**Jurnal Infortech**
Volume 1 No. 2 Desember 2019
E-ISSN: 2715-8160

# Forward Selection pada Support Vector Machine untuk Memprediksi Kanker Payudara

**Hani Harafani**

Teknik Informatika STMIK Nusamandiri
e-mail: hani.hhf@nusamandiri.ac.id

| Diterima | Direvisi | Disetujui |
|---|---|---|
| 15-12-2019 | 16-12-2019 | 17-12-2019 |

**Abstrak** - Kanker payudara merupakan masalah kesehatan yang serius, sehingga deteksi dini dari kanker payudara dapat berperan penting dalam perencanaan pengobatan. Pada penelitian ini Support Vector Machine dengan kernel (dot, polynomial, RBF) dan forward selection akan dipercobakan untuk memprediksi kanker payudara. Selain itu, hasil percobaan prediksi kanker payudara juga akan dibandingkan dengan empat algoritma lainnya yaitu *Neural Network, Random Forest, Decision Tree* dan KNN. Dataset kanker payudara Coimbra diambil dari *UCI Machine learning repository.* Hasil percobaan menunjukkan perbandingan akurasi SVM tanpa forward selection dengan menggunakan forward selection terdapat selisih yang besar. Hasil penelitian menunjukkan SVM dengan Kernel Dot unggul dalam akurasi dibandingkan ketiga kernel lainnya sebelum diterapkan forward selection, namun SVM(RBF)+FS unggul dengan akurasi 85,38% dibandingkan dengan SVM (Polynomial & dot), selain itu SVM(RBF)+FS juga unggul dibandingkan empat algoritma machine learning lainnya yang dipercobakan dalam memprediksi dataset kanker payudara Coimbra yang mana *Neural network* menempati urutan ke tiga dalam akurasi, *Random Forest* menempati urutan ke tujuh, *Decision Tree* menempati urutan ke Sembilan, dan KNN(k=5) menempati urutan terakhir dari perbandingan akurasi yaitu sebesar 50%.

Kata Kunci: Kanker, SVM, Fitur

## PENDAHULUAN

Kanker payudara menurut komunitas kanker Amerika dalam (Singh, 2019) merupakan masalah kesehatan yang serius, karena kanker payudara menyumbang lebih dari 1,6% dari total angka kematian wanita di seluruh dunia. Kanker payudara merupakan salah satu tipe kanker dimana ada pertumbuhan *cell* kanker yang tidak terkendali yang terbentuk pada jaringan payudara(Bustamam, Bachtiar, & Sarwinda, 2019). Pertumbuhan sel kanker akan membentuk benjolan yang dapat menyebar ke jaringan lain di dalam tubuh, yang sering disebut sebagai tumor ganas.

Berdasarkan data statistik GLOBOCAN 2018 dalam (Bustamam et al., 2019), terdapat 2,1 juta kasus kanker dan 626 ribu tidak selamat. Jumlah kematian diperkirakan akan terus meningkat seiring waktu jika tidak ada penanganan yang tepat(Bustamam et al., 2019). Sehingga deteksi dini dari kanker payudara dapat berperan penting dalam perencanaan pengobatan(Singh, 2019).

Sebanyak 30% hingga 50% dari kanker ini dapat dicegah dengan berbagai cara diantaranya dengan metodologi yang berhubungan dengan kedokteran dan kesehatan seperti *cancer screening*, data analisis(Shukla, Hagenbuchner, Win, & Yang, 2018) atau dengan metode komputasi seperti machine learning (Tapak et al., 2018).

Beberapa metode machine learning yang banyak dipilih untuk menyelesaikan kasus prediksi kanker payudara ini terutama algoritma klasifikasi diantaranya Artificial Neural Network(Jafari-Marandi, Davarzani, Soltanpour Gharibdousti, & Smith, 2018), Deep learning (Levine et al., 2019), Neural Network(Ellmann et al., 2019), dan Support Vector Machine(Bustamam et al., 2019; Singh, 2019; Tapak et al., 2018) bahkan Fuzzy Logic(Nilashi, Ibrahim, Ahmadi, & Shahmoradi, 2017).

Multilayer perceptron menurut Pan, Iplikci, Warwick, & Aziz dalam (Harafani & Wahono, 2015) merupakan salah satu model yang paling popular dari ANN memiliki kelebihan untuk menemukan pola dari data yang terlalu rumit untuk diketahui oleh manusia atau dengan teknik komputasi lainnya, selain itu MLP memiliki kekurangan yaitu sulit untuk menemukan pola bila data berdimensi tinggi.

Support Vector Machine (SVM) menurut Maimon & Rokach dalam (Harafani & Wahono, 2015) disebutkan memiliki keunggulan mengatasi masalah klasifikasi dan regresi linier maupun nonlinier yang dapat menjadi satu kemampuan algoritma pembelajaran untuk klasifikasi serta regresi, selain itu SVM juga memiliki akurasi yang tinggi dan tingkat kesalahan yang relative kecil, dan kemampuan untuk mengatasi *overfitting*. Meskipun SVM memiliki banyak kelebihan dibandingkan machine learning lainnya, SVM juga memiliki kelemahan salah satunya

adalah kesulitan dalam memilih fitur untuk input yang optimal(Ilhan & Tezel, 2013). Terlebih lagi, data kanker mempunyai banyak fitur yang berisi informasi tentang kanker itu sendiri, tetapi tidak semua fitur merupakan fitur yang relevant(Bustamam et al., 2019), jadi, seleksi fitur dibutuhkan untuk dapat meningkatkan kinerja machine learning dalam memprediksi kanker payudara.

Ada beberapa metode seleksi fitur yang banyak direkomendasikan oleh peneliti dunia, diantaranya *backward feature Eliminations* (Li et al., 2014), dan *forward feature selection* (Meyer, Reudenbach, Hengl, Katurji, & Nauss, 2018)(Balram, Lian, & Sebastian, 2019)(Bergman, Schrempf, Kosiol, & Vogl, 2018). Di dalam penelitian ini kami menggunakan *forward feature selection* untuk memperbaiki akurasi prediksi daripada machine learning.

## METODE PENELITIAN

### 1. Dataset Kanker Payudara Coimbra

Pada penelitian ini, data yang digunakan adalah dataset kanker payudara Coimbra yang didapat dari situs UCI Machine Learning Repository yang dapat diunduh pada laman https://archive.ics.uci.edu/ml/machine-learning-databases/00451/.

Pada dataset kanker payudara Coimbra terdapat 10 atribut yang terdiri dari 9 atribut bebas dan satu label (Patrício et al., 2018). 9 atribut bebas diantaranya *Age*, BMI(Body Mass Index), *Glucose*, Insulin, HOMA(Homeostatis Model Assesment), Leptin, Adiponectin, Resitin, dan MCP-1(Chemokine Monocyte). Sedangkan pada label attribut terdiri dari 1=*Healthy controls* dan 2=*Patients.*

Data-data ini didapat dari relawan yang didiagnosa kanker payudara yang di rekrut dari departemen Gynaecology Universitas Pusat Rumah Sakit di Coimbra Portugal antara Tahun 2003 dan 2013. Untuk setiap pasien, diagnosis berasal dari hasil mamografi dan secara histologis yang telah dikonfirmasi. Semua sampel bersifat naïve yaitu dikumpulkan sebelum melakukan operasi dan perawatan. Relawan yang sehat tanpa kanker payudara digunakan juga pada penelitian ini. Semua pasien yang sehat tidak pernah mengalami pengobatan kanker sebelumnya, dan semua peserta terbebas dari infeksi atau penyakit akut atau komorbiditas pada saat pendaftaran pada penelitian (Patrício et al., 2018).

Sampel yang ada di dataset kanker payudara Coimbra ini terdiri dari 64 wanita yang mengidap kanker payudara, dan 52 wanita sehat, jadi total records yang ada pada dataset ini berjumlah 116 samples. Dataset Brest Cancer Coimbra dapat dilihat pada Tabel 1.

Tabel 1. Contoh Dataset *Breast Cancer* Coimbra
Sumber: UCI Machine Learning Repository

| *Age* | BMI | *Glucose* | Insulin | HOMA | Leptin | Adiponectin | Resistin | MCP-1 | *Classification* |
|---|---|---|---|---|---|---|---|---|---|
| 48 | 23,5 | 70 | 2,707 | 0,467 | 8,80771 | 9,70224 | 7,99 | 417,114 | 1 |
| 83 | 20,69 | 92 | 3,115 | 0,706 | 8,8433 | 5,4229 | 4,064 | 468,78 | 1 |
| 82 | 23,124 | 91 | 4,4988 | 1,0099 | 17,93 | 22,43 | 9,27 | 554,69 | 1 |
| 68 | 21,36 | 77 | 3,2266 | 0,6127225 | 9,88827 | 7,169956 | 12,766 | 928,22 | 1 |
| 86 | 21,111 | 92 | 3,54 | 0,805 | 6,69 | 4,81 | 10,57 | 773,9 | 1 |
| 49 | 22,85 | 92 | 3,22 | 0,7332 | 6,83 | 13,67 | 10,31 | 530,41 | 1 |
| 89 | 22,7 | 77 | 4,69 | 0,89 | 6,96 | 5,58 | 12,93 | 1256,08 | 1 |

(Sumber: UCI Machine Learning Repository 2019)

Pada tahap pertama pada penelitian ini, dataset kanker payudara Coimbra dirubah record labelnya menjadi nominal. 1 diganti menjadi Healthy, dan 2 diganti menjadi Patient. Contoh perubahan attribute pada Tabel 2.

Tabel 2. Contoh Perubahan Dataset *Breast Cancer* Coimbra
Sumber: Hasil *Preprocessing Data*

| *Age* | BMI | *Glucose* | Insulin | HOMA | Leptin | Adiponectin | Resistin | MCP-1 | *classification* |
|---|---|---|---|---|---|---|---|---|---|
| 48 | 23,5 | 70 | 2,707 | 0,467 | 8,80771 | 9,70224 | 7,99 | 417,114 | *Healthy* |
| 83 | 20,69 | 92 | 3,115 | 0,706 | 8,8433 | 5,4229 | 4,064 | 468,78 | *Healthy* |
| 82 | 23,124 | 91 | 4,4988 | 1,0099 | 17,93 | 22,43 | 9,27 | 554,69 | *Healthy* |
| 68 | 21, | 77 | 3,226 | 0,61 | 9,88827 | 7,16956 | 12,766 | 928,22 | *Hea* |

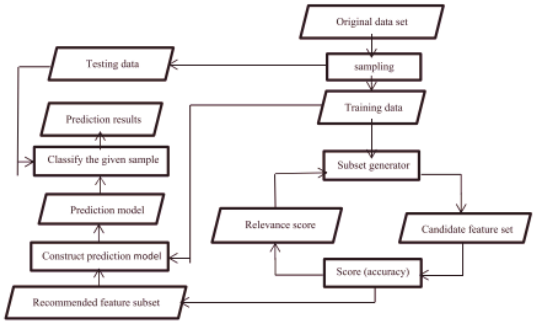
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 36 | | | 2725 | | | | | *lthy* |
| 86 | 21,1111 | 92 | 3,54 | 0,805 | 6,69 | 4,81 | 10,57 | 773,9 | *Healthy* |
| 49 | 22,85 | 92 | 3,22 | 0,732 | 6,83 | 13,67 | 10,31 | 530,41 | *Healthy* |
| 89 | 22,7 | 77 | 4,69 | 0,89 | 6,96 | 5,58 | 12,93 | 1256,08 | *Healthy* |

Selanjutnya pada penelitian ini akan menggunakan metode seleksi fitur tertanam atau *embedded method*. Pada metode tertanam, proses pencarian fitur tertanam ke dalam algoritma klasifikasi (Zhu & Song, 2013), dan proses pembelajaran dan proses pemilihan fitur tidak dapat dipisahkan. Mirip dengan metode pembungkus (*wrapper method*), metode tertanam mencakup interaksi dengan classifier pada kasus ini diterapkan SVM, sementara pada saat yang sama, metode tertanam dapat menghemat biaya komputasi (*computational cost*) lebih besar daripada metode pembungkus.

## 2. Seleksi Fitur

Seleksi fitur menurut Oreski dalam (Fallahpour, Lakvan, & Zadeh, 2017) adalah teknik yang biasa digunakan untuk memecahkan masalah dimana dataset memerlukan banyak fitur. Proses ini mengurangi banyak dimensi vector fitur untuk mengurangi fitur yang tidak perlu dan memilih fitur yang diperlukan untuk pembelajaran model, sehingga meningkatkan akurasi prediksi dan memperbaiki penjelasan dari model prediksi. Ketika vektor fitur berisi fitur tambahan yang tidak perlu, akurasi yang diperoleh dari model pengklasifikasi akan lebih rendah daripada ketika dataset hanya berisi fitur-fitur yang diperlukan untuk pembelajaran model. Teknik pemilihan fitur yang tepat dapat digunakan untuk mengatasi masalah menentukan fitur yang dibutuhkan untuk mencapai akurasi yang optimal.

Metode seleksi fitur yang dipilih adalah metode *forward selection* dimana metode *forward selection* merupakan salah satu metode dari kategori metode pembungkus (*wrap method*) dalam seleksi fitur dimana dalam seleksi fitur terdapat tiga kategori yaitu penyaring/*filter*, pembungkus/*wrapper*, dan tertanam/*embedded*(Zhu & Song, 2013). Kategori penyaring/*filter* menyeleksi fitur dengan cara memilih atribut yang berada di peringkat teratas dalam memenuhi kriteria tertentu (kriteria klasifikasi)(Kotu & Deshpande, n.d.). Metode-metode yang termasuk dalam kategori penyaring diantaranya adalah *Principal Component Analysis* (PCA), Information *gain-based filtering*, dan chi squared *based filtering*. Pemilihan atribut pada kategori *filter* dikerjakan sebelum dilakukan pemodelan. sedangkan kategori pembungkus/*wrapper* bekerja dengan cara memilih secara *iterative* melalui umpan balik perulangan, dan attribute yang dipilih adalah hanyalah attribut yang meningkatkan kinerja/*performace* suatu algoritma(Kotu & Deshpande, n.d.). Metode-metode yang termasuk dalam kategori pembungkus/*wrapper* adalah *forward selection* dan *backward elimination*. *Forward Selection* menurut Kamber, M & Han,j dalam (Astuti, 2018) adalah salah satu prosedur bertahap yang bertujuan untuk menambah variable yang dikendalikan satu per satu ke dalam persamaan yang didasarkan pada Alpa tertentu sebagai masukan. Alpha masukan merupakan nilai yang menentukan apakah salah satu prediktor yang saat ini tidak berada didalam model harus ditambahkan kedalam model. Nilai P dari masing-masing prediktor kurang dari tingkat, sehingga predictor merupakan kandidat untuk dimasukkan kedalam model.




Sumber: (Fallahpour et al., 2017)

Gambar 1. Langkah Klasifikasi dengan Metode Pembungkus (Wrapper)

Prosedur ini akan berakhir ketika semua variabel yang masuk ke dalam model memiliki nilai P kurang dari Alpha tertentu sebagai masukan, sehingga *Forward selection* akan menghilangkan atribut-atribut yang tidak relevan. Algoritma *forward selection* didasarkan pada model regresi linear. Menurut Mulyana dalam (Hasan, 2017) prosedur *forward* selection dapat dirumuskan sebagai berikut:

A. Menentukan model awal $\hat{y} = b_0$ *(1)*
B. Memasukkan variable respon dengan setiap variable berprediktor, misalnya $X_1, X_2, \ldots . X_n$ yang terkait dengan $\hat{y}$. Misalkan $X_1$sehingga membentuk model $\hat{y} = b_0 + b_1 X_1$ *(2)*
C. Uji F terhadap peubah pertama yang terpilih. Jika $F_{hitung} < F_{tabel}$ maka peubah terpilih dibuang dan proses dihentikan. Apabila $F_{hitung} > F_{tabel}$ maka peubah terpilih memiliki pengaruh nyata

terhadap peubah terkait y, sehingga layak untuk diperhitungkan di dalam model.

D. Masukan peubah bebas terpilih (yang paling signifikan) ke dalam model. Misalkan $X_2$, sehingga membentuk suatu model
$\hat{y} = b_0 + b_1X_1 + b_2X_2$ *(3)*

E. Uji F, jika $F_{hitung} < F_{tabel}$ maka proses dihentikan dan model terbaik adalah model sebelumnya. Namun jika $F_{hitung} \geq F_{tabel}$, variable peubah bebas layak untuk dimasukan ke dalam model dan kembali ke langkah C. proses akan berakhir jika tidak ada lagi peubah yang tersisa yang bias dimasukkan ke dalam model.

Pada metode tertanam/*Embedded* proses pencarian fitur tertanam kedalam algoritma klasifikasi, dan proses pembelajaran dengan proses pemilihan fitur tidak dapat dipisahkan(Zhu & Song, 2013). Mirip seperti metode pembungkus (*wrapper*), metode tertanam mencakup interaksi dengan algoritma pengklasifikasi, sementara pada saat yang sama, metode tertanam dapat menghemat biaya komputasi/ *computational cost*. metode-metode yang termasuk dalam metode tertanam adalah seluruh metode yang ada pada metode pembungkus yaitu *forward feature selection,* dan *Backward Feature Elimination.*

## 3. Support Vector Machine

Support Vector machine menurut Gorunescu dalam (Harafani & Maulana, 2019) secara konseptual adalah mesin linear yang dilengkapi dengan fitur special dan didasarkan pada metode minimalisasi resiko struktural, serta teori pembelajaran statistik. SVM telah banyak diteliti dalam komunitas data mining dan pembelajaran mesin (*Machine Learning*) selama sepuluh tahun terakhir . Dua sifat khusus dari SVM yaitu (1) mencapai generalisasi yang tinggi dengan memaksimalkan margin, dan (2) mendukung pembelajaran yang efisien dari fungsi nonlinier pada trik kernel sehingga membuat kinerja generalisasinya baik dalam menyelesaikan masalah pengenalan pola.

Untuk masalah klasifikasi SVM mencoba untuk mencari garis pemisah yang optimal yang diekspresikan sebagai kombinasi linier dari subset data pelatihan dengan menyelesaikan masalah keterbatasan linier pemrograman kuadrat (QP) dengan margin maksimum antara dua kelas.

Tujuan utama dari SVM adalah memperkirakan fungsi klasifikasi dengan menggunakan data pelatihan input-output dari dua kelas $(x_1,y_1),\ldots,(x_n,y_n) \in R^m x \{\pm 1\}$ *(4)*

Tujuan dari fungsi klasifikasi adalah untuk membentuk persamaan hyperplane yang membagi data pelatihan dan meninggalkan semua titik dari kelas yang sama pada sisi yang sama sambil memaksimalkan jarak minimum antara *hyperplane* dan masing-masing dari du akelas (w,b). Dimana w mewakili vector bobot yang mewujudkan margin fungsional 1 pada titik positif $x^+$serta titik negatif $x^-$ dan matrik geometric dapat dihitung dengan rumus berikut:

$$y_i(w.x_i + b) \geq 1, \quad i = 1,\ldots,m \qquad (5)$$

*Hyperplane* optimal $w.x + b = 0$ secara geometris setara dengan memaksimalkan margin yaitu jarak antara dua bidang sejajar $w.x + b = 1$ dan $w.x + b = -1$. Jarak panjang Euclidean dari margin adalah $2/||w||^2$, dimana $||w||^2 = \sum_{i=1}^{m} w_i^2$. Margin maksimum juga merupakan 2-norm $||w||^2$ tergantung pada kendalanya. Oleh karena itu permasalahan ini dapat diformulasikan dengan rumus:

$$\underset{w,b}{Min} \frac{||w||^2}{2} \qquad (6)$$

Tergantung pada $y_i(w.x_i + b) \geq 1$

Dikarenakan class jarang sekali dapat dipisahkan secara linier, generalisasi permasalahan bidang yang optimal dibutuhkan. Dengan demikian, satu set variabel n yang mengukur variasi kendala ditambahkan untuk setiap titik. Formulasi mutasi akhir adalah

$$\underset{w,b,\xi}{min} \frac{||w||^2}{2} + \frac{c}{m} \sum_{i=1}^{m} \xi_i \qquad (7)$$

Tergantung pada $y_i(w.x_i + b) + \xi_i \geq 1$

$\xi_i \geq 0 \;\; i = 1,\ldots,m$

Dimana b merupakan bias, parameter w dan b adalah parameter yang perlu ditentukan nilainya agar dapat memberikan fungsi yang terbaik untuk memetakan input ke data output. Unutk kasus pemisahan yang tidak linier, tidak ada *hyperplane* yang dapat digunakan secara sempurnya untuk memisahkan dua bidang. Oleh karena itu variabel slack diperkenalkan. Rumus yang harus dipecahkan menjadi:

$$\underset{w,b}{max} \frac{2}{\|w\|} + C \sum_{i=1}^{m} \xi_i \qquad (8)$$

Tergantung pada $y_i(w.x_i + b) - \xi_i \geq 0$

Dimana C merupakan parameter penalty

Untuk kasus terpisah tidak linier, ide dasarnya adalah untuk memproyeksikan dataset $x_1$ pada ruang fitur berdimensi tinggi cara yang nonlinier menggunakan fungsi kernel. Sampai saat ini banyak fungsi kernel yang disugestikan untuk mengatasi permasalahan ini

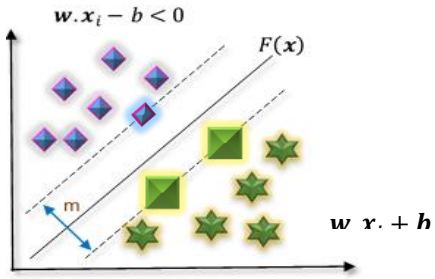
diantaranya, *Radial Basis Function* (RBF) yang paling banyak digunakan dan dilakukan pada banyak kasus. Rumus dari RBF adalah:

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (9)$$

Dimana parameter $\gamma$ menunjukkan lebar dari kernel Gaussian.

## 4. K-Fold Cross Validation

*Cross Validation* adalah estimator yang banyak digunakan untuk mengevaluasi kesalahan prediksi (Bergmeir, Costantini, & Benítez, 2014) terutama untuk model *non-linear regression*(Borra & Di Ciaccio, 2010). K-*fold cross validation* merupakan prosedur untuk menempatkan nilai hyperparameter terbaik untuk SVM(Q. Li, Salman, Test, Strack, & Kecman, 2013). Data yang tersedia dibagi menjadi k-set ukuran yang sama, sebagai contoh prosedur 10-*fold cross validation* membagi data set menjadi 10 bagian sama besar. Selama setiap kelipatan fase training, setiap set kecil digunakan sebagai data testing sekali atur dan data selebihnya digunakan untuk training. Jumlah total sample yang salah diakumulasi unruk menghitung akurasi akhir. Sebagai contoh RBF SVM mempunyai 2 Hyperparameter, nilai pinalti C dan parameter $\gamma$. Jika ada 10 perbedaan nilai C dan 5 perbedaan $\gamma$, maka 10-*fold cross validation* akan menjalankan proses sebanyak 50 kali, yang mana fase *training* menjalankan proses 50 x 10 kali, dan fase *testing* juga menjalankan proses sebanyak 50 x 10 kali. Berdasarkan semua proses tersebut dapat diketahui bahawa komputasi kernel sangat memakan waktu.




Sumber: Penelitian (2019)

Gambar 2. Margin pemisah maksimal pada klasifikasi linier SVM

## 5. Confussion Matrix

Pada penelitian ini keputusan yang diperoleh pada tahapan training dan testing akan dituliskan dalam *confusion matrix*, yang mana confusion matrix dikenal sebagai tabel kemungkinan error menurut Stehman Stephen V dalam (Hasan, 2017). Tabel kemungkinan error confusion matrix dapat dilihat pada Tabel 3. Berdasarkan Tabel kemungkinan error *confussion matrix* dapat diperoleh nilai akurasinya. Sehingga skema alur penelitian dapat digambarkan pada Gambar 3.

Tabel 3. Confussion Matrix

.

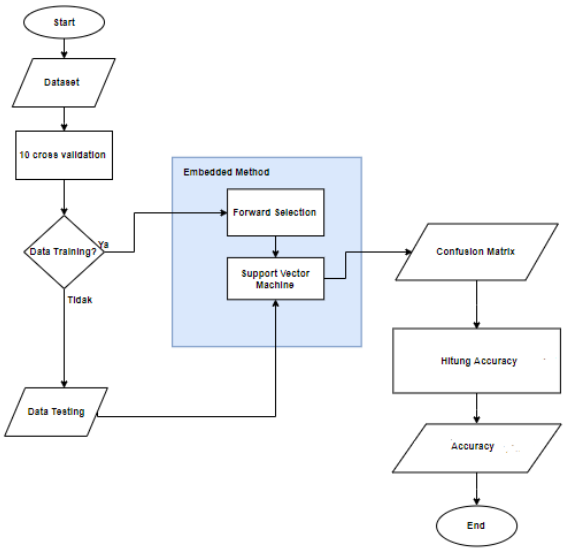
| | | KONDISI SEBENARNYA | |
|---|---|---|---|
| | **JUMLAH POPULASI** | POSITIF | POSITIF |
| HASIL PREDIKSI | **POSITIF** | TRUE POSITIF | FALSE POSITIF |
| | **POSITIF** | FALSE NEGATIF | TRUE NEGATIF |

Sumber: (Hasan, 2017)

Maksud dari *true positif* adalah mesin prediksi dapat memprediksi nilai yang positif sebagai nilai yang positif artinya *true positif* mewakili nilai yang benar, yaitu sesuai dengan kondisi sebenarnya. Begitupula dengan *true negatif*. Sementara *False negatif* maksudnya adalah mesin memprediksi nilai yang positif sebagai nilai yang negatif. Artinya nilai negatif yang dihasilkan oleh mesin prediksi bernilai salah atau bias dibilang nilai negatif yang salah, karena seharusnya nilai sebenarnya adalah positif. Rumus Akurasi bias didapatkan dengan persamaan berikut:

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (10)$$

Pada pengukuran performansi. Variable TP, TN, FP, dan FN merupakan singkatan dari True Positif, True Negatif, False Positif, dan False Negatif.




Sumber: Metode Penlitian (2019)

Gambar 3. Skema Alur Penelitian

## HASIL DAN PEMBAHASAN

Eksperimen dilakukan menggunakan komputer personal Intel Core i3, SSD 120GB, 500GB HDD, 4GB RAM, sistem operasi Windows 10, dan Rapidminer 9.0.

Pada tahap pertama pada penelitian ini, dataset kanker payudara Coimbra dirubah record labelnya menjadi nominal. 1 diganti menjadi Healthy, dan 2 diganti menjadi Patient. Pada tahapan kedua sebelum dataset dilatih (training) dan diuji (testing), dataset akan dipecah terlebih dahulu dengan menerapkan 10-fold cross validation untuk membagi data menjadi dua bagian yaitu 90% training dan 10% testing. Kemudian percobaan dilakukan dengan tiga tahapan yaitu yang pertama memasukkan training dataset ke dalam algoritma klasifikasi Support Vector Machine (SVM) dengan berbagai kernel (linear, polynomial, RBF) dan mengukur performa percobaan.

Percobaan kedua dilakukan dengan menerapkan teknik seleksi fitur dengan metode tertanam atau yang sering disebut *embedded method* pada SVM dengan ketiga kernelnya.

Tahapan ketiga pada penelitian ini adalah peneliti mencoba untuk memasukkan nilai parameter pada SVM dengan 5 kali percobaan, kemudian membandingkan akurasi SVM dengan SVM+FS.

Tahapan keempat membandingkan hasil akurasi antara machine learning yang lainnya (Decision Tree, Neural Network, K-NN, dan Random Forest), dibandingkan dengan hasil akurasi SVM, dan hasil akurasi SVM dengan metode *forward selection.*

Setelah dataset kanker payudara Coimbra dirubah labelnya dan dipisah menjadi bagian training dan testing, selanjutnya Support Vector Machine akan melakukan klasifikasi terhadapt dataset, sehingga dapat diperoleh akurasi performansinya seperti yang dapat dilihat pada Tabel 4.

Tabel 4. Akurasi SVM

| Algoritma | TP | TN | FP | FN | Accuracy |
|---|---|---|---|---|---|
| SVM (Dot) | 39 | 44 | 13 | 20 | 72,12% |
| SVM (Polynomial) | 11 | 63 | 41 | 1 | 63,71% |
| SVM (RBF) | 25 | 57 | 27 | 7 | 70,98% |

Sumber: Hasil Penelitian (2019)

Pada Tabel 4. dapat dilihat bahwa akurasi SVM dengan kernel Polynomial lebih rendah dari pada kedua kernel lainnya Pada percobaan ini, SVM memiliki nilai akurasi yang cukup tinggi, ketidakmaksimalan prediksi ini disebabkan karena SVM memiliki kelemahan dalam menentukan nilai parameter yang optimal(Zhu & Song, 2013)(Asdaghi & Soleimani, 2019).

Tabel 5. Akurasi SVM dengan Forward Selection

| Algoritma | TP | TN | FP | FN | Accuracy |
|---|---|---|---|---|---|
| SVM (Dot) | 31 | 51 | 13 | 21 | 79,32% |
| SVM (Polynomial) | 41 | 39 | 25 | 11 | 69,77% |
| SVM (RBF) | 35 | 56 | 8 | 17 | 82,42% |

Sumber: Hasil Penelitian (2019)

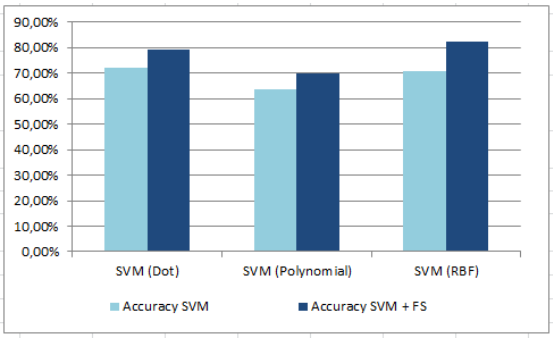
Pada Table 5. dapat dilihat bahwa akurasi SVM meningkat seiring dengan dapat diprediksinya nilai *False Positif dan False Negatif* pada masing-masing kernel SVM. Sehingga jika dibandingkan maka terdapat perbedaan akurasi yang cukup jauh antara sebelum melakukan seleksi fitur dan setelah melakukan seleksi fitur yang dapat dilihat pada Table 5, dan Gambar 2.

Tabel 6. Perbandingan akurasi antara SVM dan SVM+FS

| Algoritma | Accuracy SVM | Accuracy SVM + FS |
|---|---|---|
| SVM (Dot) | 72,12% | 79,32% |
| SVM (Polynomial) | 63,71% | 69,77% |
| SVM (RBF) | 70,98% | 82,42% |

Sumber: Hasil Penelitian (2019)

Berdasarkan Tabel 6 akurasi SVM dengan kernel Linear mencapai selisih 15,99% akurasi, sementara akurasi SVM dengan kernel polynomial mencapai selisih 14,47% akurasi, sedangkan SVM dengan kernel RBF memiliki selisih paling besar yaitu 24,32%. Grafik perbandingan akurasi SVM dapat dilihat pada Gambar 4.



Sumber: Hasil Penelitian (2019)

Gambar 4. Perbandingan Akurasi SVM dengan SVM+FS

Selanjutnya parameter pada SVM akan diatur dengan percobaan sebanyak 5kali dan akurasi SVM akan dibandingkan dengan akurasi SVM+FS seperti yang dapat dilihat pada Tabel 7.

Tabel 7. Perbandingan akurasi SVM kernel dot dengan SVM+FS setelah parameter diatur

| Parameter | | Akurasi | |
|---|---|---|---|
| C | ε | **SVM** | **SVM+FS** |
| 0,1 | 0,01 | 73,03% | 72,50% |
| 0,5 | 0,01 | 72,95% | **79,39%** |
| 0,3 | 0,001 | 72,20% | 79,32% |
| 0,6 | 0,01 | 73,86% | 74,02% |
| 0,7 | 0,01 | **76,14%** | 73,18% |

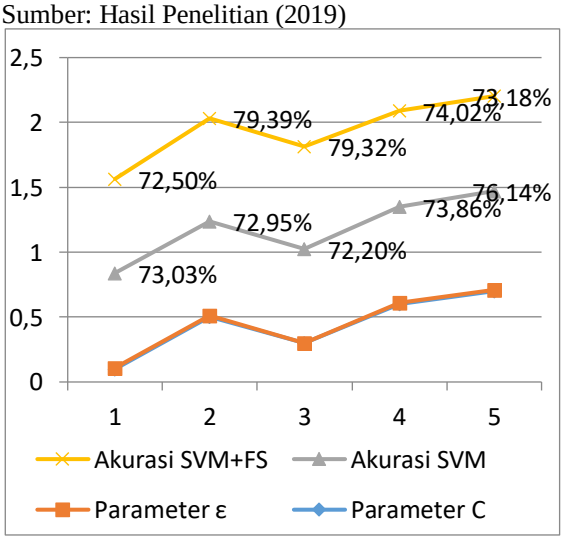
Sumber: Hasil Penelitian (2019)

Nilai Akurasi terbesar SVM pada kernel dot diperoleh dari nilai pengaturan parameter C= 0,7, dan ε= 0,01yaitu sebesar 76,14%, namun nilai akurasi yang palib besar dari seluruh percobaan dengan kernel dot diperoleh dari percobaan SVM(dot)+FS dengan pengaturan nilai parameter C= 0,5, dan ε= 0,01 yaitu sebesar 79,39% namun selisihnya tidak jauh dengan akurasi SVM+FS sebelum parameter diatur yaitu sebesar 0,07%. Grafik perbedaan Akurasi SVM berdasarkan parameter yang telah diatur dapat dilihat pada Gambar 5.

Tabel 8. Perbandingan akurasi SVM kernel polynomial dengan SVM+FS setelah parameter diatur

| Parameter | | Akurasi | |
|---|---|---|---|
| C | ε | **SVM** | **SVM+FS** |
| -0,01 | 0,1 | **65,38%** | 69,02% |
| -0,05 | 0,01 | **65,38%** | 68,86% |
| -0,07 | 0,01 | 61,97% | 68,86% |
| -0,02 | 0,3 | 64,70% | **71,36%** |
| -0,07 | 0,7 | 63,86% | 66,67% |
| | | | |

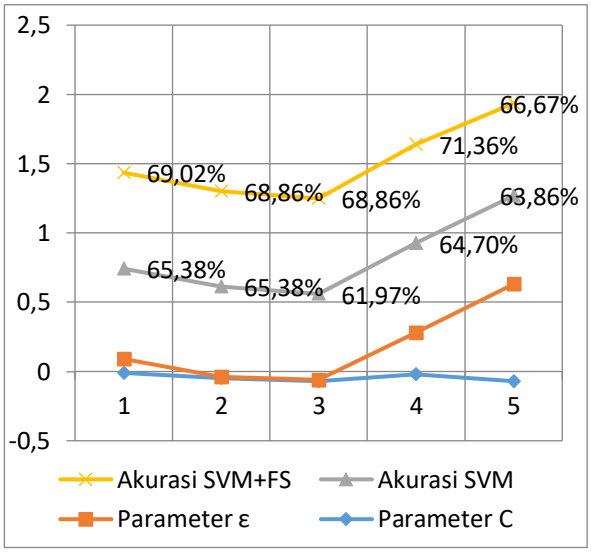
Sumber: Hasil Penelitian (2019)




Sumber: Hasil Penelitian (2019)

Gambar 5. Perbandingan Akurasi SVM dengan SVM+FS Kernel Dot setelah parameter diatur

Nilai Akurasi terbesar SVM pada kernel polynomial diperoleh dari percobaan pengaturan nilai parameter C= -0,01, dan ε= 0,1 yaitu sebesar 65,38%, namun nilai akurasi terbesar dari seluruh percobaan dengan kernel polynomial diperoleh dari percobaan SVM(polynomial)+FS dengan pengaturan nilai parameter C= -0,02, dan ε= 0,3 yaitu sebesar 71,36% dan selisihnya tidak jauh dengan akurasi SVM+FS sebelum parameter diatur yaitu sebesar 1,59% . Grafik perbedaan Akurasi SVM berdasarkan parameter yang telah diatur dapat dilihat pada Gambar 6.




Sumber: Hasil Penelitian (2019)

Gambar 6. Perbandingan Akurasi SVM dengan SVM+FS Kernel Polynomial setelah parameter diatur

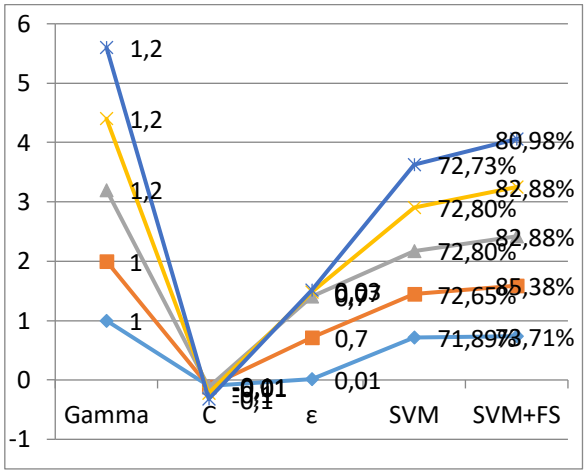
Tabel 9. Perbandingan akurasi SVM kernel RBF dengan SVM+FS setelah parameter diatur

| **Parameter** | | | **Akurasi** | |
|---|---|---|---|---|
| $\gamma$ | **C** | **ε** | **SVM** | **SVM+FS** |
| 1 | -0,1 | 0,01 | 71,89% | 73,71% |
| 1 | -0,01 | 0,7 | 72,65% | **85,38%** |
| 1,2 | -0,01 | 0,7 | **72,80%** | 82,88% |
| 1,2 | -0,1 | 0,07 | 72,80% | 82,88% |
| 1,2 | -0,1 | 0,03 | 72,73% | 80,98% |

Sumber: Hasil Penelitian (2019)

Nilai Akurasi terbesar SVM pada kernel RBF diperoleh dari percobaan pengaturan nilai parameter $\gamma = 1,2$, C= -0,01, dan ε= 0,7 yaitu sebesar 72,80%, namun nilai kaurasi terbesar dari seluruh percobaan dengan kernel RBF diperoleh dari percobaan SVM(RBF)+FS dengan pengaturan nilai parameter $\gamma = 1$, C= -0,01, dan ε= 0,7 dan perbedaannya cukup jauh dengan akurasi SVM+FS sebelum parameter

diatur yaitu sebesar 2,96% . Grafik perbedaan Akurasi SVM berdasarkan parameter yang telah diatur dapat dilihat pada Gambar 7.




Sumber: Hasil Penelitian (2019)

Gambar 7. Perbandingan Akurasi SVM dengan SVM+FS Kernel RBF setelah parameter diatur

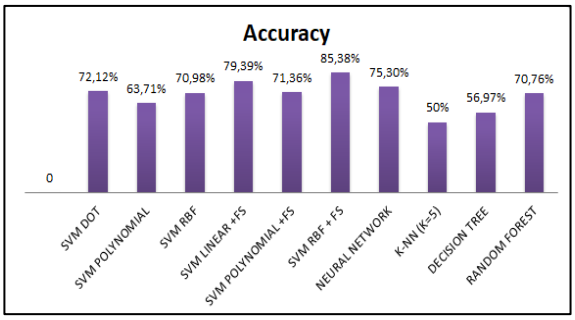
Selanjutnya, akurasi SVM, dan SVM+FS akan dibandingkan dengan akurasi algoritma *machine learning* lainnya (Decision Tree, Neural Network, K-NN, dan Random Forest) yang dapat dilihat pada Tabel 8.

Tabel 10. Perbandingan akurasi antara SVM dan SVM+FS

| Algoritma | Accuracy |
|---|---|
| SVM DOT | 72,12% |
| SVM POLYNOMIAL | 63,71% |
| SVM RBF | 70,98% |
| SVM (DOT)+FS | 79,39% |
| SVM (POLYNOMIAL) +FS | 71,36% |
| SVM (RBF) + FS | **85,38%** |
| NEURAL NETWORK | 75,30% |
| K-NN (K=5) | 50% |
| DECISION TREE | 56,97% |
| RANDOM FOREST | 70,76% |

Sumber: Hasil Penelitian (2019)

Berdasarkan Tabel 10 algoritma Neural Network memiliki nilai akurasi yang baik walaupun tanpa seleksi fitur, namun akurasi SVM(dot)+FS masih lebih unggul dibandingkan NN, dan akurasi SVM(RBF) + FS paling unggul dibandingkan semua algoritma. Grafik perbandingan akurasi antar algoritma dapat dilihat pada Gambar 8.




Sumber: Hasil Penelitian (2019)

Gambar 8. Perbandingan Akurasi SVM, SVM+FS, dengan Algoritma *Machine Learning* Lainnya

## KESIMPULAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa SVM dengan kernel Polynomial mempunyai akurasi yang paling buruk diantara ketiga kernel. Pada awalnya tanpa seleksi fitur Kernel Dot unggul dalam akurasi dibandingkan ketiga kernel lainnya, kemudian setelah diterapkan *forward selection*, Akurasi SVM dengan kernel RBF meningkat melebihi kernel lainnya, Begitupula setelah parameter-parameter SVM diatur, Kernel RBF menunjukkan selisih prosentase akurasi yang paling jauh diantara kedua kernel lainnya. Sehingga dapat dibuktikan bahwa Seleksi fitur benar-benar dapat meingkatkan performansi machine learning terutama pada Support Vector Machine dengan kernel RBF. Algoritma SVM awalnya memiliki performansi cukup baik, ketidakmaksimalan SVM dalam memprediksi disebabkan oleh lemahnya SVM dalam menentukan nilai parameter yang optimal. Namun, SVM tetap lebih unggul jika dibandingkan dengan algoritma machine learning lainnya setelah dilakukan pengaturan parameter dan menerapkan *forward selection*. Isu penelitian dimasa mendatang, peneliti percaya bahwa algoritma metaheuristic dapat menanggulangi permasalahan pada SVM, sehingga dikemudian hari kombinasi SVM dan Forward Selection jika digabungkan dengan algoritma Metaheuristic dapat lebih meningkatkan performansi SVM dalam memprediksi Kanker Payudara Coimbra.

## REFERENSI

Astuti, F. D. (2018). Seleksi Fitur Forward Selection pada Algoritma Naive Bayes untuk Klasifikasi Benih Gandum. *Jurnal Informasi Interaktif*, *3*(1), 161–166.

Balram, D., Lian, K. Y., & Sebastian, N. (2019). Air quality warning system based on a localized PM2.5 soft sensor using a novel approach of Bayesian regularized neural network via forward feature selection. *Ecotoxicology and Environmental Safety*, *182*(April), 109386. https://doi.org/10.1016/j.ecoenv.2019.109386

Bergman, J., Schrempf, D., Kosiol, C., & Vogl, C. (2018). Inference in population genetics using forward and backward, discrete and continuous time processes. *Journal of Theoretical Biology*, *439*, 166–180. https://doi.org/10.1016/j.jtbi.2017.12.008

Bustamam, A., Bachtiar, A., & Sarwinda, D. (2019). Selecting Features Subsets Based on Support Vector Machine-Recursive Features Elimination and One Dimensional-Naïve Bayes Classifier using Support Vector Machines for Classification of Prostate and Breast Cancer. *Procedia Computer Science*, *157*, 450–458. https://doi.org/10.1016/j.procs.2019.08.238

Ellmann, S., Seyler, L., Evers, J., Heinen, H., Bozec, A., Prante, O., … Bäuerle, T. (2019). Prediction of early metastatic disease in experimental breast cancer bone metastasis by combining PET/CT and MRI parameters to a Model-Averaged Neural Network. *Bone*, *120*, 254–261. https://doi.org/10.1016/j.bone.2018.11.008

Fallahpour, S., Lakvan, E. N., & Zadeh, M. H. (2017). Using an ensemble classifier based on sequential floating forward selection for financial distress prediction problem. *Journal of Retailing and Consumer Services*, *34*(March 2016), 159–167. https://doi.org/10.1016/j.jretconser.2016.10.002

Harafani, H., & Wahono, R. S. (2015). Optimasi Parameter pada Support Vector Machine Berbasis Algoritma Genetika untuk Estimasi Kebakaran Hutan. *Journal of Intelligent Systems*, *1*(2).

Ilhan, I., & Tezel, G. (2013). A genetic algorithm-support vector machine method with parameter optimization for selecting the tag SNPs. *Journal of Biomedical Informatics*, *46*(2), 328–340. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23262450

Jafari-Marandi, R., Davarzani, S., Soltanpour Gharibdousti, M., & Smith, B. K. (2018). An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals. *Applied Soft Computing Journal*, *72*, 108–120. https://doi.org/10.1016/j.asoc.2018.07.060

Kotu, V., & Deshpande, B. (n.d.). *Predictive Analytics and Data Mining*.

Levine, A. B., Schlosser, C., Grewal, J., Coope, R., Jones, S. J. M., & Yip, S. (2019). Rise of the Machines: Advances in Deep Learning for Cancer Diagnosis. *Trends in Cancer*, *5*(3), 157–169. https://doi.org/10.1016/j.trecan.2019.02.002

Li, L., Yu, S., Xiao, W., Li, Y., Li, M., Huang, L., … Yang, H. (2014). Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach. *Biochimie*, *104*(1), 100–107. https://doi.org/10.1016/j.biochi.2014.06.001

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling and Software*, *101*, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). A knowledge-based system for breast cancer classification using fuzzy logic method. *Telematics and Informatics*, *34*(4), 133–144. https://doi.org/10.1016/j.tele.2017.01.007

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, *18*(1), 1–8. https://doi.org/10.1186/s12885-017-3877-1

Shukla, N., Hagenbuchner, M., Win, K. T., & Yang, J. (2018). Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, *155*, 199–208. https://doi.org/10.1016/j.cmpb.2017.12.011

Singh, B. K. (2019). Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm. *Biocybernetics and Biomedical Engineering*, *39*(2), 393–409. https://doi.org/10.1016/j.bbe.2019.03.001

Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2018). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*. https://doi.org/10.1016/j.cegh.2018.10.003

Zhu, M., & Song, J. (2013). An embedded backward feature selection method for MCLP classification algorithm. *Procedia Computer Science*, *17*, 1047–1054. https://doi.org/10.1016/j.procs.2013.05.133