Penilaian Risiko Induksi Kehamilan Hipertensi Menggunakan Pendekatan Machine Learning

Sirinat Wanriko
Gambar Laboratorium Informasi dan Kecerdasan, Jurusan Teknik Komputer, Fakultas Teknik, Mahidol University
Nakhon Pathom, Thailand
sirinat.wan@student.mahidol.ac.th

Narit Hnoohom*
Image Information and Intelligence Laboratory, Department of Computer Engineering, Faculty of Engineering, Mahidol University Nakhon Pathom, Thailand
Corresponding author: narit.hno@mahidol.ac.th

Konlakorn Wongpatikaseree
Jurusan Teknik Komputer, Fakultas Teknik, Mahidol University Nakhon Pathom, Thailand
konlakorn.won@mahidol.ac.th

Anuchit Jitpattanakul
Pusat Penelitian Inovasi Cerdas dan Nonlinier, Departemen Matematika, Fakultas Sains Terapan, Universitas Teknologi Raja Mongkut Bangkok Utara
Bangkok, Thailand
anuchit.j@sci.kmutnb.ac.th

Olarik Musigavong
Departemen Obstetri dan Ginekologi (Endokrinologi Reproduksi dan Infertilitas)
Rumah Sakit Chaophraya Abhaibhubejhr
Prachin Buri, Thailand
dr.olarik@gmail.com

***Abstrak*— Penelitian ini bertujuan untuk mengembangkan model prediktif penilaian risiko hipertensi akibat kehamilan dengan menggunakan pendekatan pembelajaran mesin. Hipertensi yang diinduksi kehamilan merupakan komplikasi yang berdampak serius pada ibu hamil dan janin. Ini adalah tiga penyebab kematian tertinggi di dunia pada wanita hamil [1]. Saat ini, penyebab pasti hipertensi akibat kehamilan tidak diketahui dan oleh karena itu tidak dapat dicegah. Deteksi dini dan pengobatan yang diterima dapat mengurangi kematian dan bahayanya. Dataset publik Logan (2020) digunakan dalam penelitian ini [2]. Kumpulan data dikumpulkan dari studi kasus-kontrol pada 83 kasus pre-eklampsia dan lima kasus eklampsia di antara 352 wanita hamil yang melahirkan di rumah sakit daerah di Nairobi, Kenya. Menurut dataset, 75 persen wanita hamil itu sehat. Hanya 25 persen ibu hamil yang mengalami pre-eklampsia dan eklampsia. Dengan demikian, ini akan menghasilkan masalah klasifikasi yang tidak seimbang ketika salah satu dari dua kelas memiliki lebih banyak data daripada kelas lainnya. Dengan demikian, masalah ini diselesaikan dengan Teknik Pengambilan Sampel Minoritas Sintetis (SMOTE). Penilaian risiko hipertensi akibat kehamilan dilakukan pada tujuh algoritma pembelajaran mesin, yaitu regresi logistik (LR), K-nearest neighbor (KNN), pohon keputusan (DT), hutan acak (RF), jaringan saraf perceptron multilayer (MLP), mendukung mesin vektor (SVM), dan naif Bayes (NB). Pada hasil percobaan, RF memiliki akurasi tertinggi sebesar 89,62 persen dibandingkan dengan algoritma pembelajaran mesin lainnya.**

***Kata kunci** -- Hipertensi yang diinduksi kehamilan, pembelajaran mesin, klasifikasi tidak seimbang, minoritas sintetik atas teknik pengambilan sampel*

I. Sebagai pengantar

Hipertensi yang diinduksi kehamilan merupakan komplikasi yang berdampak serius pada ibu hamil dan janin [2]. Ini sebagian besar terjadi setelah minggu kedua puluh kehamilan. Selain kematian, ada risiko komplikasi serius, seperti, plasenta prematur, banjir paru, pembekuan darah yang tidak normal, kebutaan sementara atau permanen, dan pendarahan di otak. Saat ini, penyebab pasti hipertensi akibat kehamilan tidak diketahui dan oleh karena itu tidak dapat dicegah. Deteksi dini dan pengobatan yang diterima dapat mengurangi keparahan dan bahayanya. Data dari Organisasi Kesehatan Dunia (WHO) pada tahun 2010 ditemukan bahwa 99 persen ibu hamil yang meninggal tinggal di pedesaan dan miskin [3]. Selain itu, di pedesaan Thailand, terdapat kekurangan dokter dan tenaga medis. Akibatnya, beberapa instansi telah mendirikan proyek untuk mengatasi masalah pendistribusian layanan medis ke pedesaan, namun masalah ini tetap saja terjadi. Menurut statistik Thailand pada tahun 2017, Dewan Medis menemukan bahwa rata-rata satu dokter harus merawat 1.143 pasien per tahun [4-5]. Dengan demikian, tujuan akhir dari penelitian ini adalah untuk mengembangkan model prediktif penilaian risiko hipertensi akibat kehamilan menggunakan pendekatan pembelajaran mesin. Tujuannya adalah untuk mendukung dokter dalam penilaian awal pasien, perencanaan perawatan, dan perawatan pemulihan.

Sisa kertas disusun sebagai berikut. Pertama, pengenalan singkat diberikan. Pekerjaan terkait dijelaskan dalam Bagian II. Bagian III berisi ikhtisar metode yang diusulkan. Hasil percobaan dimasukkan dalam Bagian IV. Akhirnya, dalam Bagian V, kesimpulan dibahas.

II. Related Work

Saat ini ada beberapa penelitian yang menggunakan pembelajaran mesin untuk membuat model prediktif risiko hipertensi akibat kehamilan, seperti eklampsia dan preeklampsia. Tahir dkk. [6] mengusulkan prediksi risiko tingkat preeklampsia pada ibu hamil selama proses kehamilan menggunakan algoritma neural network (NN) dan deep learning (DL). Jumlah atribut dikurangi dari 17 menjadi sembilan menggunakan pengoptimalan segerombolan partikel. DL memberikan akurasi paling tinggi dengan 95,68 persen. Nikolaides et al. [7] mengembangkan indikator tahap awal untuk risiko pre-eklampsia. Dataset disusun menggunakan 6.838 kasus ibu hamil di Inggris yang melibatkan 24 variabel. Dataset hanya berisi 116 kasus ibu hamil dengan pre-eklampsia. Jaringan saraf multi-slab memberikan hasil paling akurat sebesar 93,8 persen. Selain itu, Tahir et al. [8] mengusulkan deteksi pre-eklampsia dengan menggunakan jaringan saraf dibandingkan dengan algoritma lainnya. Dataset preeklampsia diambil dari Rumah Sakit Umum Haji Surabaya, Indonesia. Algoritma jaringan saraf dan validasi LOO memberikan akurasi paling tinggi

2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering | 978-1-6654-1569-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ECTIDAMTNCON51128.2021.9425764

with 96.6 percent. Likewise, Lacmaq et al. [9] presented a novel preeclampsia prediction model focusing on the convergence of multiple models. They analyzed that Bayes' theorem could be used to integrate multiple models. The integrated model provided the most accuracy, where 81 percent was accurately identified at 20 weeks of gestation.

In addition to the above research, machine learning has been used to create a predictive model for the risk of other diseases. Khalilia et al. [10] proposed the Random Forest (RF) for predicting the risk of eight chronic diseases by using the database of the nationwide inpatient sample from samples of hospitals in the United States. The ensemble learning approach was used to solve the problem of the imbalanced data. The results from testing the accuracy was 88.79 percent. In addition, Pattekari and Parveen [11] proposed a prediction system for heart disease. They used mining techniques consisting of DT, Naive Bayes (NB), and NN. This system could intelligently answer complex questions in diagnosing heart disease and help medical practitioners. This further assisted in enhancing and reducing the treatment costs. Moreover, Akhil jabbar et al. [12] proposed the K-nearest neighbor (KNN) and genetic algorithms to analyze heart disease using six datasets from the UCI Repository (UCI Machine Learning Repository) and one dataset from various hospitals in Andhra Pradesh, India. To enhance the accuracy of the research, KNN was used to collect all cases and classify new patients using a genetic algorithm to calculate the similarities. The results showed that using both methods together gave 95.73 percent accuracy. Nayeem et al. [13] proposed a multilayer perceptron neural network (MLP) to predict heart disease, liver disease, and lung cancer using a feed-forward backpropagation neural network algorithm and MLP to distinguish between infected and uninfected individuals. They used the MIT-BIH Arrhythmia dataset, and the results of the research showed that the accuracy of the heart disease prediction was 82 percent, liver disease was 82 percent, and lung cancer was 91 percent, respectively.

II. The Proposed Method

In this section, the process was the development of the risk prediction model of pregnancy-induced hypertension as shown in Figure 1. The proposed method consisted of five processes: (1) pregnancy data, (2) imbalanced data, (3) data preparation, (4) feature selection, and (5) evaluate machine learning algorithms.

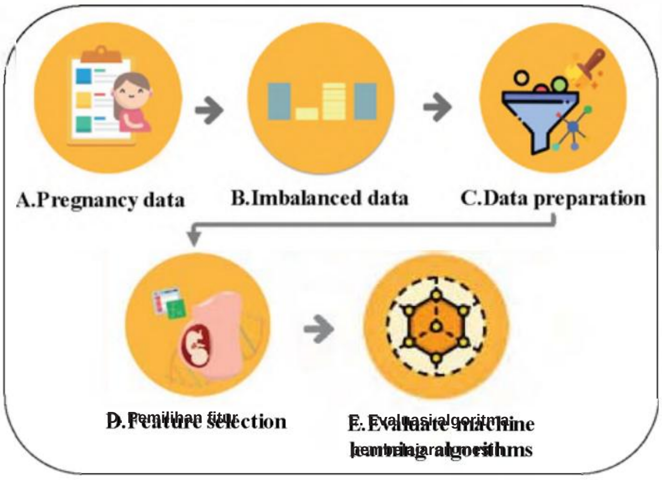
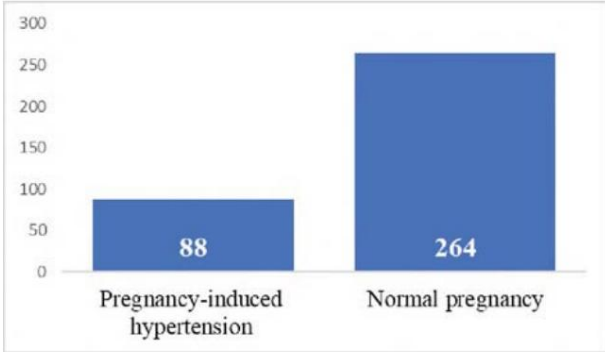


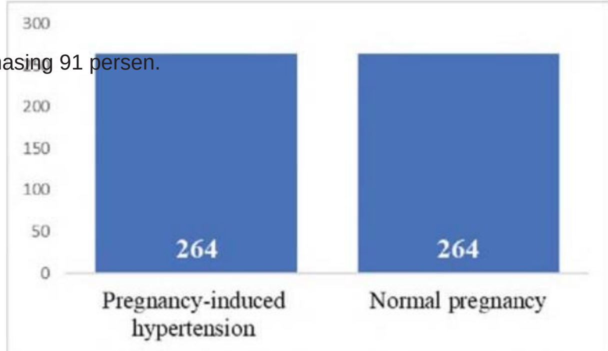
Figure 1. Proposed method development process.

A. Pregnancy data

The researcher explored the required public dataset relating to the risk of pregnancy-induced hypertension. The public dataset of Logan (2020) was used in this research [2]. The dataset was collected from a case-control study to determine the risk of 83 pre-eclampsia and eclampsia cases among 352 pregnant women delivering in county hospitals in Nairobi, Kenya. The researcher selected 17 attributes, which were matched to the attributes of the maternal and child handbook of Thailand, and this was used to record the mother and child health [14]. The 17 attributes comprised maternal age, age of the first pregnancy, diabetes family history, diabetes personal history, hypertension family history, hypertension personal history, first antenatal care visit, number of antenatal care visits, distance, twin pregnancy, caesarean section delivery, multiple pregnancy, gravidity, parity, province, residence, alcohol use, and tobacco use. The researcher then converted the data for appropriateness in the imbalanced data process.



(a) Imbalanced data.



(b) Balanced data.

Figure 2. Comparison of the imbalanced data and balanced data.

B. Imbalanced data

In the pregnancy data [13], 75 percent of pregnant women were healthy and only 25 percent were pre-eclampsia and eclampsia. The problem of the imbalanced classification arose when one of the two classes had more data than the other class, and this data affected the classification of the minority class. Typical classification was effective when each data class had a similar number, and the prediction was inclined to the majority class.

To compare the performance of the imbalanced data and balanced data approaches, the researcher performed the Synthetic Minority Over-sampling Technique (SMOTE) algorithm to the dataset [15]. The SMOTE works by utilizing

the KNN algorithm to construct synthetic data. The SMOTE began by selecting random data from the minority class after which the KNNs were created from the data. The random data and the randomly chosen KNN were then merged to construct synthetic data. This process was replicated until the minority class's proportion equaled that of the majority class.

The researcher plotted the class distribution to show the imbalanced data in the selected dataset as shown in Figure 2(a). Figure 2(b) shows the balanced data was handled with the SMOTE.

C. Data preparation

In the data preparation process, the researcher used data transformation in order to convert the data to better reveal the structure of the classification problem for the risk of pregnancy-induced hypertension. The Python program and scikit-learn library were utilized to develop the predictive model in this research. The 17 attributes were redistributed or rescaled, data tidying the MinMaxScaler, StandardScaler, and preprocessed using three preprocessing methods, which consisted of the MinMaxScaler, StandardScaler, and Normalizer.

D. Feature selection

Feature selection for machine learning is the method of choosing features that are relevant to each other to reduce the number of features to only the features necessary in order to develop a predictive model. Principal component analysis (PCA) is a data reduction technique that is commonly used in combination with linear algebra to minimize the dimensionality of a dataset by compressing a dataset of attributes while preserving most of the information in the dataset. PCA was used to extract the three principal components on the 17 attributes, imbalanced data, and balanced data.

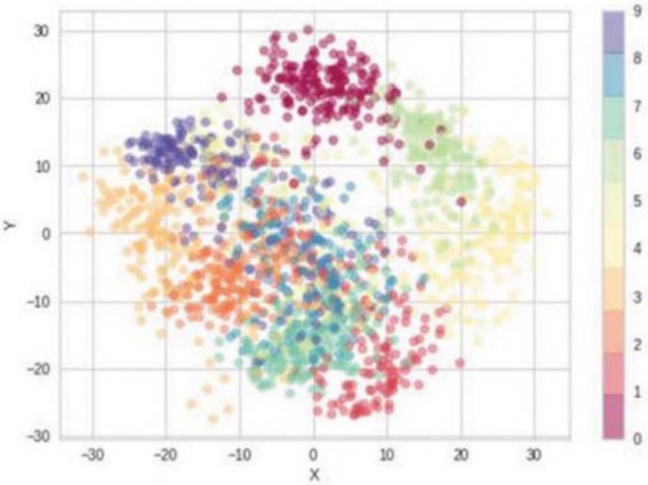


Fig. 3. Post-processing data with PCA.

E. Evaluate machine learning algorithms

To evaluate machine learning algorithms on the dataset, the predictive model was performed on seven machine learning algorithms, which were LR, KNN, DT, RF, MLP, SVM, and NB.

To estimate the performance of the machine learning algorithm on the unseen data, the researcher first split the dataset both the imbalanced data and balanced data into two parts: training and testing. For this case study, the researcher took a 77:33 ratio keeping 77 percent of the dataset for training and the remaining 33 percent for testing. The researcher trained a machine learning algorithm on the first part then evaluated the predictions on the test set against the expected results.

IV. Experimental Results

To obtain the predictive model of the risk assessment of pregnancy-induced hypertension, there were three main steps. Firstly, the three preprocessing methods were performed on the imbalanced data and balanced data using scikit-learn with the MinMaxScaler, StandardScaler, and Normalizer. Next, PCA was used to extract the three principal components on the imbalanced data and balanced data. Finally, the three principal components were performed to seven machine learning algorithms, which were logistic regression (LR), K-nearest neighbor (KNN), decision tree (DT), random forest (RF), multilayer perceptron neural network (MLP), support vector machines (SVM), and naive Bayes (NB). The evaluation results are shown in Tables I-VI.

TABLE I. Results of the Imbalanced Data with StandardScaler.

Machine Learning	Feature Selection
	PCA
Logistic Regression	69.01
K-Nearest Neighbor	64.79
Decision Tree	61.97
Random Forests	70.42
Multilayer Perceptron	<u>71.83</u>
Naive Bayes	28.17
Support Vector Machines	70.42

The results can be seen in Table I. The highest accuracy rate was 71.83 percent, shown by the underlined letter, which belonged to the MLP algorithm.

TABLE II. Results of the Imbalanced Data with MinMaxScaler.

Machine Learning	Feature Selection
	PCA
Logistic Regression	<u>71.83</u>
K-Nearest Neighbor	67.61
Decision Tree	60.56
Random Forests	<u>71.83</u>
Multilayer Perceptron	<u>71.83</u>
Naive Bayes	28.17
Support Vector Machines	<u>71.83</u>

As shown in Table II, the highest accuracy rate was 71.83 percent, shown by the underlined letter, which belonged to the LR, RF, MLP and SVM algorithms.

The results can be seen in Table III. The highest accuracy rate was 73.24 percent, shown by the underlined letter, which belonged to the KNN algorithm.

As shown in Table IV, the highest accuracy rate was 89.62 percent, shown by the underlined letter, which belonged to the RF algorithm.

TABEL III. HASIL DATA YANG TIDAK SEIMBANG DENGAN NORMALIZER.

Pembelajaran mesin	Seleksi Fitur
	PCA
Regresi Logistik	71.83
K-Tetangga Terdekat	73.24
Pohon Keputusan	59.15
Hutan Acak	67.61
Perceptron berlapis-lapis	71.83
Naif Bayes	28.17
Mendukung Mesin Vektor	71.83

TABEL IV. HASIL DATA BALANCED DENGAN STANDARDSCALER.

Pembelajaran mesin	Seleksi Fitur
	PCA
Regresi Logistik	69.81
K-Tetangga Terdekat	72.64
Pohon Keputusan	82.08
Hutan Acak	89.62
Perceptron berlapis-lapis	45.28
Naif Bayes	48.11
Mendukung Mesin Vektor	78.30

TABEL V. HASIL BALANCED DATA DENGAN MINMAXSCALER.

Pembelajaran mesin	Seleksi Fitur
	PCA
Regresi Logistik	66.04
K-Tetangga Terdekat	68.87
Pohon Keputusan	79.25
Hutan Acak	85.85
Perceptron berlapis-lapis	45.28
Naif Bayes	50.00
Mendukung Mesin Vektor	59.43

Hasilnya dapat dilihat pada Tabel VI. Tingkat akurasi tertinggi adalah 85,85 persen, ditunjukkan oleh huruf yang digarisbawahi, yang termasuk dalam algoritma RF.

TABEL VI. Hasil Data Seimbang Dengan Normalizer.

Pembelajaran mesin	Seleksi Fitur
	PCA
Regresi Logistik	47.17
K-Tetangga Terdekat	69.81
Pohon Keputusan	73.58
Hutan Acak	85.85
Perceptron berlapis-lapis	45.28
Naif Bayes	48.11
Mendukung Mesin Vektor	45.28

Seperti yang ditunjukkan pada Tabel VI, tingkat akurasi tertinggi adalah 85,85 persen, ditunjukkan oleh huruf yang digarisbawahi, yang termasuk dalam algoritma RF.

Menurut percobaan yang dilakukan dengan dataset, algoritma RF menghasilkan hasil terbaik berdasarkan data seimbang. Tingkat akurasi tertinggi adalah 89,62 persen.

V. Co clus io n

Makalah ini mempresentasikan klasifikasi risiko hipertensi yang diinduksi kehamilan menggunakan pendekatan pembelajaran mesin. Dataset publik Logan (2020) digunakan dalam penelitian ini. Dalam menangani data yang tidak seimbang, peneliti melakukan algoritma SMOTE pada dataset. Dalam proses penyiapan data, peneliti menggunakan transformasi data yang terdiri dari MinMaxScaler, StandardScaler dan Normalizer untuk mengkonversi data agar lebih mengungkap struktur data. masalah prediksi. PCA digunakan untuk mengekstrak tiga komponen utama pada data yang tidak seimbang dan data yang seimbang. Pada hasil percobaan, algoritma RF mencapai kinerja terbaik jika dibandingkan dengan algoritma pembelajaran mesin lainnya. Model prediksi tersebut menghasilkan tingkat akurasi tinggi 89,62 persen yang didasarkan pada data seimbang. Di masa depan, peneliti harus menerapkan model prediktif pada dataset hipertensi yang diinduksi kehamilan dari Rumah Sakit Chaophraya Abhaibhubejhr, Prachin Buri.

Akui ujungnya

Penelitian ini didukung oleh Departemen Teknik Komputer, Fakultas Teknik, Universitas Mahidol. Penelitian ini juga didukung oleh National Research Council of Thailand untuk proyek "Mom's buddy: A chatbot for pregnancy health information". Selanjutnya, penelitian ini didukung oleh Rumah Sakit Chaophraya Abhaibhubejhr tentang pengetahuan medis.

Referensi

[1] Thanomrat Prasitthimet, and Kasem Wetsutthanon, "Penyebab kematian ibu di Dinas Kesehatan Daerah 4 Tahun Anggaran 2014-2016," Jurnal Ilmu Kesehatan, vol.26, hlm.5, 2017.

[2] Logan Gorber, "Data Replikasi untuk: Penentu preeklampsia dan eklampsia di antara wanita yang melahirkan di rumah sakit daerah di Nairobi, Kenya", Harvard Dataverse, versi 1.0, 2020. [online] Tersedia: http://www.doi.org/10.7910/DVN/BYFU3J. [Diakses: 18 November 2020].

[3] WHO, UNICEF, UNFPA, "The Bank Dunia dalam kematian ibu: 1990 hingga 2010", 2010. [online] Tersedia: https://www.who.int/reproductivehealth/publications/monitoring/9789241503631/en/ [Diakses: Jan.1, 2019].

[4] Dewan Medis Thailand, "Statistik Medis", 2017. [online] Tersedia: http://www.tmc.or.th/pdf/01_stat_med2560.pdf. [Diakses: Jan.1, 2019].

[5] Sistem Registrasi Statistik Resmi, "Jumlah penduduk di setiap umur, Nasional", [online] 2017. Tersedia: http://stat.dopa.go.th/stat/statnew/upstat_age_disp.php. [Diakses: Jan.1, 2019].

[6] Muhlis Tahir, Tessy Badriyah, Iwan Syarif, "Klasifikasi Algoritma Deteksi Risiko Maternal Untuk Preeklampsia Dengan Hipertensi Selama Kehamilan Menggunakan Particle Swarm Optimization," EMITTER International Journal of Engineering Technology, vol.6, pp.236-250, 2018.

[7] Costas K. Neocleous, Panagiotis Anastasopoulos, Kypros H. Nikolaides, Christos N. Schizas, Kleanthis C. Neokleous, "Jaringan saraf untuk memperkirakan risiko terjadinya preeklampsia," dalam Proc. Konferensi Bersama Internasional IEEE tentang Jaringan Neural, 2009, hlm.2221-2224.

[8] Muhlis Tahir, Tessy Badriyah, dan Iwan Syarif, "Algoritma Neural Networks untuk Menanyakan Faktor Preeklampsia Sebelumnya pada Wanita Hipertensi Kronis Selama Kehamilan dalam Proses Persalinan," dalam Prod. Simposium Elektronik Internasional IEEE tentang Penciptaan Pengetahuan dan Komputasi Cerdas (IES-KCIC), 2018, hlm.51-55.

[9] SY Leemaqz, GA Dekker dan CT Roberts "Sistem Prediksi Berjenjang untuk Preeklampsia: aplikasi integratif dari berbagai model," dalam Proc. Kongres Internasional ke-20 tentang Pemodelan dan Simulasi (MODSIM), 2013, hlm.2041-2045.

[10] Mohammed Khalilia, Sounak Chakraborty, dan Mihail Popescu, "Memprediksi risiko penyakit dari data yang sangat tidak seimbang menggunakan hutan acak," BMC Medical Informatics and Decision Making, Vol.11, pp.1-13, 2011.

[11] Shadab Adam Pattekari dan Asma Parveen, "Sistem prediksi penyakit jantung menggunakan naif Bayes," Penelitian Biomedis, vol.29, pp.2646-2649, 2018.

[12] M.Akhil jabbar, BL Deekshatulu, dan Priti Chandra, "Klasifikasi Penyakit Jantung Menggunakan K-Nearest Neighbor dan Genetic Algorithm," Konferensi Internasional tentang Kecerdasan Komputasi: Pemodelan Teknik dan Aplikasi, CIMTA, Kalyani, Kolkata, India, 27 September 2013, hlm.85-94.

[13] Md Osman Goni Nayeem, Maung Ning Wan, dan Md Kamrul Hasan, "Prediksi Tingkat Penyakit Menggunakan Multilayer Perceptron Jaringan Syaraf Tiruan untuk Pemantauan Pasien," International Journal of Soft Computing and Engineering (IJSCE), vol.5., hlm. 17-23, September 2015.

[14] Departemen Kesehatan, dan kantor keamanan kesehatan nasional (nhso) thailand, "catatan kesehatan ibu dan anak atau buku catatan merah muda rumah sakit umum di Thailand," 2018. [online] Tersedia: http://www.oic.go.th/FILEWEB/CABINF0CENTER17/DRAWER002/GENERAL/DATA0001/00001375.PDF. [Diakses: Jan.1, 2019].

[15] NV Chawla, KW Bowyer, LO Hall, WP Kegelmeyer, "SMOTE: Teknik Pengambilan Sampel Minoritas Sintetis," jurnal Penelitian Kecerdasan Buatan, vol.16, hlm. 321-357, 2002.