Prediksi Kepercayaan Pengguna Internet dalam Paradigma Komputasi Pervasif dengan Regresi Logistik Berganda

Puguh Jayadi
Magister Teknik Informatika, Universitas Islam Indonesia,
Jalan Kaliurang Km.14,5, Sleman 55584, Indonesia
e-mail: **puguh.jayadi@students.uii.ac.id**

Abstract - Prediction on internet user confidence in the pervasive computing paradigm has been done before, but does not rule out using other methods. The multiple logistic regression prediction model is used as a development. The outcome variable (dependent) in the dataset has a binominal data type which indicates the trust status of internet users on several predictor variables (independent) which are determined by a number of polynominal data type factors. The use of a multiple logistic regression prediction model was carried out several times with different percentages of dataset distributions resulting in different regression coefficients. This is done to find the maximum predicted truth value through the output values of the Confusion Matrix and the ROC curve. The results of this study indicate that the multiple logistic regression model can be used effectively to predict the trust of internet users in a pervasive computing paradigm through the stages of data preparation and the distribution of training data and testing data accordingly. From some data sharing, this model predicts an accuracy of almost 100%

Keywords *- Multiple logistic regression, preprocessing, Pervasive Computing, Confusion Matrix, ROC*

Abstrak - Prediksi terhadap data kepercayaan pengguna internet dalam paradigma pervasive computing pernah dilakukan sebelumnya, namun tidak menutup kemungkinan untuk menggunakan metode lain. Model prediksi regresi logistik berganda digunakan sebagai pengembangannya. Variabel hasil (dependen) pada dataset memiliki tipe data binominal menunjukkan keadaan kepercayaan pengguna internet terhadap beberapa variabel prediktor (independen) yang didefinisikan dengan sejumlah faktor-faktor bertipe data polynominal. Penggunaan model prediksi regresi logistik berganda dilakukan beberapa kali dengan persentase pembagian dataset berbeda sehingga menghasilkan koefisien regresi yang berbeda-beda pula. Hal ini dilakukan untuk menemukan nilai kebenaran prediksi yang maksimal melalui nilai keluaran dari Confusion Matrix dan kurva ROC. Hasil penelitian ini menunjukkan bahwa model regresi logistik berganda dapat digunakan secara efektif untuk memprediksi kepercayaan pengguna internet dalam paradigma pervasive computing dengan terlebih dahulu melalui tahap preprosessing data serta pembagian data training dan data testing yang sesuai. Dari beberapa pembagian data, model memprediksi hampir mencapai tingkat akurasi 100%

Kata kunci *- Regresi logistic berganda, preprosessing, Pervasive Computing, Confusion Matrix, ROC*

I. PENDAHULUAN

Dalam pengembangan teknologi komputer istilah *Ubiquitous/Pervasive Computing* yaitu gagasan baru dalam perkembangan di dunia komputer setelah adanya *mainfame* (satu komputer, banyak pengguna) dan *PC* (satu komputer, satu pengguna) [1]. *Pervasive Computing* dianggap suatu paradigma bahwa komputer digunakan tanpa disadari dimana saja, kapan saja, oleh siapa saja, tanpa mengurangi fungsi utamanya dalam meningkatkan efektifitas kerja pada lingkungan penggunanya dengan tingkat visibilitas yang rendah [1], [2].

Pervasive Computing bertujuan untuk menciptakan layanan secara langsung antara pengguna dan lingkungannya, dengan sangat mengurangi panduan manusia langsung [3]. *Pervasive Computing* biasanya menggunakan teknologi komunikasi nirkabel yang umum tersedia, seperti *WiFi*, *Bluetooth*, dan teknologi sensor RFID [4]. Teknologi tersebut lebih bernilai dan memberikan peran penting dalam penyimpanan, transaksi keuangan, serta informasi pribadi lainnya [5].

Namun keadaan ini menjadi rawan ketika ada beberapa komputer saling terhubung secara bersamaan dengan sistem komputasi tidak terlihat dan luas, menjadi sulit untuk mengetahui apa mengendalikan apa, terhubung dengan apa, di mana informasi mengalir, bagaimana digunakan dan apa konsekuensi dari tindakan yang diberikan [6]. Itulah yang menjadikan alasan adanya perbedaan kepercayaan pengguna internet pada penerapan *pervasive computing.*

Penelitian yang akan dilakukan ini merupakan pengembangan serta menggunakan data yang dihimpun dari penelitian mengenai model kepercayaan berbasis kecerdasan buatan untuk komputasi pervasif [6] dan model kepercayaan untuk komputasi pervasif berdasarkan pembelajaran aturan asosiasi apriori dan klasifikasi bayesian [7]. Secara khusus memfokuskan pada pembuatan model prediksi kepercayaan pengguna internet dalam penerapan *pervasive computing.* Dataset tersebut memiliki beberapa atribut

yang dapat menggambarkan keadaan pengguna internet dan digunakan sebagai parameter prediksi [8]. Kelas/label sebagai variabel hasil/ *dependen* pada dataset terdiri dari dua kelas (*binominal*) sehingga peneliti berupaya untuk mengembangkannya dengan metode prediksi lain.

Penggunaan metode statistik dan pendekatan data mining yang digunakan untuk prediksi yaitu Regresi Logistik Berganda. Tahapan *preprosessing* yang dilakukan adalah mengubah bentuk/tipe data, pembagian dataset/*split data* menjadi dua bagian yaitu data *training* dan data *testing* serta dengan persentase berbeda. Pembagian data tersebut digunakan untuk membangun model regresi dan melakukan pengujian prediksi menggunakan *Confusion Matrix* dan kurva *ROC*. Dari penelitian ini diharapkan menghasilkan tingkat kebenaran prediksi yang maksimal dari penelitian yang sudah dilakukan.

Penelitian tentang pengembangan model kepercayaan penerapan *pervasive computing* oleh Palmieri, et al, [6] berdasarkan pada berbagai konteks dan sumber informasi kepercayaan yang berbeda. Model tersebut mengintegrasikan aturan asosiasi algoritma Apriori dengan klasifikasi Naive Bayes. Dengan penelitian ini menghasilkan fitur/model yang mewakili tanda dari pola perilaku suatu entitas yang melakukan aktifitas tertentu pada jaringan internet [6].

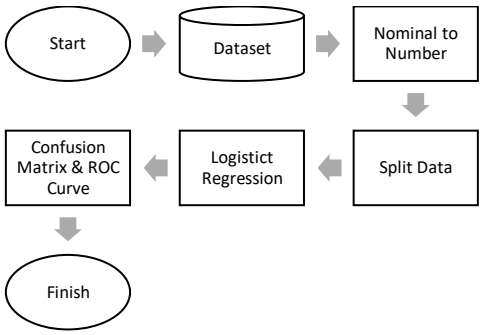
Kelanjutan penelitian tersebut dilakukan oleh Angelo, et al, [7] dengan menggunakan metode dan data yang tidak jauh berbeda dengan penambahan validasi model menggunakan *Performance Metrics* [7]. Dari kedua penelitian yang sudah dilakukan mengenai prediksi penggunaan internet pada *pervasive computing* tidak menutup kemungkinan untuk menggunakan metode lain dalam melakukan prediksi pada dataset yang sama.

Pemodelan prediksi adalah teknik menggunakan informasi historis pada atribut atau acara tertentu untuk mengidentifikasi pola lama yang akan membantu menemukan pola prediksi baru [9]. Regresi logistik merupakan salah satu model prediksi analisis diskriminan yang bisa digunakan pada data beratribut banyak. Altman adalah pelopor regresi pada tahun 1968 namun regresi berganda lebih populer digunakan oleh Ohlson pada tahun 1980 [10]. Probabilitas suatu peristiwa berada antara 0 dan 1, tidak praktis dan tergolong sulit untuk memodelkan probabilitas dengan teknik regresi linier, karena model regresi linier memungkinkan variabel dependen mengambil nilai > 1 atau < 0 [11]. Model regresi logistik memperluas model regresi linier dengan menghubungkan kisaran angka nyata ke kisaran 0 sampai 1 [11].

Banyak penelitian menggunakan regresi logistik berupaya membandingkan atau mengkombinasikan dengan berapa metode lain seperti neural network [10], CHAID regression trees [12], Naive Bayes [13], C4.5, KNN, Decission Stump, Random Forest [14]. Dari beberapa penelitian tersebut menemukan hasil kebenaran prediksi penggunaan regresi logistik terhitung tinggi dibanding dengan menggunakan metode lainnya. Oleh karena itu metode yang digunakan pada penelitian ini adalah regresi logistik berganda yaitu perluasan metode regresi biasa untuk melakukan prediksi pada dataset yang memiliki atribut atau variabel berupa data biner/binominal lebih dari satu [8], [15]. Sebelum pada bagian pengujian tingkat kebenaran prediksi, terlebih dahulu dilakukan pembagian data berbeda untuk mengukur masing-masing nilai keluaran yang akan dihasilkan oleh *Confusion Matrix* dan kurva *ROC* [7], [16].

II. METODE PENELITIAN

Bab ini menjelaskan materi atau bahan yang digunakan untuk menyelesaikan masalah. Pada **Gambar 1** menjelaskan beberapa rangkaian alur dalam melakukan penelitian. Dimulai dataset bertipe polinominal diubah dalam bentuk angka/*numeric* terlebih dahulu agar dapat diolah dengan model Berganda Regresi Logistik. Kemudian data akan dibagi sesuai dengan beberapa bagian untuk menguji tingkat kebenaran prediksi.



Gambar 1. Alur Penelitian

Kemudian perhitungan *Performance* dari penerapan model Regresi Logistik Berganda dengan *Confusion Matrix* dan *ROC Curve* untuk menghasilkan beberapa keluaran validasi seperti: *Accuracy*, *Sensitivity*, *Specificity*, dan *Precision.*

A. Persiapan Dataset

Penelitian ini menggunakan dataset berasal dari *UCI Repository* yang digunakan pada penelitian [6],[7]. Data tersebut merupakan data penerapan *pervasive computing* yang berisi pengguna internet berinteraksi kemudian pada suatu kondisi pengguna tidak dapat memperoleh informasi sehingga menimbulkan rasa saling tidak percaya satu sama lain [8]. Rangkuman detail tentang dataset ditunjukkan pada **Tabel** 1.

Penulis dalam penelitian ini tidak membahas mengenai bagaimana dataset tersebut dikumpulkan dan diolah, karena data tersebut didapatkan dari *UCI Repository* [8]. Seperti yang dijelaskan dibagian depan, data tersebut digunakan untuk menguji suatu keadaan berdasarkan model dan data yang digunakan untuk memprediksi tingkat kepercayaan pengguna yang berinteraksi di lingkungan *pervasive computing.*

Tabel 1. Dataset *Pervasive Computing*

Atribut	Tipe	Keterangan
Counting Trust (CT) - Menghitung Kepercayaan	Polinominal	Menghitung berapa banyak transaksi yang dapat dipercaya terjadi setelah transaksi terakhir yang tidak dapat dipercaya. {*CT_range_1,CT_range_2,CT_range_3,CT_range_4*}
Counting Un-trust (CU) - Menghitung Ketidakpercayaan	Polinominal	Menghitung berapa banyak transaksi yang tidak dapat dipercaya terjadi setelah transaksi terakhir yang dapat dipercaya. *{CU_range_1,CU_range_2, CU_range_3, CU_range_4}*
Last Time (LT) - Waktu Terakhir	Polinominal	Memperhitungkan tanggal di mana pengalaman terakhir dalam konteks tertentu terjadi. *{LT_range_1,LT_range_2,LT_range_3,LT_range_4}*
Transactions Context (TC) - Transaksi Konteks	Polinominal	Mengidentifikasi jenis transaksi, seperti game, e-commerce, jejaring sosial dan lain-lain. *{sport, game, ECommerce, holiday}*
Trust Score (TS) - Skor Kepercayaan	Binominal	Skor yang diberikan entitas kepada entitas lain di akhir setiap interaksi langsung. *{trustworthy, untrustworthy}*

B. Regresi Logistik Berganda

Analisis regresi kerap digunakan untuk mencari ketergantungan atau hubungan variabel prediktor (*independen*) dengan variabel hasil (*dependen*) dengan maksud membangun model/pola dari data yang sudah diketahui yang bisa digunakan memprediksi kelas/label data baru yang belum diketahui [17]. Model regresi memberikan hasil yang lebih baik pada nilai-nilai data bersifat angka/numerik, tetapi juga memungkinkan untuk memprediksi variabel bersifat diskrit/nominal dari nilai campuran variabel prediktor angka dan diskrit [13][18].

Selaian itu Regresi Logistik lebih banyak digunakan ketika variabel respon adalah biner yang artinya hanya dapat terdapat nilai 1 atau 0 [10]. Menurut [19] persamaan regresi logistik berganda yang digunakan secara umum dikutip dari [20] ditunjukkan pada **Persamaan 1**.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i)} \quad \textbf{(1)}$$

Dimana $\pi(x)$ merupakan peluang terjadinya benar dengan nilai kemungkinan $0 \leq \pi(x) \leq 1$ berarti 1 adalah benar dan 0 adalah salah. Dari rumus tersebut diubah dalam bentuk logit menjadi **Persamaan 2** dan **Persamaan 3** [17].

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i) \quad \textbf{(2)}$$

Maka menghasilkan model regresi logistik:

$$g(x) = logit[\pi(x)] = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i \quad \textbf{(3)}$$

Dengan demikian *g(x)* merupakan fungsi hubungan dari model regresi logistik yang disebut sebagai fungsi hubungan logit [17]. Regresi logistik dapat dijadikan sebagai metode pendekatan prediksi karena tidak membutuhkan hubungan linier antara variabel hasil (*dependen*) dengan variabel prediktor (*independen*), tidak membutuhkan asumsi *error varians (residual)* terdistribusi normal seperti halnya regresi linier. Selain itu, bisa digunakan pada data dengan jumlah relatif besar, minimum dibutuhkan 50 sampel data yang dijadikan variabel prediktor (*independen*) [21].

C. Pengubahan Nominal ke Angka

Ada beberapa metode yang digunakan dalam prediksi hanya dapat diterapkan pada variabel prediktor (*independen*) bersifat angka/numerik dan tidak mendukung pada tipe data nominal/kategoris [22]. Hal ini disebabkan karena variabel skala nominal/kategoris tidak memiliki skala yang didefinisikan dengan baik dengan interval tetap, sehingga tidak cocok sebagai variabel prediktor (*independen*) dalam model regresi [23]. Kenyataan itu menjadi alasan perlunya tahap *pre-processing* dataset sebelum menerapkan algoritma tertentu suatu penerapan model. Teknik yang biasa digunakan untuk mengubah semua variabel nominal ke dalam bentuk angka adalah dengan menjadikan sebagai *dummy (binary)*[22]. Variabel *dummy* merupakan variabel *independen* yang bernilai 0 atau 1. Dengan menggunakan teknik ini dari fitur *nominal* yang memiliki *n* nilai yang berbeda *n* dapat menghasilkan fitur *dummy* baru [22].

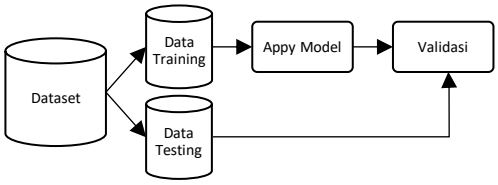
Sebagai contoh suatu kolom variabel prediktor (*independen*) dalam dataset adalah "jenis kelamin" yang berisi nilai *pria dan wanita* dengan tipe data nominal. Dengan *variabel dummy,* variabel ini menambahkan kolom baru yang tadinya "jenis kelamin" menjadi "jenis_kelamin_pria" dan "jenis_kelamin_wanita" dengan nilai 0 atau 1 mewakili nilai asli setiap barisnya. Jika suatu baris data asli berisi nilai "pria" maka nilai dari kolom/variabel *dummy* "jenis_kelamin_pria" adalah 1

dan secara otomatis kolom atau variabel "jenis_kelamin_wanita" adalah 0.

Dalam model regresi logistik, mengubah semua variabel independen sebagai *variabel dummy* memungkinkan kemudahan menghitung dan memahami setiap variabel prediktor karena lebih konsisten dengan nilai biner dan memudahkan dalam pengambilan keputusan serta meningkatkan stabilitas dan signifikansi dari koefisien [24].

D. Pembagian Data

Pada saat menerapkan model pada prediksi atau klasifikasi, suatu dataset pada umumnya dibagi menjadi 2 bagian yaitu data *training* (data pembentuk model) dan data *testing* (data penguji model) lihat **Gambar 2**. Untuk perolehan data *testing* menurut [18] dibedakan menjadi internal dan eksternal. Internal merupakan model dikembangkan dengan beberapa bagian dataset sebagai data *training* dan divalidasi dengan dataset sisanya. Atau dengan kata lain ada sebgaian dataset dijadikan *training* dan sisanya sebagai *testing*.



Gambar 2. Ilustrasi Pembagian Data

Metode yang paling banyak digunakan untuk melakukan validasi internal yang baik adalah *data-spliting, repeated data-splitting, jackknife technique* dan *bootstrapping*. Perolehan data *testing* eksternal adalah pengujian validitas dengan kumpulan dataset baru dari populasi yang sama. Untuk mendapatkan kumpulan data baru sebagai data *testing* diperlukan pemeriksaan model dalam konteks yang berbeda [18].

Nilai akurasi hasil penerapan model prediksi terdapat suatu bias sesuai dengan ukuran dataset yang digunakan [25]. Bias cenderung relatif besar jika ukuran data *training* kecil namun ketika ukuran data meningkat bias menurun hingga tidak lagi mendominasi. Hal ini karena keakuratan prediksi meningkat karena ukuran dataset *training* meningkat serta mendekati akurasi maksimum [25]. Ketika data yang dijadikan data *training* dengan persentase 40% - 80% dan sisanya dijadikan data *testing* maka akan menghasilkan proporsi untuk menguji akurasi yang optimal. Persentase tersebut bisa juga menggunakan ukuran 2/3 sebagai data *training* dan 1/3 sebagai data *testing* untuk mengukur akurasi model [25].

E. Pengujian Model

Setelah model regresi logistik dibangun, sehingga perlu untuk diuji tingkat akurasi prediksinya [26]. Pengukuran tingkat keakuratan prediksi dapat dilihat dari nilai akurasi model yang diterapkan. Semakin tinggi akurasi berarti semakin baik performa model klasifikasi/prediksi [27]. *Confusion Matrix* memberikan gambaran tentang kesesuaian antara data hasil prediksi dan data asli (data *training*) [28]. Atau dengan kata lain pada Tabel 2 juga digunakan untuk menguji nilai keluaran variabel hasil (*dependen*) dan variabel prediktor (*independen*) [18]. Informasi yang digambarkan pada *Confusion Matrix* berupa tabulasi silang tentang label data asli pada baris matriks dan kelas data hasil prediksi pada kolom [27].

Secara umum, *Confusion Matrix* ditunjukkan pada **Tabel 2** dimana terdiri dari:

(i) True Positives (TP) – label hasil prediksi adalah positif & label aslinya adalah positif,
(ii) False Positives (FP) – label hasil prediksi adalah negatif & label aslinya adalah positif
(iii) True Negatives (TN) – label hasil prediksi adalah negatif & label aslinya adalah negatif,
(iv) False Negatives (FN) label hasil prediksi adalah positif & label aslinya adalah negatif.

Tabel 2. Confusion Matrix

Nilai Asli	Nilai Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Dari Tabel *Confusion Matrix* dapat diperoleh beberapa nilai yang memrepresentasikan kulaitas dari penerapan model prediksi dengan menggunakan beberapa rumus rumus [8], [14]. *Accuracy* adalah nilai benar yang diperoleh dari semua proses prediksi [8] dihitung berdasarkan **Persamaan 4**.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad \textbf{(4)}$$

Sensitivity menentukan kinerja prediksi atau uji diagnostik dalam memprediksi nilai positif dengan benar di antara semua sampel positif yang tersedia selama pengujian dimana dihitung berdasarkan **Persamaan 5**.

$$\text{Sensitivity} = \text{TPRate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \textbf{(5)}$$

Specificity (*True Negative Rate - TNR*) digunakan untuk mengukur hasil negatif yang diidentifikasi dengan benar [8] dimana dihitung berdasarkan **Persamaan 6**. *Precision* (nilai prediksi positif) merupakan ukuran positif sebenarnya dibandingkan dengan semua bagian positif [8] dimana dihitung berdasarkan **Persamaan 7**.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \textbf{(6)}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \textbf{(7)}$$

Selain itu untuk mengetahui dan menguji tingkat kebenaran model dalam melakukan prediksi adalah dengan kurva *ROC (Receiver Operating Characteristic)*. *ROC* merupakan ukuran visual lebih sering dikenal sebagai kurva ukuran bawah atau *AUC* (*Area Under Curve)* [26]. Kurva ROC didasarkan pada ukuran dari nilai *true positive rate* dan *false positive rate* [26]. AUC memiliki nilai bervariasi dari 0,5 (tidak ada kemampuan prediktif) hingga 1,0 (kemampuan memprediksi yang sempurna). AUC yang lebih besar menunjukkan kemampuan model prediksi yang lebih baik [18].

FPRate digunakan untuk menentukan berapa banyak hasil positif yang salah, yang sebenarnya negative ada di antara semua sampel negatif yang ada selama pengujian [29] dimana dihitung berdasarkan **Persamaan 8**. Visualisasi kurva diperoleh dengan memplot *TPRate* atau $Sensitivity$ pada sumbu y dan *FPRate* pada sumbu x [26]. Untuk jarak nilai dan kategori prediksi pada AUC seperti pada **Tabel 3**.

$$\text{FPrate} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (8)$$

III. HASIL DAN PEMBAHASAN

Pada bab ini menjabarkan hasil penelitian yang telah dilakukan beserta pembahasannya. Pembahasan dilakukan sesuai dengan alur penelitian yang dijelaskan pada bagian metodologi.

Tabel 3. Nilai AUC [14][16]

Nilai AUC	Keterangan
0.90 – 1.0	Prediksi Paling Baik
0.80 – 0.90	Prediksi Baik
0.70 – 0.80	Prediksi rata-rata / sama
0.60 – 0.70	Prediksi rendah
< 0.60	Prediksi gagal

A. Persiapan Dataset

Persiapan data dimulai dengan mengunduh dataset dari [8], kebanyakan tipe data yang ada pada dataset tersebut adalah *polynominal* (Tabel 1). Pada penelitian ini yang dijadikan sebagai variabel hasil (*dependen*) adalah *Trust Score* (TS) dengan tipe data *binominal* (*trustworthy*, *untrustworthy*). Dan yang dijadikan variabel prediktor (*independen*) adalah *Counting Trust* (CT), *Counting Un-trust* (CU), *Last Time* (LT), *Transactions Context* (TC) berupa tipe data polinominal sehingga harus diubah dulu dalam bentuk angka/numerik.

B. Pengubahan Nominal ke Angka

Dataset pada Tabel 1 tersebut terdiri dari 322 baris data dan 4 atribut (*polynominal*) dangan 1 label kelas (*binominal*) bernilai (*trustworthy*, *untrustworthy*). Dari data bertipe nominal tersebut kemudian diubah menjadi data bertipe angka dengan menggunakan teknik variabel *dummy*. Pengubahan dataset awal pada ditunjukkan pada **Tabel 4** dimana diubah menjadi dataset *dummy* yang ditunjukkan pada **Tabel 5**.

Tabel 4. Dataset Awal

No	*CT*	*CU*	*LT*	*TC*	*TS*
1	CT_range_4	CU_range_1	LT_range_4	sport	untrustworthy
2	CT_range_4	CU_range_1	LT_range_4	sport	untrustworthy
3	CT_range_1	CU_range_4	LT_range_4	sport	trustworthy
: :	: :	: :	: :	: :	: :

Tabel 5. Dataset Dummy

No	*CT_range_4*	*CT_range_1*	*CT_range_2*	*CU_range_1*	*CU_range_4*	*CU_range_3*	*LT_range_4*	*LT_range_1*	*LT_range_4*	*TC_sport*	*TC_games*	*TC_ECommerce*	*TS*
1	0	0	1	1	0	0	1	0	0	1	0	0	untrustworthy
2	0	0	1	1	0	0	1	0	0	1	0	0	untrustworthy
3	0	1	0	0	0	1	1	0	0	1	0	0	trustworthy
: :	: :	: :	: :	: :	: :	: :	: :	: :	: :	: :	: :	: :	: :

C. Pembagian Data

Pada pembahasan sebelumnya untuk melakukan penerapan model prediksi menggunakan regresi logistik dan menguji nilai kebenaran prediksi yang dihasilkan maka dataset terlebih dahulu dibagi menjadi 2 bagian yaitu data *training* yaitu untuk membangun model, dan data *testing* untuk menguji kebenaran model yang dibuat. Pada penelitian ini menggunakan *split data* dengan persentase berkisar antara 40% - 80% untuk menghasilkan tingkat akurasi yang tinggi [25]. Berikut ini persentase pembagian data yang digunakan:

Pada **Tabel 6** terdapat kolom *kode pembagian* data untuk mempermudah saat membandingkan nilai akurasi prediksi. Model regresi logistik berganda dilakukan beberapa kali sesuai dengan banyaknya kode pembagian data. Hal ini dilakukan untuk menghasilkan nilai kebenaran prediksi dan koefisien berbeda jika diterapkan pada jumlah data *training* dan data *testing* yang berbeda. Dari nilai yang berbeda itu dapat diketahui pada pembagian data yang menghasilkan tingkat kebenaran prediksi paling tinggi pada penerapan regresi logistik berganda.

Tabel 6. Persentase Pembagian Dataset

Kode Pembagian	Data training (%)	Data testing (%)
A	60	40
B	65	35
C	70	30
D	75	25
E	80	20

D. Penerapan Model

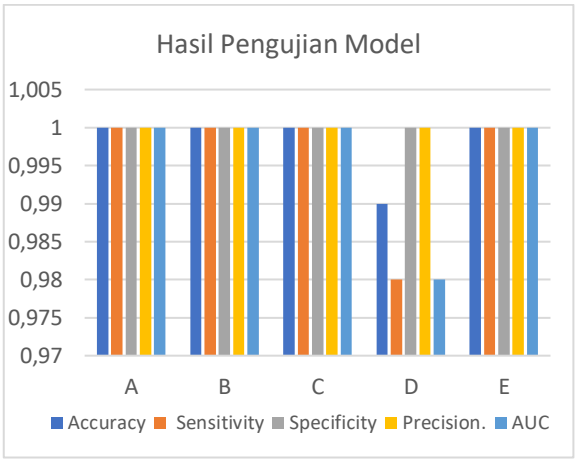
Penerapan model yang dilakukan dengan regresi logistik berganda terhadap pembagian data *training* yang berbeda. Oleh karena itu koefisien dan *p-value* yang dihasilkan juga berbeda-beda dimana ditunjukkan pada **Tabel 7**. Pengujian parsial pada Regresi Logistik Berganda dilakukan untuk melihat apakah tiap parameter layak digunakan dalam model [17]. Hipotesis yang digunakan adalah sebagai berikut :

$H_0 : \delta_i = 0$ (variabel δ_i tidak layak dalam model)
$H_1 : \delta_i \neq 0$ (variabel δ_i layak dalam model)

Pengujian hipotesisnya adalah H_0 ditolak jika *p-value* kurang dari tingkat signifikan 0.05 atau 5%. Kesimpulan yang didapatkan adalah semua variabel prediktor (*independen*) signifikan terhadap model yang dibuat dengan walaupun menggunakan pembagian data yang berbeda. Dari kesimpulan tersebut, dapat diputuskan bahwa tidak ada penghapusan atau pengurangan variabel prediktor (*independen*) dari dataset sehingga layak untuk dijadikan model prediksi.

E. Pengujian Model

Item penilaian yang dihasilkan oleh *Confusion Matrix dan kurva ROC* ditunjukkan pada **Tabel 8**. Dari beberapa pembagian data *training* sebagai pembangun model dan data *testing* sebagai penguji model yang dibangun menghasilkan nilai *Accuracy*, *Sensitivity*, *Specificity*, *Precision* dan *AUC* yang bervariasi. Dengan menggunakan variabel *dummy* menghasilkan nilai keluaran model prediksi yang cukup baik dimana ditunjukkan pada **Gambar 3**.



Gambar 3. Hasil Pengujian

Tabel 7. Hasil Penerapan Regresi Logistik

Variabel	Split data									
	A		B		C		D		E	
	Coef	*p-Value*	*Coef*	*p-Value*	*Coef*	*p-Value*	*Coef*	*p-Value*	*Coef*	*p-Value*
CT_range_4	-32.896	0.927	-32.759	0.927	-32.781	0.9219	-32.833	0.919	-32.792	0.917
CT_range_1	-28.727	0.956	-28.436	0.954	-28.423	0.9517	-28.475	0.950	-28.335	0.948
CT_range_2	-2.613	0.995	-2.586	0.995	-2.619	0.9942	-2.356	0.994	-2.340	0.994
CU_range_1	-28.727	0.981	-30.554	0.979	-30.242	0.9738	-30.081	0.974	-29.754	0.973
CU_range_4	-0.000	1.000	-2.117	0.998	-1.819	0.9981	-1.606	0.998	-1.420	0.998
CU_range_3	-0.000	1.000	-2.117	0.999	-1.819	0.9983	-1.606	0.999	-1.420	0.999
LT_range_4	59.392	0.924	56.133	0.912	56.407	0.9097	57.012	0.908	57.101	0.903
LT_range_1	59.392	0.872	59.094	0.864	59.075	0.8583	59.470	0.856	59.312	0.845
LT_range_3	2.946	0.995	2.965	0.995	2.888	0.9946	2.887	0.995	2.792	0.995
sport	-28.727	0.965	-25.475	0.963	-25.754	0.9616	-26.017	0.960	-26.124	0.958
game	60.615	0.885	60.451	0.877	60.546	0.8737	60.940	0.870	60.977	0.862
ECommerce	-63.816	0.922	-62.681	0.917	-62.663	0.9146	-62.711	0.912	-62.632	0.909

Tabel 8. Nilai keluaran Confusion Matrix dan ROC

Hasil	Split data				
	A	**B**	**C**	**D**	**E**
Accuracy	1.00	1.00	1.00	0.99	1.00
Sensitivity	1.00	1.00	1.00	0.98	1.00
Specificity	1.00	1.00	1.00	1.00	1.00
Precision	1.00	1.00	1.00	1.00	1.00
AUC	1.00	1.00	1.00	0.98	1.00

Terlihat bahwa keakuratan model didapatkan akurasi rata-rata sebesar 0.998 atau 99,8%. Kemudian juga terdapat *Specificity* atau *true positive* rata-rata sebesar 0.996 atau 99,6% dan *Sensitivity* atau *true negative* rata-rata sebesar 1 atau 100%. Nilai *Precision* atau *recall* didapatkan rata-rata sebesar 1 atau 100%. Yang terakhir adalah nilai *AUC* yang didapatkan dari kurva *ROC* rata-rata sebesar 0.996 atau 99,6% yang berarti proses prediksi telah berhasil dilakukan.

Variabel prediktor (*independen*) yang diterima dalam membangun model regresi logistik berganda menghasilkan nilai signifikan terhadap model yang dikembangkan.

IV. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa model regresi logistik berganda dapat digunakan secara efektif untuk memprediksi kepercayaan pengguna internet dalam paradigma *pervasive computing* dengan terlebih dahulu melalui tahap *preprosessing* data serta pembagian data *training* dan data *testing* yang sesuai. Dari beberapa pembagian data, model memprediksi hampir mencapai tingkat akurasi 100% atau 1.00.

Perubahan dataset menjadi dataset *dummy* pada proses *preprosessing* meningkatkan nilai hasil pengujian prediksi paling baik yang dibuktikan dengan nilai AUC 0.90 – 1.0. Pembagian dataset yang memiliki nilai prediksi rendah dengan nilai *Sensitivity* dan AUC yaitu 75% sebagai data training dan 25% sebagai data testing dari dataset awal 322 baris.

DAFTAR PUSTAKA

[1] M. Weiser, "The computer for the 21st century," in *Scientific American*, 1991, pp. 94–101.

[2] C. Setiawan, "Pervasive Computing / Ubiquitous Computing (Ubicomp) di Indonesia."

[3] M.-Y. Cho and T. Thom Hoang, "Feature Selection and Parameters Optimization of SVM Using Particle Swarm Optimization for Fault Classification in Power Distribution Systems," 2017.

[4] S. Kurkovsky, "Pervasive computing: Past, present and future," in *2007 ITI 5th International Conference on Information and Communications Technology*, 2007.

[5] J. Ye, S. Dobson, and P. Nixon, "An Overview of Pervasive Computing Systems," in *Ambient Intelligence with Microsystems*, 2009.

[6] G. Dangelo, S. Rampone, and F. Palmieri, "An Artificial Intelligence-Based Trust Model for Pervasive Computing," in *Proceedings - 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2015*, 2015, pp. 701–706.

[7] G. D'angelo, S. Rampone, F. Palmieri, and G. D. ' Angelo, "Developing a Trust Model for Pervasive Computing Based on Apriori Association Rules Learning and Bayesian Classification," *Soft Comput. - A Fusion Found. Methodol. Appl.*, vol. 21, no. 21, pp. 6297–6315, 2017.

[8] G. Dangelo, "Dishonest Internet users Dataset Data Set," *UCI Machine Learning Repository*, 2018. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Dishonest+Internet+users+Dataset.

[9] V. Sampath, A. Flagel, and C. Figueroa, "A Logistic Regression Model To Predict Freshmen Enrollments."

[10] M. Zaidi and A. Amirat, "Forecasting Stock Market Trends By Logistic Regression And Neural Networks Evidence From Ksa Stock Market," *Int. J. Econ. Commer. Manag.*, vol. IV, no. 6, pp. 220–234, 2016.

[11] N. Srivastava, "A logistic regression model for predicting the occurrence of intense geomagnetic storms," 2005.

[12] D. Regresi, L. Biner, I. Ketut, P. Suniantara, and M. Rusli, "Klasifikasi Waktu Kelulusanmahasiswa Stikom Bali Menggunakan Chaid Regression Trees Dan Regresi Logistik Biner," vol. 5, no. 1, 2017.

[13] L. Mary Gladence, M. Karthi, and V. Maria Anu, "A statistical comparison of logistic regression and different bayes classification

methods for machine learning," *ARPN J. Eng. Appl. Sci.*, 2015.

[14] A. Yumalia, R. E. Indrajit, and M. A. Ri, "Penerapan Konsep Business Intelligence Untuk Percepatan Penyelesaian Perkara Pada PANMUD Perdata Khusus Mahkamah Agung RI," vol. 1, no. 2, 2017.

[15] D. H. Ismunarti, "Regresi Logistik Binomial , Model untuk Toksisitas Logam Berat Timbal Pb terhadap Larva Udang Vannamae," *Bul. Oseanografi Mar. Oktober*, vol. 1, pp. 47–52, 2012.

[16] A. Ilham, "Komparasi Algoritma Klasifikasi Dengan Pendekatan Level Data Untuk Menangani Data Kelas Tidak Seimbang," *J. Ilm. Ilmu Komput.*, vol. 3, no. 1, pp. 2442–4512, 2017.

[17] A. N. Manthovani, "Analisis Perbandingan Klasifikasi Metode Regresi Logistik Biner Dan Random Forest Pada Big Data," 2018.

[18] H. A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *J. Korean Acad. Nurs.*, 2013.

[19] A. Rohmadi, "Penerapan Metode Regresi Logistik Pada Aplikasi Pemilihan Organisasi Mahasiswa," 2013.

[20] S. L. David W. Hosmer Jr., *Applied logistic regression*. 1998.

[21] A. T. Basuki, "Regresi Logistik biner."

[22] E. Zdravevski, P. Lameski, and A. Kulakov, "Advanced Transformations for Nominal and Categorical Data into Numeric Data in Supervised Learning Problems," *10th Conf. Informatics Inf. Technol.*, vol. 7, no. Ciit, pp. 142–146, 2013.

[23] M. te Grotenhuis and P. Thijs, "Dummy variables and their interactions in regression analysis: examples from research on body mass index," 2015.

[24] S. Garavaglia, A. Sharma, and M. Hill, "A Smart Guide To Dummy Variables : Four Applications and a Macro."

[25] K. K. Dobbin and R. M. Simon, "Optimally splitting cases for training and testing high dimensional classifiers," *BMC Med. Genomics*, vol. 4, 2011.

[26] N. Šarlija, A. Bilandžić, and M. Stanić, "Logistic regression modelling: procedures and pitfalls in developing and interpreting prediction models," *Croat. Oper. Res. Rev.*, vol. 8, pp. 631–652, 2017.

[27] A. Salim, "Pengoptimalan Naïve Bayes Dan Regresi Logistik Menggunakan Algoritma Genetika Untuk Data Klasifikasi," 2017.

[28] Y. Jiang, "Using Logistic Regression Model to Predict the Success of Bank Telemarketing," *Int. J. Data Sci. Technol.*, vol. 4, no. 1, pp. 35–41, 2018.

[29] A. I. Abdelrahman, "Applying Logistic Regression Model to The Second Primary Cancer Data," *InterStat J.*, 2010.