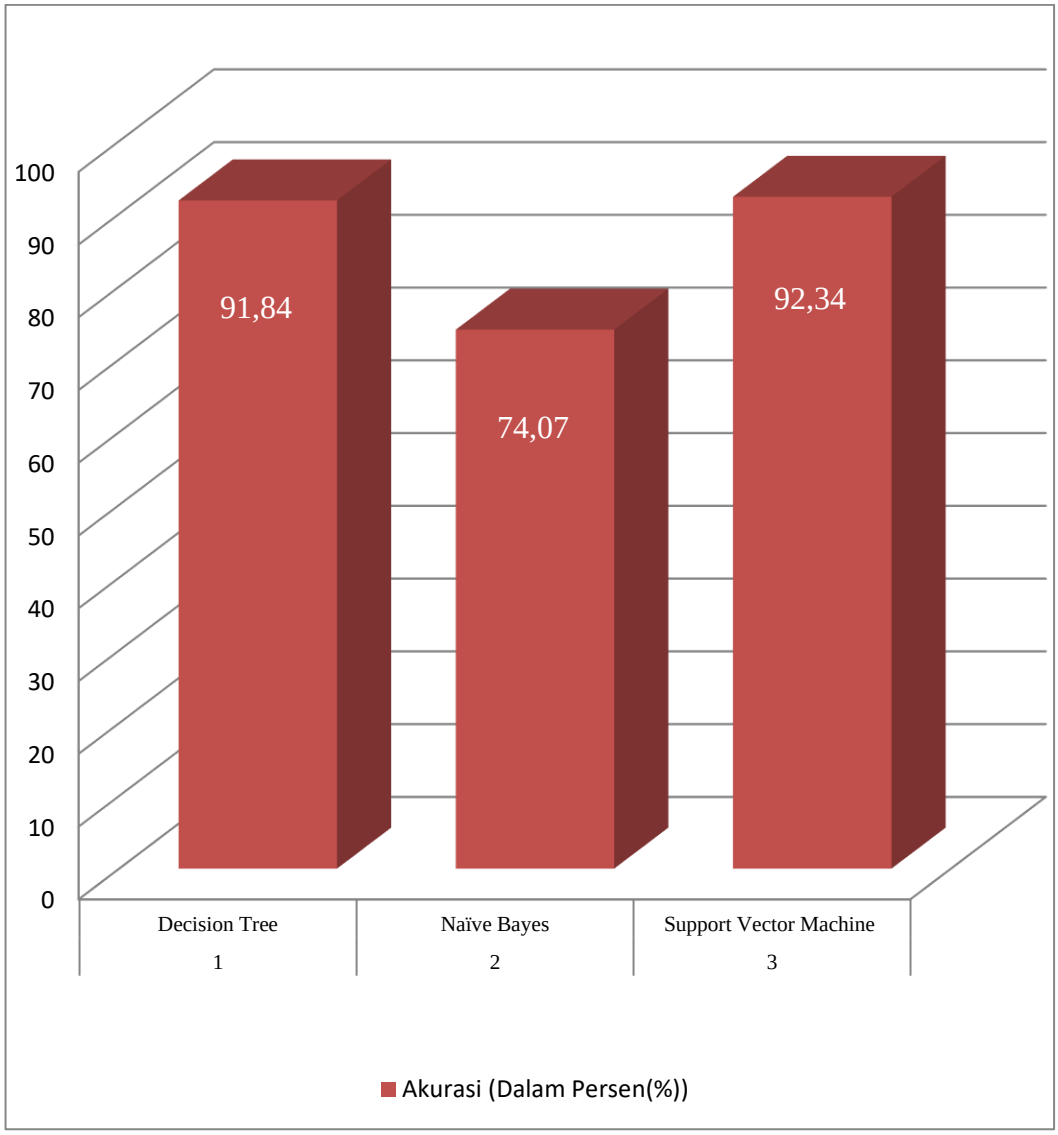


Gambar 4.8. Kurva ROC dengan Metode *Support Vector Machine*

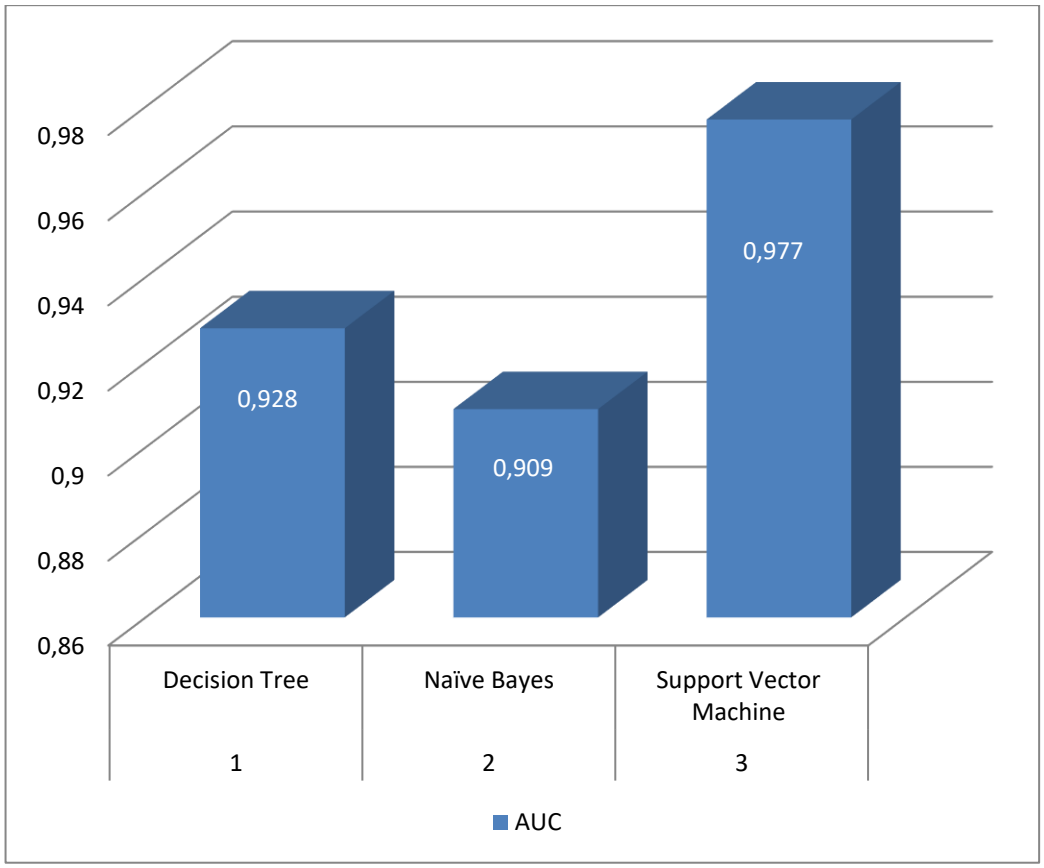
4.2. Pembahasan

Berdasarkan hasil perhitungan yang dilakukan untuk memecahkan masalah prediksi website *phishing* dapat menggunakan metode *Decision Tree* mempunyai tingkat akurasi sebesar 91,84 % dan mempunyai nilai AUC sebesar 0,928, kemudian dicoba dengan metode *Naïve Bayes* mempunyai tingkat akurasi sebesar 74,07 % dan mempunyai nilai AUC sebesar 0,909, dan kemudian dicoba dengan metode *Support Vector Machine* mempunyai tingkat akurasi sebesar 92,34% dan mempunyai nilai AUC sebesar 0,977 disimpulkan bahwa hasil perhitungan menyatakan bahwa menggunakan metode *Support Vector Machine* mempunyai tingkat akurasi lebih baik dibandingkan metode *Decision Tree dan*

metode Naïve Bayes. Hal ini menunjukan bahwa metode *Support Vector Machine* mempunyai tingkat akurasi yang paling tinggi.



Gambar 4.9. Grafik Akurasi Metode *Support Vector Machine, Naïve Bayes dan Decision Trees*



Gambar 4.10. Grafik perbandingan hasil AUC Metode *Support Vector Machine, Naïve Bayes dan Decision Trees*

4.3. Implikasi Penelitian

Hasil dari penelitian ini mempunyai implikasi terhadap penilaian kredit yaitu meliputi implikasi terhadap aspek system pendukung keputusan penilaian kredit, terhadap aspek manajerial dan terhadap aspek penelitian-penelitian selanjutnya yang akan diuraikan di bawah ini.

1. Aspek Sistem

   Penerapan hasil prediksi terhadap Website Phishing dapan menjadi acuan untuk meningkatkan kewaspadaan terhadap phishing melalui penerepan setiap variable yang digunakan sebagai label pengujian, dimana phishing sudah menjadi ancaman yang sangat berbahaya khususnya di Institusi keuangan dimana data yang sangat rahasia bisa dicuri dan berdampak pada beberapa pihak yang dirugikan

2. Aspek Manajerial

   Dari hasil penelitian ini diketahui bahwa metode *Support Vector Machine* mendukung pengambilan keputusan dan pengembangan sistem informasi manajemen pada lembaga keuangan dan perbankan dengan menggunakan bantuan software *RapidMiner*, untuk itu diperlukan peningkatan kemampuan manajerial dari seorang manajer dan juga karyawan yang bersangkutan agar mampu membuat perencanaan secara formal, mengerjakan dan mengoperasikan sistem dengan baik dan benar. Hal tersebut dapat dilakukan dengan pelatihan atau *training*.

3. Penelitian Lanjutan

   Penelitian dapat menjadi acuan selanjutnya dalam membangun *system* yang lebih aman terhadap serangan *phishing* sehingga tidak ada kerugian secara material dan non material yang dialami oleh pengguna layanan *website*

BAB V

KESIMPULAN DAN SARAN

5.1. Kesimpulan

Phishing sudah menjadi masalah yang sangat rentan di dunia, dalam penelitian ini dilakukan pengujian model berbasiskan metode *Decision Tree,* metode *Naïve Bayes, dan* metode *Support Vector Machine* menggunakan framework RapidMiner Versi 7.0 didapat hasil eksperiment menggunakan metode *Decision Tree* mempunyai tingkat akurasi sebesar 91,84 % dan mempunyai nilai AUC sebesar 0,928, kemudian dicoba dengan metode *Naïve Bayes* mempunyai tingkat akurasi sebesar 74,07 % dan mempunyai nilai AUC sebesar 0,909, dan kemudian dicoba dengan metode *Support Vector Machine* mempunyai tingkat akurasi sebesar 92,34% dan mempunyai nilai AUC sebesar 0,977, Maka dapat disimpulkan pengujian pengujian dataset website phishing UCI menggunakan metode *Decision Tree,* metode *Naïve Bayes, dan* metode *Support Vector Machine* didapat bahwa pengujian *Support Vector Machine* lebih baik dari pada *Decision Tree* dan *Naïve Bayes,* Dengan demikian dari hasil pengujian model di atas dapat disimpulkan bahwa *Support Vector Machine* memberikan pemecahan untuk permasalahan prediksi *Website Phishing* lebih akurat. Hal ini karena metode *Support Vector Machine* mempunyai tingkat akurasi yang paling tinggi.

5.2. Saran

Agar penelitian ini bisa ditingkatkan, pengukuran kinerja sebuah algoritma *data mining* dapat dilakukan berdasarkan beberapa kriteria antar lain akurasi, kecepatan komputasi, *robustness*, skalabilitas dan interpretabilitas. Penelitian ini menggunakan satu kriteria yaitu berdasarkan akurasi, akan lebih baik jika semua kriteria diuji coba agar algoritma yang diteliti lebih teruji kinerjanya. Akurasi sebuah algoritma bisa ditingkatkan dengan menggunakan beberapa teknik antara lain teknik *bagging* dan *boosting*.

Penelitian ini juga belum menggunakan kedua teknik tersebut untuk meningkatkan akurasi karena penelitian ini hanya terbatas pada perbandingan algoritma *Suppot Vector Machine, Decision Tree* dan *Naïve Bayes.*

DAFTAR PUSTAKA

[1]. Antonio San Martino, X. P. (2010), "Phishing Secrets: History, Effects, and Countermeasures. International Journal of Network Security", 11, 163–171.

[2]. Aboli Bhanji, S. B. (2013), "Secure Server Verification By Using RSA Algorithm And Visual Cryptography. International Journal of Engineering Research & Technology (IJERT).

[3]. Lance James and Joe Stewart, (2005). "Phising Exposed" (Book)

[4]. He Chunjian dan Zhang Cuilian Zhao Yan (2009), "A New SVM Merged into Data Information".

[5]. Vladimir N. Vapnik (1999), "An Overview of Statistical Learning Theory" IEEE Transactions On Neural Networks, Vol. 10.

[6]. Jianchao Han, Juan C. Rodriguze, Mohsen Beheshti, (2008), "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner" ,Second International Conference on Future Generation Communication and Networking.

[7]. Sholom M. Weiss, Indurkhya, & Zhang "Text Mining: Predictive Methods for Analyzing Unstructured Information" Associate Professor at the Department of Statistics and Biostatistics at Rutgers University, New Jersey, Book Springer Science & Business Media, 2010

[8]. Zhao, Fujiang, & Zhou, (2011) Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes, Expert Systems with Applications 38 (2011) 5197–5204

[9]. Shih Wei Lin,Yeou Ren Shiue, Shih Chi Chen, Hui Miao Cheng, (2009), "Applying enhanced data mining approaches in predicting bank performance A case of Taiwanese commercial banks". Expert Systems with Applications 36 (2009) 11543–11551

[10]. Oded Maimon, Lior Rokach, (2010). "Data Mining and Knowledge Discovery Handbook" (Second Edition, Book)

[11]. YuanningLIU, GangWANG, XiaodongZHU, HuilingCHEN, ZhenLIU, Zhengdong ZHAO, (2011). "An Adaptive Fuzzy Ant Colony Optimization for Feature Selection". Journal of Computational Information Systems 7:4 (2011) 1206-1213

[12]. V.Ramakanth, Neela Megha Shyam Desai And T.Shyam Prasad "A Survey On Attacks And Defense Mechanisms In Phishing" International Journal Of Research And Applications, 2014

[13]. James Luke, Suharjito, “Data mining of automatically promotion tweet for products and services using Naïve Bayes algorithm to increase twitter engagement followers at PT. Bobobobo”, International Conference on Computer Science and Computational Intelligence (ICCSCI), 2015

[14]. Dharm Singh, Naveen Choudhary & Jully Samota “Analysis of Data Mining Classification with Decision tree Technique”,Maharana Pratap University of Agriculture and Technology, India, 2013

[15]. Isredza Rahmi A Hamid1,2, Jemal Abawajy1 , Tai-hoonn Kim “Using Feature Selection and Classification Scheme for Automating Phishing Email Detection”, School of Information Technology, Deakin University, Waurn Ponds, VIC., 3217, Australia, 2016

[16]. Ian H. Witten, Eibe Frank, Mark A. Hall. (2011), “Data Mining Practical Machine Learning Tools and Techniques” (Third Edition, Book).

[17]. Veronika S. Moertini, (2002), “Data Mining sebagai Solusi Bisnis”. INTEGRAL, Majalah Ilmiah Matematika dan Ilmu Pengetahuan Alam, Vol. 7 No. 1, April 2002, ISSN 1410-1335

[18]. Bellazzi & Zupan, (2008) “Medical data analysis and construction of predictive models”. International Journal of Medical Informatics 77 (2008) 81-97

[19]. Jiawei Han and Micheline Kamber, (2000), “Data Mining- Concepts and Techniques”

[20]. Ethem Alpaydin, (2014), “Introduction to Machine Learning” (Third Edition, Book).

[21]. Daniel t. Larose, (2007), “Data Mining Methods And Models”. Department of Mathematical Sciences Central Connecticut State University. (Book)

[22]. Florin Gorunescu, (2010), “Data Mining Concepts Models”. (Book)

[23]. Yulin Dong, Zhonghang Xia, Shandong Qingdao, Manghui Tu Guangming Xing, (2007). “An optimization method for selecting parameters in support vector machines”. Sixth International Conference on Machine Learning and Applications.

[24]. Christopher J.C. Burges, (1998), “A Tutorial on Support Vector Machines for Pattern Recognition” Data Mining and Knowledge Discovery, 2, 121–167 (1998) Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

[25]. Zenglin Xu, Kaizhu Huang, Jianke Zhu, Irwin King, Michael R Lyu, (2009), “Novel Kernel Based Maximum A Posteriori Classification Method”

[26]. Anto Satriyo Nugroho, Arief Budi Witarto, Dwi Handoko, (2003), “Support Vector MachineTeori dan Aplikasinya Dalam Bioinformatika”. Kuliah Umum IlmuKomputer.Com

[27]. Carlo Vercellis, (2009), “Business Intelligence”. Data Mining and Optimizationfor Decision Making” (Book)

[28]. Gregor Polancic, Marjan Hericko, Ivan Rozman, (2010). “An empirical examination of application frameworks success based on technology acceptance model”. The Journal of Systems and Software 83 (2010) 574–584

[29]. Mikael Berndtsson, Jörgen Hansson, Björn Olsson, Björn Lundell, (2008), “Thesis Projects A Guide for Students in Computer Science and Information Systems. (Book)

[30]. Dr. Catherine Dawson, (2009), “Introduction to Research Methods”. (Fourth Editon) A Practical guide for undertaking a research project.