Deteksi Dini Penyakit Kanker Leher Rahim (Serviks) di Kota Bogor Mengunakan Regresi Logistik Biner dan *Support Vector Machine* (SVM)

Agil Darmawan dan Santi Wulan Purnami
Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111
Email : santiwulan08@gmail.com

***Abstrak*— Negara-negara berkembang menyumbang 370.000 dari total 466.000 kasus serviks kanker yang diperkirakan terjadi di dunia dalam tahun 2000. Sebagian besar kasus kanker serviks disebabkan oleh infeksi *Human Papilloma Virus* (HPV). Kanker serviks tidak akan terdiagnosa secara langsung karena ada fase pra-ganas selama beberapa tahun, maka dibutuhkan deteksi dini untuk mencegah munculnya fase ganas pada kanker serviks. Untuk melakukan deteksi dini tersebut digunakan metode klasifikasi *Support Vector Machine* (SVM) yang akan dibandingkan dengan Regresi Logistik Biner. Selain untuk melihat ketepatan klasifikasi Regresi Logistik Biner juga digunakan untuk mengetahui variabel prediktor yang paling berpengaruh terhadap respon. Dalam penelitian ini, data diambil dari Studi Kohort Faktor Risiko Penyakit Tidak Menular di Kota Bogor. Variabel prediktor yang digunakan adalah sebanyak 13 variabel. Faktor resiko yang berpengaruh signifikan pada taraf signifikasi 90% (α=0,1) terhadap Kanker Serviks pada Analisis Regresi Logistik Biner adalah Lama penggunaan kontrasepsi, Riwayat Keluarga dan Tes *Pap Smear*. Performansi klasifikasi menggunakan SVM pada semua kombinasi baik 90:10, 70:30, dam 50:50 adalah sebesar 100%, sedangkan nilai specificity semua 0%. Akurasi klasifikasi menggunakan Logistik Biner tertinggi adalah kombinasi 90:10 sebesar 100%, kombinasi 70:30 sebesar 87,7%, sedangkan kombinasi 50:50 sebesar 55,5%. Nilai specificity Logistik Biner semua 100%, jadi responden yang terjangkit semua bisa diprediksi terjangkit. Nilai *sensitifity* SVM sebesar 100%. Hal ini menunjukkan bahwa prediksi menuju kepada prediksi kategori "tak terjangkit". Terjadi demikian karena proporsi kategori yang tidak seimbang antara "terjangkit" dengan "tak terjangkit".**

Kata kunci—kanker serviks, klasifikasi, Regresi Logistik, SVM

I. PENDAHULUAN

Kanker serviks merupakan suatu masalah kesehatan masyarakat di negara-negara berkembang di Asia Tenggara, Amerika Tengah dan Selatan, Afrika. Sebagian besar kasus kanker serviks disebabkan oleh infeksi Human Papilloma Virus (HPV), virus menular yang menginfeksi sel dan dapat menyebabkan lesi prakanker dan kanker invasif. Di seluruh dunia, kanker serviks diklaim menjangkit 231.000 wanita per tahun, lebih dari 80% yang terjadi di negara berkembang [1].

Kanker serviks memiliki tahap pra-ganas dimana ia tumbuh, namun tidak akan menjalar. Tahap pra-ganas berlangsung beberapa tahun. Oleh karena itu untuk mendeteksi dini adanya kanker serviks dianjurkan untuk melakukan pemeriksaan *Pap Smear* [2].

Data-data ini diperkuat dengan penelitian Yayasan Kanker Indonesia yang memperkirakan, ada sekitar 52 juta perempuan Indonesia memiliki risiko terkena kanker serviks. Semua data tersebut seolah mempertegas asumsi bahwa setiap perempuan berisiko terkena infeksi Human Papilloma Virus (HPV), virus penyebab kanker serviks [3].

Faktor resiko adalah faktor yang menjadikan resiko terkena penyakit kanker menjadi besar. Penggunaan kontrasepsi jenis hormonal seperti pil dan suntik juga meningkatkan resiko terserang kanker serviks, terutama untuk penggunaan yang lama . Faktor lain, perempuan yang sering melahirkan anak (paritas) dan ganti-ganti pasangan seksual meningkatkan resiko kanker ini [4]. Adanya riwayat kanker pada keluarga juga meningkatkan resiko terjangkit kanker serviks [5]. Faktor resiko lain adalah merokok, karena rokok dapat mengganggu sistem imun tubuh dalam melawan virus [6].

Beragamnya faktor resiko penyakit kanker serviks, maka untuk mengetahui faktor resiko yang berpengaruh signifikan terhadap kanker serviks tersebut peneliti menggunakan Regresi Logistik Biner. Regresi logistik biner bermanfaat untuk penelitian dengan variabel respon biner (dua outcome), seperti ya-tidak, benar-salah, normal-abnormal, dan lain-lain. Pada penelitian tentang kanker serviks oleh Intansari (2012) [7] menggunakan *Bagging Logistik* menunjukkan faktor yang paling berpengaruh signifikan terhadap kanker serviks adalah usia, jumlah anak, usia pertama melahirkan, dan penggunaan kontrasepsi. Pada penelitian tersebut didapatkan akurasi ketepatan klasifikasi menggunakan *Bagging Logistic* sebesar 70,74%. Nilai tersebut masih tergolong rendah. Dibutuhkan metode klasifikasi lain yang memiliki ketepatan klasifikasi tinggi, yaitu *Support Vector Machine* (SVM).

SVM mampu menemukan pemisah (*hyperplane*) terbaik yang memisahkan dua buah *class* pada *input space*. Keunggulan SVM adalah memiliki tingkat akurasi klasifikasi yang tinggi dibanding metode lain seperti *Logistic Regression*, *Neural Netwotk* (NN) dan *Discriminant Analysis* [8]. Pada penelitian Rahman (2012) [9] tentang Kanker Payudara menggunakan Regresi Logistik Ordinal dan SVM, hasil pengukuran klasifikasi kedua metode, akurasi SVM

sebesar 98,11 % jauh lebih tinggi dari pada Regresi Logistik Ordinal.

II. TINJAUAN PUSTAKA

A. Regresi Logistik Biner

Analisis Regresi adalah suatu metode yang mendiskripsi-kan antara variabel respon dan satu atau lebih variabel penjelas atau prediktor [11]. Regresi Logistik Biner adalah metode regresi yang mampu menyelesaikan kasus di mana variabel respon berupa *dichotomous*, ya-tidak, sukses-gagal, normal-cacat, benar-salah, laki-laki-perempuan, dan sebagainya. Variabel respon adalah data kategorik [12].

Outcome variabel *y* yang terdiri dari 2 kategori, yaitu "sukses" dan "gagal" dinotasikan dengan $y = 1$ (sukses) dan $y = 0$ (gagal). Variabel *y* tersebut mengikuti distribusi *Bernaulli* untuk setiap observasi tunggal. Fungsi probabilitas untuk setiap observasi adalah :

$$f(y) = \pi^{y}(1-\pi)^{1-y} \quad y = 0, 1 \tag{1}$$

Di mana jika $y = 0$ maka $f(y) = 1 - \pi$ dan jika $y = 1$ maka $f(y) = \pi$. Fungsi regresi logistiknya dapat ditulis sebagai berikut :

$$f(z) = \frac{1}{1+e^{-z}} \text{ ekuivalen } f(z) = \frac{e^{z}}{1+e^{z}} \tag{2}$$

dengan $z = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$.

Nilai *z* antara $-\infty$ sampai $+\infty$ sehingga nilai *f(z)* terletak antara 0 dan 1 untuk setiap nilai *z* yang diberikan. Hal tersebut menunjukkan bahwa model Logistik sebetulnya menggambarkan probabilitas atau resiko dari suatu objek. Model regresi logistik-nya adalah sebagai berikut :

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}, \tag{3}$$

di mana *p* = banyaknya variabel prediktor.

Untuk mempermudah pendugaan parameter regresi maka persamaan (3) di atas dapat diuraikan menggunakan transformasi logit dari π(x) sebagai berikut :

$$g(x) = \ln\left(\frac{\pi(\vec{x})}{1-\pi(\vec{x})}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \tag{4}$$

model tersebut merupakan fungsi linear dari parameter – parameternya.

Estimasi parameter pada regresi Logistik menggu-nakan *Maximum Likelihood*. Metode ini menduga parameter β dengan memaksimumkan fungsi *Likelihood* dan men-syaratkan data harus mengikuti distribusi tertentu. Pada regresi Logistik biner, setiap percobaan mengikuti distribusi Bernaulli sehingga dapat ditentukan fungsi *Likelihood*-nya.

Jika x_i dan y_i adalah pasangan variabel respon dan prediktor pada pengamatan ke-*i* dan diasumsikan bahwa setiap pengamatan saling independen dengan pasangan pengamatan lainnya, $i = 1, 2, \ldots, n$ maka fungsi probabilitas untuk setiap pasangan adalah sebagai berikut :

$$f(x_i) = \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad y_i = 0, 1 \tag{5}$$

dengan, $\pi(x_i) = \dfrac{e^{\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\sum_{j=1}^{p}\beta_j x_j}}$ (6)

Ketika $j = 0$ maka $x_{ij} = x_{i0} = 1$ Fungsi *Likelihood*-nya :

$$l(\beta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \tag{7}$$

Fungsi *Likelihood* tersebut lebih mudah dimak-simumkan dalam bentuk log l(β) dan dinyatakan dalam L($\vec{\beta}$).

$$\mathrm{L}(\vec{\beta}) = \log l(\vec{\beta}) = \sum_{j=0}^{p}\left(\sum_{i=1}^{n} y_i x_{ij}\right)\beta_j - \sum_{i=1}^{n}\log\left(1+e^{\sum_{j=1}^{p}\beta_j x_j}\right) \tag{8}$$

Nilai β didapatkan melalui turunan L($\vec{\beta}$) terhadap β dan hasilnya disamadengankan 0.

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij}\left(\frac{e^{\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\sum_{j=1}^{p}\beta_j x_j}}\right) \tag{9}$$

Sehingga, $\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij}\hat{\pi}_i(x_i) = 0 \quad j=0,1, \ldots,p$ (10)

Untuk mencari turunan dari persamaan (10) seringkali tidak mendapatkan hasil yang eksplisit sehingga digunakan metode iterasi *Newton Raphson* untuk mengatasinya.

Berikutnya adalah melakukan pengujian secara serentak untuk mengetahui keberartian koefisien β secara serentak terhadap respon.

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1 : minimal ada satu $\beta_j \neq 0 \qquad j = 1, 2, \ldots, p$

Statistik uji : $G = -2\ln \dfrac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}}$ (13)

di mana, $n_1 = \sum_{i=1}^{n} y_i \qquad n_0 = \sum_{i=1}^{n}(1-y_i)$. (14)

Statistik uji *G* merupakan *Likelihood Rasio Test* yang mengikuti distribusi *Chi Square* sehingga tolak H_0 jika $G > \chi^2_{(v,\alpha)}$ dengan *v* derajat bebas banyaknya parameter dalam model tanpa β_0.

Kemudian dilakukan pengujian keberartian terhadap koefisien β secara univariat.

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0 \qquad i = 1, 2, \ldots, p$

Statistik uji : $W = \dfrac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$ (11)

Rumus di atas biasa disebut Uji Wald, yang mengikuti distribusi normal sehingga tolak H_0 jika $|W| > Z_{\alpha/2}$

Uji berikutnya adalah Uji Kesesuaian Model. Ini dimaksudkan untuk mendapatkan informasi apakah terdapat perbedaan antara hasil pengamatan dengan kemungkinan hasil prediksi model.

H_0 : Model sesuai
H_1 : Model tidak sesuai

Statistik uji : $\chi^2 = \sum_{k=1}^{g} \frac{(o_k - n_k' \bar{\pi}_k)^2}{n_k' \bar{\pi}_k (1-\bar{\pi}_k)}$, (15)

di mana;
o_k : observasi pada grup ke – k
$\bar{\pi}_k$: rata-rata taksiran peluang
g : jumlah grup
n_k' : banyak observasi grup ke – k,

daerah penolakannya adalah, tolak H_0 jika $\chi^2 < \chi^2_{(p-1,\alpha)}$

Performansi dalam melakukan klasifikasi kanker serviks diuji ketepatannya menggunakan data testing. Pengukuran ketepatan klasifikasi menggunakan sensitivitas, spesivisitas, dan akurasi berdasarkan model yang terbentuk.

Tabel 1. Ketepatan klasifikasi

Observasi	Prediksi	
	Gagal	Sukses
Gagal	n_{11}	n_{12}
Sukses	n_{21}	n_{22}

n_{11}: kategori gagal yang diprediksi gagal
n_{12}: kategori gagal yang diprediksi sukses
n_{21}: kategori sukses yang diprediksi gagal
n_{22}: kategori sukses yang diprediksi sukses

Akurasi : $\frac{n_{11}+n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}}$

Specificity : $\frac{n_{22}}{n_{21}+n_{22}}$

Sensitifity : $\frac{n_{11}}{n_{11}+n_{12}}$

B. *Support Vector Machine* (SVM)

SVM adalah metode *learning machine* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan pemisah (*hyperplane*) terbaik yang memisahkan dua buah *class* pada *input space* [8]. Prinsip dasar SVM adalah *linear classifier*, dan selanjutnya dikembangkan agar dapat digunakan untuk kasus *non-linear* dengan memasukkan konsep Kernel. Dengan begitu, ada suatu jaminan bahwa klasifikasi menggunakan SVM akan menghasilkan pemetaan yang sangat akurat [13].

Data yang ada dinotasikan sebagai $\vec{x}_i \in \Re^d$ sedangkan untuk respon/target masing – masing dinotasikan $y_i \in \{-1,+1\}, i = 1,2,...,l$, yang mana l adalah banyaknya data. Diketahui bahwa X memiliki pola tertentu, yaitu apabila $\mathbf{x}_i$ termasuk kedalam *class* maka $\mathbf{x}_i$ diberikan label (target) $y_i = +1$ dan $y_i = -1$. Diasumsikan +1 dan –1 dapat terpisah secara sempurna oleh *hyperplane* berdimen-si d, yang didefinisikan :

$$\vec{w}.\vec{x} + b = 0 \quad (16)$$

Pattern $\vec{x}_i$ yang termasuk *class* –1 (sampel negatif) dapat dirumuskan sebagai *Pattern* yang memenuhi pertidaksamaan sebagai berikut :

$$\vec{w}.\vec{x} + b \le -1 \quad (17)$$

Sedangkan *Pattern* $\vec{x}_i$ yang masuk *class* +1(sampel positif) dapat dirumuskan sebagai *Pattern* yang memenuhi pertidaksamaan :

$$\vec{w}.\vec{x} + b \ge +1 \quad (18)$$

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara *hyperplane* dan titik terdekatnya, yaitu $1/\|\vec{w}\|$. Hal ini dapat dirumuskan sebagai *Quadratic Programming* (QP) *problem*, yaitu mencari titik minimal persamaan (19), dengan memperhatikan *constraint* persamaan (20).

$$\min_{\vec{w}} \tau(\vec{w}) = \frac{1}{2}\|\vec{w}\|^2 \quad (19)$$

$$y_i[\vec{w}.\vec{x} + b] \ge 0, i = 1,...,l \quad (20)$$

Problem ini diselesaikan dengan metode *Lagrange Multiplier*.

$$L(\vec{w}, b, \alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i \{y_i[\vec{w}.\vec{x}_i + b] - 1\} (i=1,2,\ldots) \quad (21)$$

Di mana α_i adalah *Lagrange Multiplier* yang bernilai nol atau positif ($\alpha_i \ge 0$). Nilai optimal dari persamaan (21) dapat dihitung dengan meminimalkan L terhadap $\vec{w}$ dan b, dan memaksimalkan L terhadap α_i. Dengan memperhatikan sifat bahwa pada titik optimal gradient L = 0, persamaan (21) dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung saja α_i, sebagaimana persamaan (22) di bawah.

Maximize :

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i \vec{x}_j , \quad (22)$$

yang mana, $\alpha_i \ge 0$ (i=1,2,..l) $\sum_{i=1}^{l} \alpha_i y_i = 0.$ (23)

Untuk menyelesaikan problem non linear, SVM dimodifikasi dengan memasukkan fungsi Kernel.

Fungsi kernel dirumuskan sebagai berikut [14] :

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i).\Phi(\vec{x}_j) \quad (24)$$

Fungsi Kernel memberikan berbagai kemudahan, karena dalam proses pembelajaran SVM, untuk menentukan *support vector*, kita hanya cukup mengetahui fungsi kernel yang dipakai, dan tidak perlu mengetahui wujud dari fungsi non linear Φ.

Tabel 2. Fungsi Kernel yang umum pada SVM

Jenis Kernel	Fungsi
Polynomial	$K(\vec{x}_i, \vec{x}_j) = \left((\vec{x}_i, \vec{x}_j) + 1\right)^p$ p=1,…
Gaussian Radial Basis Function (RBF)	$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{2\sigma^2}\right)$
Sigmoid	$K(\vec{x}_i, \vec{x}_j) = \tanh(\alpha.\vec{x}_i \vec{x}_j + \beta)$

Selanjutnya hasil klasifikasi dari data $\vec{x}$ diperoleh dari persamaan berikut :

$$f[\Phi(\vec{x})] = \vec{w}\,\Phi(\vec{x}) + \vec{b} \quad (25)$$

$$f[\Phi(\vec{x})] = \sum_{i=1,SVs} \alpha_i y_i \Phi(\vec{x}).\Phi(\vec{x}_i) + \vec{b} \quad (26)$$

$$f[\Phi(\vec{x})] = \sum_{i=1,SVs} \alpha_i y_i \, K(\vec{x}_i, \vec{x}_j) + \vec{b} \quad (27)$$

Nilai SV pada persamaan di atas adalah subset dari training set yang terpilih sebagai *support vector*, dengan kata lain data $\vec{x}_i$ yang berkorespondensi pada $\alpha_i \geq 0$.

C. Kanker Serviks

Leher rahim (serviks) adalah bagian dari sistem reproduk-si perempuan yang terletak di bagian bawah yang sempit dari rahim (uterus atau *womb*) [15]. Kanker ini merupakan kanker ganas yang terbentuk dalam jaringan ser-viks (organ yang menghubungkan uterus dengan vagina). Ada beberapa tipe kanker serviks. Tipe yang paling umum dikenal adalah *squamous cell carcinoma* (SCC), yang merupakan 80 hingga 85 persen dari seluruh jenis kanker serviks. Infeksi *Human Papilloma Virus* (HPV) merupakan salah satu faktor utama tumbuhnya kanker jenis ini [16].

Faktor resiko bukanlah penyebab mutlak akan terjangkitnya kanker rahim, namun faktor – faktor tersebut mampu meningkatkan resiko ter-kena penyakit kanker serviks menjadi lebih besar. Faktor resiko tersebut adalah Riwayat keluarga [5], *Human Papploma Virus* (HPV), Tes *Pap Smear*, Merokok [6], Usia [16], Menikah dini, banyak anak, hubungan seksual, kontrasepsi, hamil muda [17].

III. METODOLOGI

A. Data Penelitian

Data yang digunakan merupakan data sekunder mengenai Kanker Leher Rahim yang didapatkan dari Studi Kohort Faktor Resiko Penyakit Tidak Menular di Kota Bogor 2011 oleh Kemen-kes RI dengan banyak data 729 responden. Unit penelitian adalah perempuan usia 25-65 tahun di Bogor. Data awal sebanyak 1032 responden, peneliti mengurangi jumlah responden karena adanya *missing* data :

Tabel 3. Variabel penelitian

Kode	Variabel	Definisi
Y	Diagnosa kanker serviks	1: Terjangkit 2: Tidak terjangkit
X_1	Usia	Usia responden saat survey
X_2	Status Pernikahan	1: Iya 2: Tidak
X_3	Jumlah pasangan seksual	1: 1 pasangan 2: > 1 pasangan
X_4	Pendarahan saat menstruasi	1: Iya 2: Tidak
X_5	Usia pertama melahirkan	Usia melahirkan anak pertama
X_6	Banyak anak	Jumlah anak yang dilahirkan
X_7	Jenis kontrasepsi	1: Hormonal 2: Tidak hormonal
X_8	Waktu kontrasepsi	Lama penggunaan kontrasepsi
X_9	Riwayat kangker pada keluarga	1: Ada 2: Tidak
X_{10}	Vaksinasi HPV	1: Pernah 2: Tidak pernah
X_{11}	Usia menikah	Usia pertama menikah
X_{12}	Uji *Pap Smear*	1: Pernah 2: Tidak
X_{13}	Merokok	1: Iya 2: Tidak

Penelitian ini menggunakan variabel respon (Y) biner, yaitu Terjangkit Kanker Serviks (y=1) ada sebanyak 4 responden dan Tidak Terjangkit Kanker Serviks (y=2) sebanyak 725 responden. Pada variabel Status Pernikahan (X_2) dalam Kuesioner Kohort terdapat empat pilihan jawaban :

1: Belum Menikah [dikoding 2]
2: Menikah [dikoding 1]
3: Cerai hidup [dikoding 1]
4: Cerai mati [dikoding 1],

untuk pilihan 3 dan 4 peneliti memasukkan ke pilihan Menikah karena pada dasarnya yang telah bercerai telah menikah sebelumnya. Variable Pendarahan [X_4] saat Menstruasi dihilangkan observasi yang memiliki nilai 0.

Variabel Banyak anak yang dilahirkan (X_6) pada Kuesioner Kohort terdapat pilihan pengisian :

1: Lahir Hidup
2: Lahir Mati,

dalam penelitian ini dua pilihan tersebut digabungkan, karena sama-sama memiliki informasi yang dibutuhkan, yaitu jumlah anak yang telah lahir.

Variabel Penggunaan Kontrasepsi [X_7] dihapus karena semua responden menggunakan kontrasepsi, atau semua observasi bernilai X_7=1. Variabel Usia pertama menikah (X_{12}) pada Kuesioner diwakili oleh pertanyaan "umur pertama kali berhubungan intim" pada kode Gc.04.

B. Langkah Analisis

Langkah-langkah analisis data yang digunakan dalam penelitian ini adalah :

1. Melakukan pengumpulan data sekunder dari penelitian Kohort Litbangkes 2011.
2. Melakukan statistika deskriptif untuk melihat karakteristik data.
3. Mendapatkan faktor-faktor yang mempengaruhi terjangkitnya penyakit kanker serviks di Kota Bogor menggunakan Regresi Logistik Biner, dengan langkah analisis:
   a. Seleksi kandidat dengan Uji Univariabel
   b. Estimasi Parameter β_j
   c. Melakukan Pengujian Parameter
   d. Uji Kesesuaian Model
4. Membuat model klasifikasi kanker serviks menggu-nakan *Regresi Logistik Biner* dengan kombinasi :
   a. Training-testing 90:10
   b. Training-testing 70:30
   c. Training-testing 50:50
5. Menghitung performansi klasifikasi Regresi Logistik Biner dengan pengukuran *Accuracy*, *Specificity*, dan *Sensitifity*.
6. Membuat model klasifikasi penyakit kanker serviks menggunakan *Support Vector Machine* (SVM). Dengan langkah analisis :
   a. Menentukan data training-testing dengan 3 kombinasi;
      I. Training-testing 90:10
      II. Training-testing 70:30
      III. Training-testing 50:50
   b. Menentukan Fungsi Kernel yang dipakai, dalam penelitian kali ini menggunakan *Gaussian Radial Basis Function* (RBF).
   c. Menentukan parameter *C=10* dan *σ=2*.

7. Menghitung klasifikasi beserta ketepatan akurasinya dengan pengukuran *Accuracy*, *Specificity*, dan *Sensitifity*..
8. Membandingkan performansi antara ketepatan klasifikasi Regresi Logistik Biner dengan *Support Vector Machine* (SVM).

IV. ANALISIS DAN PEMBAHASAN

A. Deskripsi Faktor Resiko kanker Serviks

Statistika deskriptif digunakan untuk variabel yang berskala interval/rasio, sedangkan tabulasi silang digunakan untuk variabel berskala nominal/ordinal.

Tabel 4. Statistika Deskriptif variabel berskala rasio

Variabel	Diagnosa	Mean	Stdev	Min	Max
Usia (tahun) [X1]	Terjangkit	43	12,57	26	53
	Tidak terjangkit	42,19	10,01	25	65
Usia pertama melahirkan (tahun) [X5]	Terjangkit	23,5	3,12	19	26
	Tidak terjangkit	22,35	3,92	13	40
Jumlah anak (anak) [X6]	Terjangkit	3,5	2,38	2	7
	Tidak terjangkit	3,22	1,87	1	13
Lama kontra-sepsi (hari) [X9]	Terjangkit	22,75	25,81	0	48
	Tidak terjangkit	97,74	92,02	0	926
Usia saat menikah (tahun) [X12]	Terjangkit	23	3,37	18	25
	Tidak terjangkit	21,05	5,24	12	88

Dari tabel di atas didapatkan informasi bahwa usia rata-rata responden yang terjangkit adalah 43 tahun, sedangkan yang tidak terjangkit adalah 42,19 tahun. Rata-rata usia pertama kali responden yang terjangkit melahirkan adalah 23,5 tahun, sedangkan yang tidak terjangkit adalah 22,35 tahun. Rata-rata banyaknya anak yang dimiliki responden yang terjangkit adalah 4 anak, sedangkan yang tidak terjangkit 3 anak.

Dari responden yang terjangkit, usia pertama melahirkan yang paling muda adalah 19 tahun, sedangkan paling tua berusia 26 tahun. Dari responden yang tidak terjangkit, usia pertama melahirkan yang paling muda adalah 13 tahun, sedangkan paling tua berusia 40 tahun. Dari respon-den yang terjangkit, usia saat menikah yang paling muda adalah 18 tahun, sedangkan paling tua berusia 25 tahun. Responden yang tidak terjangkit, usia saat menikah yang paling muda adalah 12 tahun, yang paling tua 88 tahun.

Berikut ini hasil tabulasi silang variabel independen yang berskala nominal terhadap Diagnosa Penyakit Kanker Serviks.

Tabel 5. Tabulasi silang variabel berskala nominal

Variabel		Diagnosa	
		Terjangkit	Tidak terjangkit
Status pernikahan	Nikah	0,549%	99,314%
	Tidak	0,000%	0,137%
Jumlah pasangan	1 pasang	0,412%	88,889%
	> 1 pasang	0,137%	10,562%
Pendarahan saat mens	Iya	0,000%	6,447%
	Tidak	0,549%	93,004%
Kontrasepsi	Iya	0,412%	81,481%
	Tidak	0,137%	17,970%
Riwayat Keluarga	Ada	0,137%	3,978%
	Tidak	0,412%	95,473%
Vaksinasi HPV	Pernah	0,000%	0,274%
	Tidak	0,549%	99,177%
Tes *Pap Smear*	Pernah	0,137%	6,584%
	Tidak	0,412%	92,867%
Merokok	Iya	0,274%	27,160%
	Tidak	0,274%	72,291%

Tabel tersebut memperlihatkan karakteristik hubu-ngan antara variabel faktor resiko dengan diagnosa penyakit kanker serviks. Responden yang terjangkit kanker serviks 0,549% belum pernah vaksinasi HPV. Terlihat pula bahwa responden yang terjangkit kanker serviks 0,274% adalah perokok dan tidak pernah mengalami pendarahan saat menstruasi. Responden yang tidak terjangkit kanker serviks 92,87% belum pernah uji *Pap Smear* dan 95,5% tidak memiliki riwayat kanker pada keluarga.

B. Analisis dengan Regresi Logistik Biner

Tabel 6. Hasil uji univariabel

Variabel	B	df	*P-value*
Y dengan X_1			
Usia	-0,008	1	0,870
Constant	5,547	1	0,012
Y dengan X_2			
Status.nikah	-16,004	1	1,000
Constant	21,203	1	1,000
Y dengan X_3			
Jumlah.pasangan.seks	1,031	1	,374
Constant	4,344	1	0,000
Y dengan X_4			
Pendarahan.mens	16,070	1	0,998
Constant	5,133	1	0,000
Y dengan X_5			
Usia.melahirkan1	-0,068	1	0,557
Constant	6,761	1	0,015
Y dengan X_6			
Banyak.anak	-0,072	1	0,765
Constant	5,441	1	0,000
Y dengan X_7			
Jenis.kont	0,413	1	0,721
Constant	4,875	1	0,000
Y dengan X_8			
Lama.kont	0,031	1	**0,119**
Constant	3,796	1	0,000
Y dengan X_9			
Riwayat.keluarga	-2,079	1	**0,076**
Constant	5,447	1	0,000
Y dengan X_{10}			
Vaksin.HPV	16,006	1	1,000
Constant	5,197	1	0,000
Y dengan X_{11}			
Usia.nikah	-0,033	1	0,474
Constant	5,928	1	0,000
Y dengan X_{12}			
PapSmear	-1,548	1	**0,184**
Constant	5,419	1	0,000
Y dengan X_{13}			
Merokok	-0,979	1	0,329
Constant	5,574	1	0,000

Dari tabel di atas terlihat bahwa variabel yang signifikan pada taraf nyata 80% (α = 0,2) dalam Uji Univariabel adalah Lama Pemakaian Kontrasepsi (X_8), Riwayat Kanker pada Keluarga (X_9), dan Tes *Pap Smear* (X_{12}). Ketiga variabel yang signi-fikan tersebut akan

dimasukkan dalam model dan diuji secara serentak dan parsial..

Tabel 7. Nilai *Overall test*

	Chi-square	df	P-value
Step	10,057	3	0,018
Block	10,057	3	0,018
Model	10,057	3	0,018

Hipotesis yang digunakan adalah :

H_0 : $\beta_8=\beta_9=\beta_{12}=0$(Variabel indepependen tidak mempengaruhi variabel dependen)

H_1 : Minimal satu $\beta_i \neq 0$ (Minimal satu variabel independen yang berpengaruh)

i = 8,9,12

Statistik uji : $$G = -2\ln \frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\sum_{i=1}^{n} \hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}}$$

Tolak H_0 jika $G > \chi^2_{(v,\alpha)}$ atau nilai *P-value* < 0,1.

Keputusan : Terlihat dari tabel bahwa nilai P-value 0,018 yang berarti kurang dari α=0,1. Maka keputusannya adalah tolak H_0. Jadi seluruh prediktor secara bersama-sama berpengaruh terhadap terjangkitnya penyakit kanker serviks. Juga bisa disimpulkan minimal ada satu variabel independen yang mempengaruhi variabel dependen.

Tabel 8. Hasil Uji Individu dan estimasi parameter

	β	Wald	P-value	Exp (B)
Lama Kontrasepsi	0,035	2,826	0,093	1,036
Riwayat keluarga	-2,354	3,668	0,055	0,095
Tes *PapSmear*	-2,218	3,253	0,071	0,109
Constant	4,265	29,737	0,000	71,170

Terlihat dari Tabel 6 nilai koefisien parameter (β_i) adalah :

β_1^T : [β_0, β_8, β_9, β_{12}]

: [4.265, 0.035, -2.354, -2.218]

Setiap penambahan satu satuan waktu Lama kontrasepsi akan menambah peluang terjangkitnya kanker serviks sebesar 0,035. Seorang wanita yang tidak memiliki riwayat kanker pada keluarga kemungkinan terserang kanker serviks adalah 10,5 kali (lebih besar) dari pada yang memiliki riwayat keluarga. Seorang wanita yang tidak rutin tes *Pap Smear* kemungkinan terserang kanker serviks adalah 9,2 kali (lebih besar) dari pada yang pernah tes *Pap Smear*.

Dari nilai tabel di atas juga terlihat bahwa seluruh variabel memiliki nilai *P-value* < α=0,1. Jadi semua variabel yang lolos Uji Univariabel, yaitu Lama Kontrasepsi, Riwayat Keluarga, dan Tes *Pap Smear*, berpengaruh signifikan terhadap diagnosa kanker serviks.

Tabel 9. Uji *Goodness of fit*

Chi-square	df	P-value
2,639	8	0,955

Hipotesis yang digunakan adalah :

H_0: Model telah sesuai (tidak ada perbedaan signifikan antara hasil pengamatan dengan kemungkinan nilai prediksi)

H_1: Model tidak sesuai (ada perbedaan signifikan antara hasil pengamatan dengan kemungkinan nilai prediksi)

Daerah penolakan; tolak H_0 jika nilai *P-value* < 0,2

Statistik Uji: $$G^2 = 2\sum_{i=0}^{i}\sum_{j=0}^{j} n_{ijk}\log\frac{n_{ij}}{\hat{m}_{ij}}$$

Keputusan : Terlihat bahwa nilai P-value lebih dari 0.1, maka keputusannya adalah gagal tolak H_0. Jadi pada keyakinan 90% model yang terbentuk telah sesuai, atau model tersebut mampu menjelaskan data dan (tidak ada perbedaan signifikan antara hasil pengamatan dengan kemungkinan nilai prediksi.

Variabel yang dimasukkan sebagai model akhir Regresi Logistik Biner adalah yang signifikan, yaitu konstanta, X_{10} dan $X_{14.}$ Model yang terbentuk adalah :

Model Logit :

$$\hat{g}(x)=4.265+0.035X_{8(1)}-2,354X_{9(1)}-2,218X_{12(1)}$$

Model regresi logistiknya adalah sebagai berikut :

$$\hat{\pi}_1(x) = \frac{e^{4.265+0.035X_{8(1)}-2,354X_{9(1)}-2,218X_{12(1)}}}{1+e^{4.265+0.035X_{8(1)}-2.354X_{9(1)}-2.218X_{12(1)}}}$$

$$\hat{\pi}_0(x) = 1-\hat{\pi}_1(x)$$

- $\hat{\pi}_1(1,1,1) = 0,432$, $\hat{\pi}_0(1,1,1) = 0,568$

  Peluang seorang perempuan yang lama menggunakan kontrasepsi, mempunyai riwayat kanker pada keluarga, dan tes PapSmear, untuk terjangkit kanker serviks sebesar 0,43. Sedangkan peluang untuk tidak terserang sebesar 0,57.

- $\hat{\pi}_1(0,0,0) = 0,987$, $\hat{\pi}_0(0,0,0) = 0,013$

  Peluang seorang perempuan yang tidak lama menggunakan kontrasepsi, tidak mempunyai riwayat kanker pada keluarga, dan tidak tes PapSmear, untuk terjangkit kanker serviks sebesar 0,987. Sedangkan peluang untuk tidak terserang sebesar 0,013.

- $\hat{\pi}_1(1,1,0) = 0,875$, $\hat{\pi}_0(1,1,0) = 0,124$

  Peluang seorang perempuan yang lama menggunakan kontrasepsi, mempunyai riwayat kanker pada keluarga, dan tidak tes PapSmear, untuk terjangkit kanker serviks sebesar 0,88. Peluang tidak terserang sebesar 0,12.

- $\hat{\pi}_1(0,0,1) = 0,886$, $\hat{\pi}_0(0,0,1) = 0,114$

  Peluang seorang perempuan yang tidak lama menggunakan kontrasepsi, tidak mempunyai riwayat kanker pada kel-uarga, dan tes PapSmear, untuk terjangkit kanker serviks sebesar 0,43. Sedangkan peluang untuk tidak terserang sebesar 0,57.

Uji Parameter dan Kelayakan Model tiap kombinasi

Tabel 10. Hasil uji data Training set 50%

	B	df	Sig.	Overall Test	Hosmer-Lemeshow Test
Lama Kontrasepsi	-0,006	1	0,861	0,085	1,000
Riwayat Keluarga	-3,183	1	0,053		
Tes Pap Smear	-2,988	1	0,076		
Constant	6,759	1	0,000		

Dari tabel terlihat bahwa dengan data training 50% ketiga variabel secara serentak berpengaruh terhadap respon. Secara individu yang berpengaruh signifikan adalah Riwayat Keluarga dan Tes Pap Smear. Diketahui juga bahwa model yang terbentuk telah dianggap baik/layak.

Tabel 11. Hasil uji data Training set 70%

	B	df	Sig.	Overall Test	Hosmer-Lemeshow Test
Lama Kontrasepsi	0,024	1	0,287	0,085	0,076
Riwayat Keluarga	-2,774	1	0,035		
Tes Pap Smear	-2,619	1	0,048		
Constant	5,085	1	0,000		

Dari tabel terlihat bahwa dengan data training 70% ketiga variabel secara serentak berpengaruh terhadap respon. Secara individu yang berpengaruh signifikan adalah Riwayat Keluarga dan Tes Pap Smear. Diketahui juga bahwa model yang terbentuk telah dianggap baik/layak.

Tabel 12. Hasil uji data Training set 90%

	B	df	Sig.	Overall Test	Hosmer-Lemeshow Test
Lama Kontrasepsi	0,027	1	0,166	0,046	0,987
Riwayat Keluarga	-2,777	1	0,035		
Tes Pap Smear	-2,627	1	0,048		
Constant	5,007	1	0,000		

Dari tabel terlihat bahwa dengan data training 90% ketiga variabel secara serentak berpengaruh terhadap respon. Secara individu yang berpengaruh signifikan adalah Riwayat Keluarga dan Tes Pap Smear. Diketahui juga bahwa model yang terbentuk telah dianggap baik/layak

Untuk mengetahui ketepatan hasil klasifikasi pada penelitian ini ada beberapa cara, yaitu dengan *sensitivity, specificity*, dan *accuracy*. Untuk mengetahui kombinasi mana yang menghasilkan ketepatan klasifikasi paling tinggi perlu dibandingkan ketiga kombinasi tersebut.

Tabel 13. Perbandingan Ketepatan Klasifikasi

Kombinasi	Akurasi	*Specivicity*	*Sensitivity*
50-50	55,5%	100%	55,2%
70-30	87,7%	100%	87,6%
90-10	100%	100%	100%

Dari tabel di atas terlihat bahwa tingkat akurasi klasifikasi paling tinggi dihasilkan kombinasi *traning:testing* 90:10 yaitu sebesar 100%. Kombinasi 70:30 menghasilkan akurasi sebesar 87,7%. Sedangkan untuk kombinasi 90:10 menghasilkan akurasi 55,5%. Nilai *specificity* semua 100%. Dari hasil perhitungan di atas ada indikasi bahwa semakin banyak data training maka akan menghasilkan performansi klasifikasi yang lebih tinggi.

C. Analisis menggunakan *Support Vector Machine* (SVM)

Analisis SVM pada penelitian ini menggunakan fungsi kernel Polinomial dengan parameter σ=2. Parameter SVM sebagai titik penalt dengan C=10. Agar bisa dibandingkan dengan ketepatan klasifikasi Regresi Logistik Biner, maka analisis SVM ini juga menggunakan kombinasi data *training-testing* 50:50, 70:30, dan 90:10.

Pada fungsi pengalih *Lagrange Multiplier*

$L(\vec{w},b,\alpha)=\frac{1}{2}\|\vec{w}\|^2-\sum_{i=1}^{l}\alpha_i\{y_i[\vec{w}.\vec{x}_i+b]-1\}$ terdapat nilai α_i yang bernilai positif. Nilai optimal persamaan tersebut adalah dengan memaksimalkan L terhadap α_i.

Tabel 14. Ketepatan Klasifikasi SVM

Kombinasi	Akurasi	*Specivicity*	*Sensitivity*
50-50	100%	0%	100%
70-30	100%	0%	100%
90-10	100%	0%	100%

Dari tabel tersebut terlihat bahwa tingkat akurasi klasifikasi semua kombinasi *training-testing* adalah sama yaitu sebesar 100%. Nilai *Specificity* sebesar 0% karena tidak ada observasi Terjangkit yang diprediksi Terjangkit. Semua observasi (pada data *testing*) dan prediksi menunjuk pada kategori Tidak Terjangkit.

D. Perbandingan akurasi SVM dengan Logistik Biner

Dari tabel 4.11 dan 4.12 terlihat bahwa pada penelitian kali ini tingkat akurasi *Support Vector Machine* mempunyai nilai akurasi yang sangat tinggi jika dibandingkan Logistik Biner, baik proporsi 90:10, 70:30, maupun 50:50. Hal ini terjadi *overfitting* karena proporsi kategori respon yang tidak seimbang. Dari total 729 responden, hanya 4 orang yang terjangkit. Selebihnya 725 responden tidak terjangkit. Data dengan proporsi respon yang tidak seimbang ini menyebabkan prediksi secara keseluruhan mengarah kepada prediksi prediksi bahwa responden tidak terjangkit kanker serviks..

V. KESIMPULAN DAN SARAN

Variabel yang berpengaruh signifikan terhadap Kanker Serviks pada Analisis Regresi Logistik Biner adalah Lama Kontrasepsi, Riwayat Keluarga, dan tes PapSmear.

Performansi klasifikasi menggunakan SVM pada semua kombinasi baik 90:10, 70:30, dam 50:50 adalah sebesar 100%, sedangkan nilai specificity semua 0%. Akurasi klasifikasi menggunakan Logistik Biner tertinggi adalah kombinasi 90:10 sebesar 100%, kombinasi 70:30 sebesar 87,7%, sedangkan kombinasi 50:50 sebesar 55,5%. Nilai specificity Logistik Biner semua 100%, jadi responden yang terjangkit semua bisa diprediksi terjangkit. Pada SVM nilai *sensitifity* sebesar 100%. Hal ini menunjukkan bahwa prediksi menuju kepada prediksi kategori tak terjangkit. Terjadi demikian karena proporsi kategori yang tidak seimbang.

Saran dari penulis, jika terdapat kasus dengan kategori respon yang tidak seimbang, maka untuk mendapatkan hasil yang lebih baik dan tidak terjadi *over fitting* perlu digunakan metode SVM untuk *inballanced data*

VI. UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih kepada Litbangkes Kemenkes RI yang telah memberikan dukungan perihal penyediaan data.

DAFTAR PUSTAKA

[1] World Health Organization. (2001). *Effective screening programmes for cervical cancer in low- and middle-income developing countries.* India : Bulletin of WHO.

[2] Susanti, Desi. (2012). *Pemeriksaan Pap Smear*. Riau : STIKES Tuanku Tambusai Bakinang.

[3] Bogor, Kota. (2011). *Seminar Kesehatan "Peduli Perempuan: Cintai Diri, Cegah, Dan Deteksi Kanker Serviks Sejak Dini"*. Retrieved March, 2014, from Web Site: http://www.kotabogor.go.id.

[4] Junita. (2014). *Faktor Resiko Kanker Rahim*. Retrieved March, 2014, from Web Site: www.health.detik.com.
[5] Modern Cancer Hospital Guangzhou. (2014). *Faktor Resiko Kanker Rahim*. Retrieved March, 2014, from Web Site: www.asiancancer.com.
[6] Mc Cormick, C.,C., Giuntoli, R., L. (2011). *Patient's Guide to Cervical Cancer*. Baltimore : The John Hopkins Health Corporation
[7] Intansari, I.A.S. (2012). *Klasifikasi Pasien Hasil pap Smear Test sebagai Pendeteksi Awal Upaya Penanganan Dini pada Penyakit Kanker Serviks di RS "X" Surabaya dengan metode Bagging Logistc Regression*. Surabaya: Institut Teknologi Sepuluh Nopember.
[8] Nugroho, A.S., Handoko, D., Witarto, A.B. (2003). *Support Vector Machine – Teori dan Aplikasinya dalam Bioinformatika*. BPPT.
[9] Rahman, Farizi. (2012). *Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine*. Surabaya: Institut Teknologi Sepuluh Nopember.
[10] Walpole, R. E. (1995). *Pengantar Statistika Edisi ke-3*. Jakarta: PT Gramedia Pustaka Utama.
[11] Hosmer, D.,W., Lemeshow, S. (2000). *Applied Regression Logis-tic, Second Edition*. Canada: John Wiley & Son's, Inc.
[12] Agresti, Alan. (2002), *Categorical Data Analysis Second Edition*. New York: John Wiley & Son's, Inc.
[13] Hsu, C.W., Chang, C.C., Lin, C.J. (2003). *A Practical Guide to Support Vector Classification*. England : University of Southampton.
[14] Gunn, Steve. (1998). *Support Vector Machine for Classification and Regression*. Taiwan : National Taiwan University.
[15] Evennet, Karen. (2003). *Pap Smear, Apa yang Perlu Anda Ke-tahui*. Jakarta : Arcan Publisher.
[16] Canhope. (2014). *Apa itu Kanker Serviks?*. Retrieved March, 2014, from Web Site: http://www.parkwaycancercentre .com.
[17] Rouzeau, Vanessa. (2012). *Cervical Cancer : A Review*. Florida : Herzing University.