

ABSTRAK

Latar Belakang: Kanker payudara merupakan salah satu kanker yang paling banyak ditemui pada wanita. Permasalahan kanker payudara saat ini yaitu tingginya angka kematian dan rendahnya kesadaran masyarakat dalam mengenali risiko dan mengidentifikasi gejala yang dimiliki secara dini, sehingga kanker payudara di Indonesia lebih banyak teridentifikasi pada stadium lanjut. Tantangan yang ada dalam menjawab permasalahan tersebut yaitu belum adanya model prediksi yang mengidentifikasi risiko kanker payudara pada wanita di Indonesia. Kondisi inilah yang menyebabkan perlu dilakukan kalkulasi faktor risiko kanker payudara untuk membantu peningkatan kesadaran masyarakat dalam mengenali risiko kanker payudara di Indonesia.

Metode: Penelitian ini dilakukan dengan menggunakan desain *case control study*. Lokasi penelitian di RSUP Dr. M Djamil Padang, RSUP Sardjito Yogyakarta dan RS Kanker Dharmas Jakarta dari bulan Juli 2018-Juli 2019. Jumlah sampel pada penelitian ini yaitu 1.000 orang kelompok kasus (kanker payudara pada wanita) dan 1.000 orang kelompok kontrol (bukan penderita kanker payudara) yang dimatchingkan usia dan jenis kelamin. Teknik pengambilan sampel pada penelitian ini yaitu *convenience sampling*. Data dikumpulkan melalui telaah rekam medis dan pengumpulan data secara primer dengan menggunakan kuesioner penelitian. Data dianalisis secara bivariat dengan menggunakan uji *chi-square* dan dilanjutkan dengan algoritma *machine learning naive bayes (NB)*, *decision tree (DT)*, *artificial neural network (ANN)*, *support vector machine (SVM)* dan *logistic regression (LR)*. Penentuan pemilihan algoritma dengan membandingkan nilai akurasi, *true positive rate*, *false positive rate* dan *area under curve (AUC)* yang tertinggi. Analisis data menggunakan program R software 4.0.2 *Waikato Environment for Knowledge Analysis (WEKA) version 3.6.4* dan *STATA*.

Hasil: Lima algoritma yang digunakan dalam memprediksi risiko kanker payudara di Indonesia yaitu *NB*, *DT*, *ANN*, *SVM* dan *LR*, yang menjadi algoritma terbaik dalam memprediksi risiko kanker payudara di Indonesia adalah *LR*. Tujuh belas variabel prediktor pada tahap awal diseleksi dalam memprediksi risiko kanker payudara, hasil seleksi variabel kandidat tersebut mengidentifikasi 10 variabel yang dapat digunakan untuk konstruksi model kalkulasi faktor risiko kanker payudara di Indonesia yaitu usia menopause, usia pertama kali hamil, riwayat keluarga tingkat pertama dan kedua menderita kanker payudara, penggunaan kontrasepsi oral, riwayat merokok, overweight, obesitas, diet tinggi lemak, diet tinggi kalori dan aktifitas fisik. Akumulasi total skor yang diperoleh dari variabel prediktor berada diantara skor 0 (apabila tidak memiliki risiko) hingga skor 30 (apabila memiliki semua risiko). Batasan skor yang membedakan subjek berisiko rendah kanker payudara (apabila jumlah skor yang diperoleh < 7), berisiko sedang kanker payudara (apabila jumlah skor yang diperoleh 7-8) dan berisiko tinggi kanker payudara (apabila jumlah skor yang diperoleh ≥ 9) dengan akurasi 79,9%, sensitivitas 76,9% dan spesifisitas 70,4%.

Kesimpulan: Model kalkulasi faktor risiko kanker payudara ini dikategorikan cukup baik dalam mengklasifikasikan risiko kanker payudara di Indonesia.

Kata Kunci: Kanker Payudara, Faktor Risiko, Prediksi, Klasifikasi, *Machine Learning*

ABSTRACT

Background: Breast cancer is the most common cancer among women. The current breast cancer problems are the high mortality rate and low perception in early recognizing the risk of breast cancer, this condition leads to the identification of breast cancer at an advanced stage. Therefore, it is necessary to calculate the risk factors for breast cancer risk to help increasing public awareness in assessing the risk of breast cancer in Indonesia.

Methods: This research was quantitative which was conducted using a case control study design. Data were collected in Dr. M. Djamil General Hospital Padang, Sardjito General Hospital Yogyakarta and Dharmas Cancer Hospital Jakarta from July 2018-July 2019. The number of samples in this study were 1,000 cases groups (breast cancer in women) and 1,000 control groups (not breast cancer patients) matching by aged and sex. The sampling technique in this study was convenience sampling. Data were extracted from medical records and primary data collection used a questionnaire. Chi-square test used for bivariate analysis and risk factors calculation were used machine learning algorithms naive bayes (NB), decision tree (DT), artificial neural network (ANN), support vector machine (SVM) and logistic regression. Determination of algorithm selection by comparing the highest accuracy, true positive rate, false positive rate and area under curve (AUC). R software 4.0.2 *Waikato Environment for Knowledge Analysis (WEKA) version 3.6.4* and *STATA* program were used to process the data.

Results: Five algorithms are used in predicting breast cancer risk in Indonesia namely NB, DT, ANN, SVM and LR. LR algorithm is the best algorithm to calculate breast cancer risk with better accuracy than other machine learning algorithms. Seventeen predictor variables were selected in the early stages in predicting breast cancer risk, the results of the selection of these candidate variables identified 10 variables that can be used for calculating risk factors for breast cancer risk in Indonesia is based on predictors of menopause age, first age of pregnancy, first and second-degree family history of breast cancer, use of oral contraceptives, history of smoking, overweight, obesity, high-fat diet, high-calorie diets and physical activity. The accumulated total score obtained from the predictor variables is between 0 (if there is no risk) to 30 (if it has all risks). The cut-off point that distinguish subjects with low risk of breast cancer (if the total score obtained <7), moderate risk of breast cancer (if the total score obtained 7-8) and high risk of breast cancer (if the total score obtained ≥ 9) with an accuracy of 79.9%, sensitivity 76.9% and specificity 70.4%.

Conclusion: This breast cancer risk factor calculation can be categorized quite well in classifying breast cancer risk in Indonesia.

Keywords: Breast Cancer, Risk Factors, Prediction, Classification, *Machine Learning*