Prediksi Website Pemancing Informasi Penting Phising Menggunakan Support Vector Machine (SVM)

Zuhri Halim [1,*]
[1] Teknik Informatika; Universitas Muhammadiyah Prof. DR. HAMKA Jakarta; Jl. Tanah Merdeka No. 6 Kampung Rambutan, Pasar Rebo, Jakarta Timur (021) 87782739;
e-mail: zuhri@uhamka.ac.id

Korespondensi: email: zuhri@uhamka.ac.id

Diterima: 17 November 2017 ; Review: 23 November 2017 ; Disetujui: 27 November 2017

Cara sitasi: Halim Z. 2017. Prediksi Website Pemancing Informasi Penting Phising Menggunakan Support Vector Machine (SVM). Information System for Educators and Professionals. 2 (1): 71 – 82.

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

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

E-ISSN: 2548-3587 ; 71 – 82

1. Pendahuluan

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 [Martino and Perramon, 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* [Long, 2008] 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 [Bhanji et al., 2013], 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 [James, 2005].

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 Chunjiang 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 et al., 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 et al., 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 [Weiss, 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 et al., 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 et al., 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 et al., 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 and Rokach, 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, 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 and Rokach, 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.*

2. Metode Penelitian

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. 4) Evaluasi dan validasi, pada tahap ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

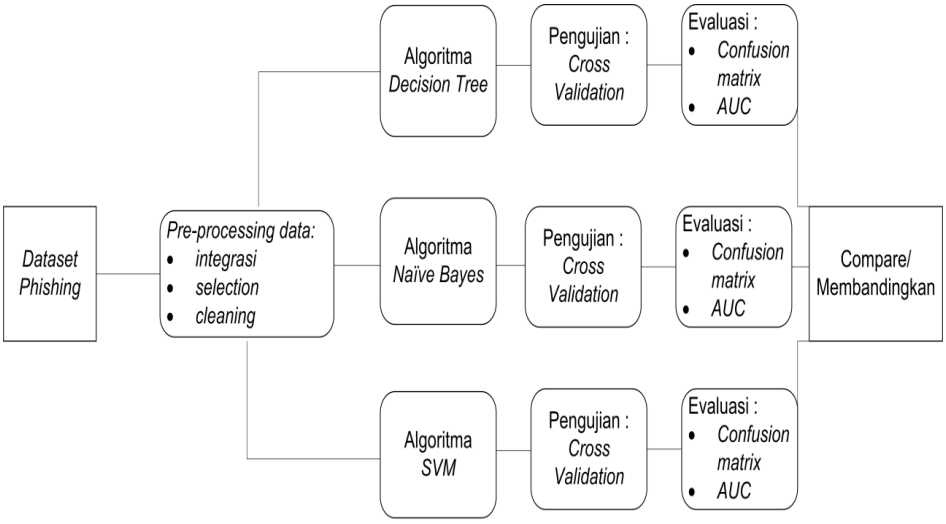
Pengolahan Data Awal

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 [Vercellis, 2009]: Pertama, Data *validation*, untuk mengidentifikasikan dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*). Kedua, Data *integration and transformation,* untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan kedalam *software RapidMiner*. Ketiga, *Data size reduction and discritization*, untuk memperoleh data set dengan jumlah atribut dan *record* yang lebih sedikit tetapi bersifat informative.

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 1 di bawah ini.

E-ISSN: 2548-3587 ; 71 – 82



Sumber: Hasil Penelitian (2016)

Gambar 1. Metode yang diusulkan

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 adalah 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 1 di bawah ini:

Tabel 1. 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

Sumber: Hasil Penelitian (2016)

Evaluasi dan Validasi Hasil

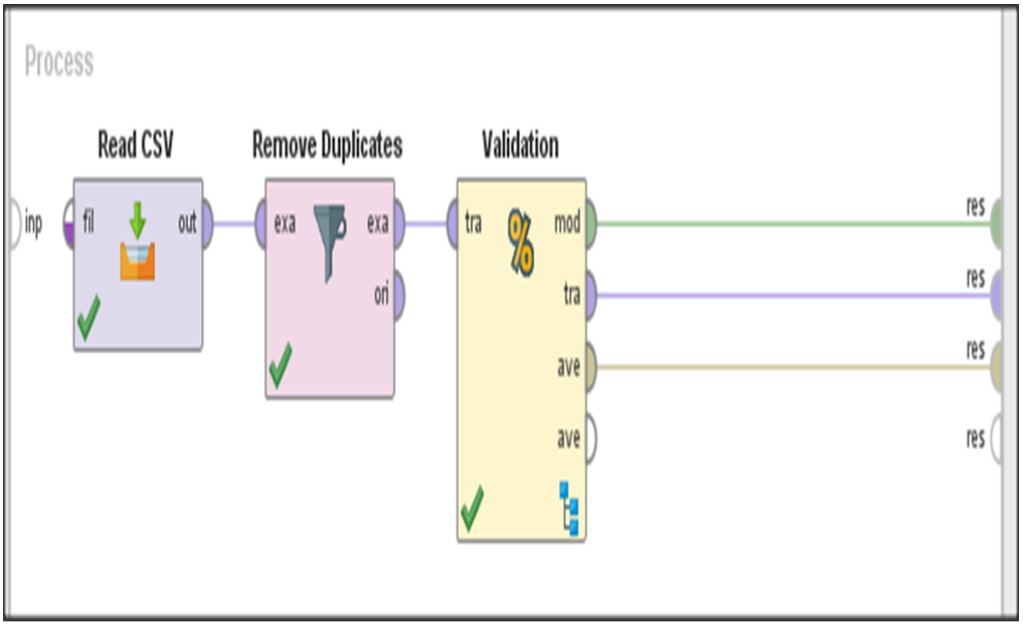
Model yang diusulkan pada penelitian tentang prediksi Situs *Phishing* adalah dengan menerapkan *Support Vector Machine, Decision Tree* dan *Naïve Bayes.* Penerapan algoritma *Support Vector Machine*, *Decision Tree* dan *Naïve Bayes* dengan menentukan nilai *weight* terlebih dahulu. Setelah didapatkan 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.

3. Hasil dan Pembahasan

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

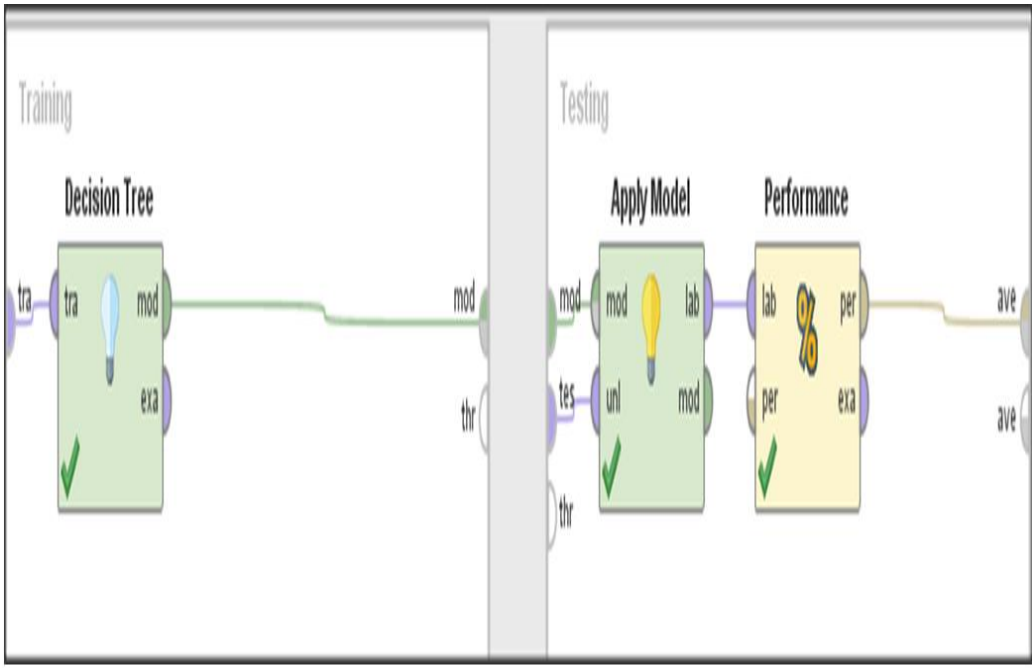
Pengujian model *Decision Tree*

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 2, selanjutnya dataset yang sudah diseleksi dilakukan *Cross Validation* untuk menemukan performa dari pengujian seperti pada gambar 3.



Sumber : Hasil Penelitian (2016)

Gambar 2 Pengujian *Decision Tree Pada Rapidminer*



Sumber : Hasil Penelitian (2016)

Gambar 3 Model pengujian validasi *Decision Tree*

E-ISSN: 2548-3587 ; 71 – 82

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

Sumber : Hasil Penelitian (2016)

Gambar 4 Nilai *accuracy*, *precision*, dan *recall* Pengujian *Decision Tree*

1. *Confusion Matrix*
   Tabel 2 menunjukkan hasil dari confusion matrix metode *Decission Tree*.

Tabel 2 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%	

Sumber : Hasil Penelitian (2016)

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:

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

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

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

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

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

Hasil perhitungan terlihat pada tabel 3 dibawah ini:

Tabel 3 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

Sumber : Hasil Penelitian (2016)

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 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)

Sumber: Hasil Penelitian (2016)

Gambar 5 Nilai *accuracy*, *precision*, dan *recall* Pengujian *Naïve Bayes*

E-ISSN: 2548-3587 ; 71 – 82

1. *Confusion Matrix*
   Tabel 4 menunjukkan hasil dari *confusion matrix* metode *Naïve Bayes*.

Tabel 4 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%	

Sumber : Hasil Penelitian (2016)

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:

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

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

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

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

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

Hasil perhitungan terlihat pada tabel 5 dibawah ini:

Tabel 5 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

Sumber : Hasil Penelitian (2016)

Pengujian model *Support Vector Machine*

Nilai *accuracy*, *precision*, dan *recall* dari *data training* dapat dihitung dengan menggunakan RapidMiner. Hasil pengujian dengan menggunakan model *Support Vector Machine* didapatkan hasil pada gambar 6.

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)

Sumber : Hasil Penelitian (2016)

Gambar 6. Nilai *accuracy*, *precision* dan *recall* Pengujian *Support Vector Machine*

1. *Confusion Matrix*
   Tabel 6. menunjukkan hasil dari *confusion matrix* metode *Support Vector Machine*

Tabel 6 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:

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

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

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

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

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

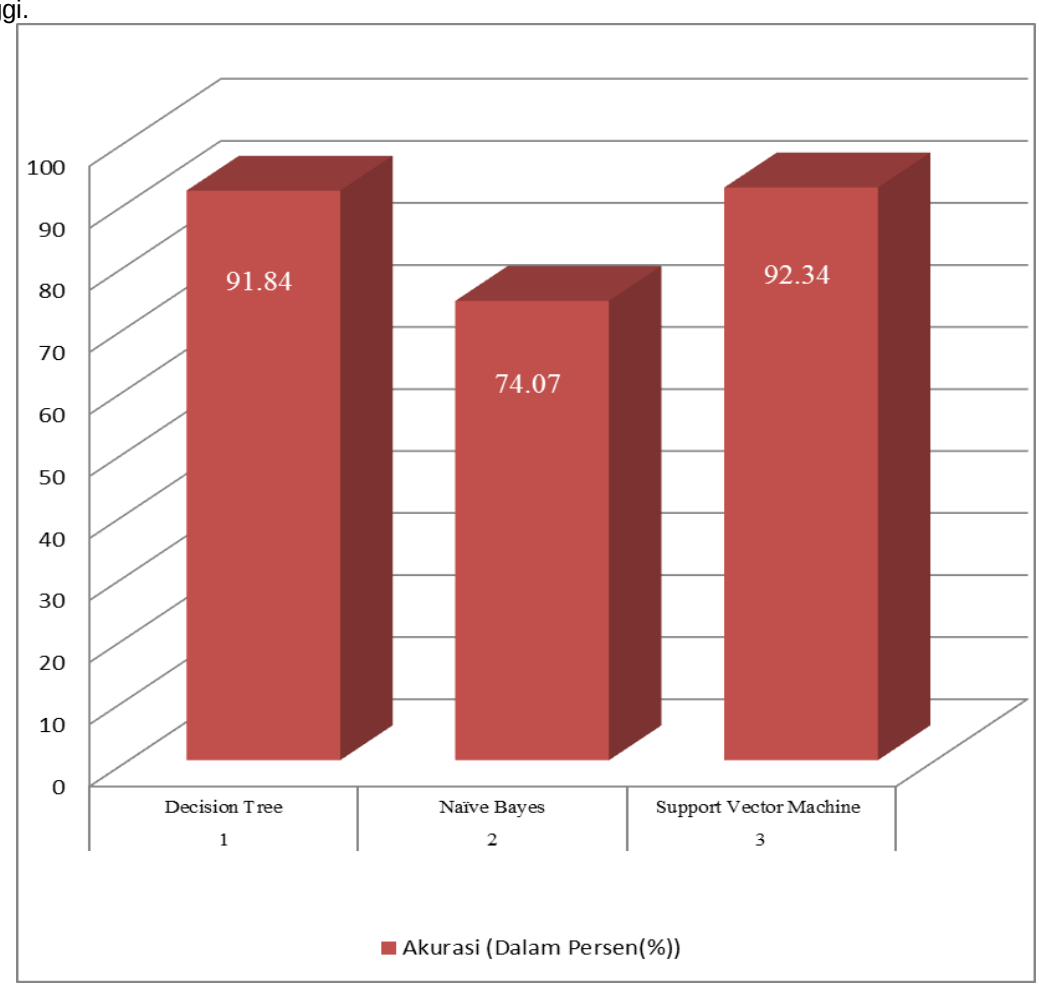
E-ISSN: 2548-3587 ; 71 – 82

Hasil perhitungan terlihat pada tabel 7 dibawah ini:

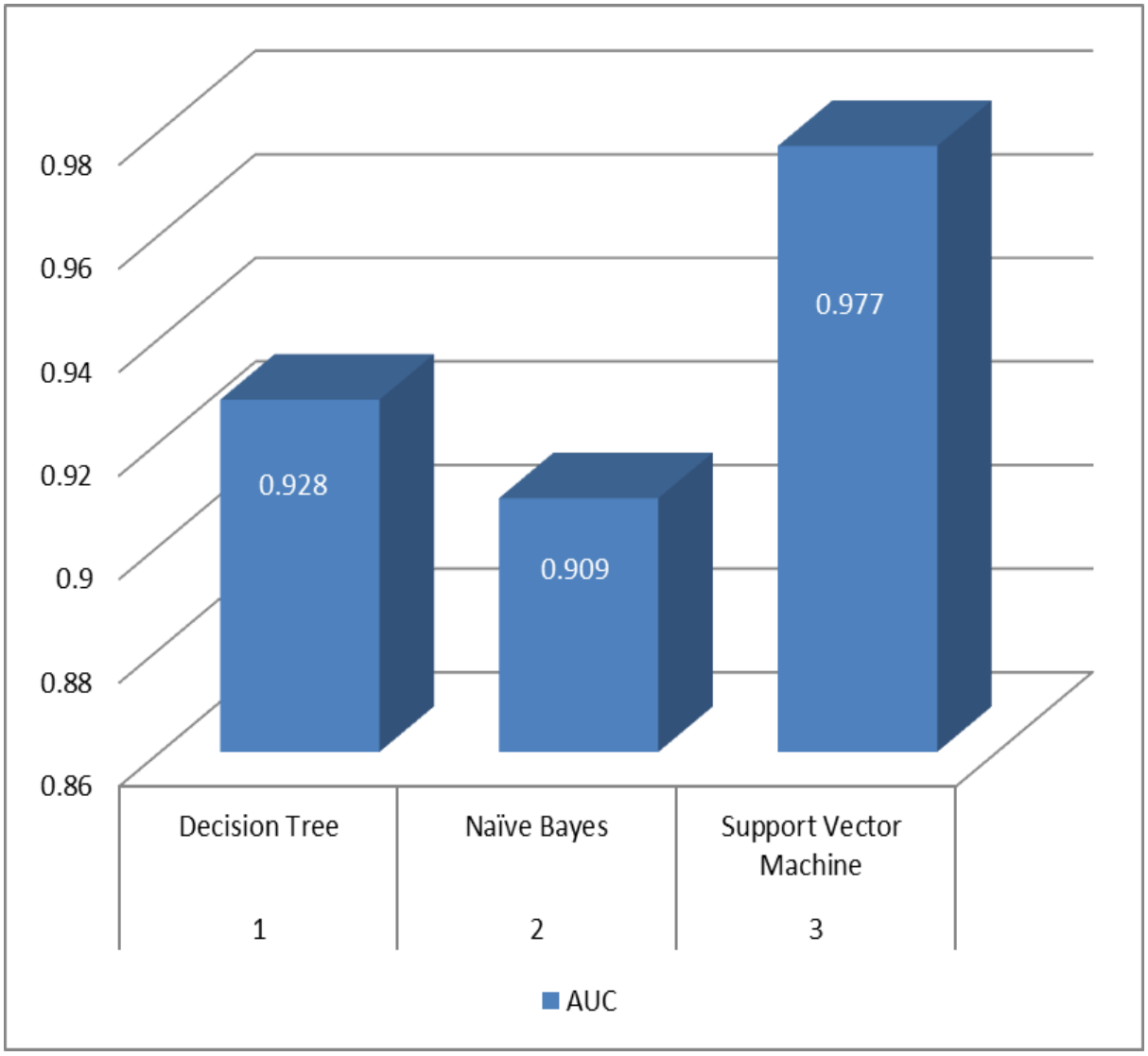
Tabel 7 Nilai *accuracy, sensitivity, specificity, ppv* dan *npv* metode *Decission Tree*

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

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



Gambar 8. Grafik perbandingan hasil AUC Metode *Support Vector Machine, Naïve Bayes dan Decision Trees*

4. 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.

E-ISSN: 2548-3587 ; 71 – 82

Referensi

Bhanji A, Jadhav P, Bhujbal S, Mulak P, Phishing K-, Introduction I. 2013. ER ER. 2: 2340–2347.

Chunjiang H, Cuilian Z, Yan Z. 2009. A New SVM Merged into Data Information. 2009 Asia-Pacific Conf. Inf. Process. I: 14–17.

Han J, Rodriguze JC, Beheshti M. 2008. Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner.

James L. 2005. Phising Exposed. Stewart J, editor. United States. 1-382 p.

Lin S, Shiue Y, Chen S, Cheng H. 2009. Expert Systems with Applications Applying enhanced data mining approaches in predicting bank performance : A case of Taiwanese commercial banks. 36: 11543–11551.

Liu Y. 2011. An adaptive fuzzy ant colony optimization for feature selection An Adaptive Fuzzy Ant Colony Optimization for Feature Selection. 1–8.

Long J. 2008. No Tech Hacking: A Guide to Social Engineering, Dumpster Diving, and Shoulder Surfing. Pinzon Scott, editor. United States: Andrew Williams. 1-285 p.

Maimon O, Rokach L. 2010. Data Mining and Knowledge Discovery Handbook, Second. Rokach L, editor. 21-36 p.

Martino AS, Perramon X. 2010. Phishing Secrets : History , Effects , and Countermeasures. 11: 163–171.

Vapnik VN. 1999. An Overview of Statistical Learning Theory. 10: 988–999.

Vercellis C. 2009. Business Intelligence: Data Mining and Optimization for Decision Making. Italy. 1-417 p.

Weiss S. 2010. Text Mining : Predictive Methods for Analysis and Prediction Model Discovery Using RapidMiner. Indurkhya, editor. New Jersey: Springer Science & Business Media. 1-237 p.

Zhao M, Fu C, Ji L, Tang K, Zhou M. 2011. Expert Systems with Applications Feature selection and parameter optimization for support vector machines : A new approach based on genetic algorithm with feature chromosomes. 38: 5197–5204.