ISSN 2085-4552

Perbandingan Regresi Linear dengan Heaviside Activation Function dengan Logistic Regression untuk Klasifikasi Diabetes

Felix Indra Kurniadi[1], Vinnia Kemala Putri[2]
[1] School of Engineering and Technology: Informatics Engineering, Tanri Abeng University, Jakarta, Indonesia
[2] Fakultas Ilmu Komputer, Universitas Indonesia, Depok, Indonesia
felixindra@tau.ac.id
vinnia.kemala51@ui.ac.id

Diterima 15 Februari 2018
Disetujui 8 Juni 2018

***Abstract* —Diabetes is one of the diseases that rapidly increase in the world. One of the most used dataset for diabetes is Pima indian dataset. Pima indian have 8 features such as pregnancies, glucose, blood pressure, insulin, BMI, diabetes pedigree function and age. In this research we are comparing between Linear Regression using Heaviside Activation Function and Logistic Regression. Logistic regression gives better result compare linear regression using Heaviside Activation Function.**

Index Terms*—Diabetes, Regresi, *Heaviside Activation Function, Logistic Regression

I. PENDAHULUAN

Diabetes adalah salah satu penyakit yang disebabkan dikarenakan ketidak mampuan pankreas untuk menghasilkan insulin yang mengontrol gula dalam darah. Hal ini jika dibiarkan akan berbahaya dikarenakan dapat terjadi komplikasi seperti jantung, stroke, glaukoma ataupun kanker[1].

Berdasarkan data yang didapatkan oleh WHO (World Health Organization) bahwa[2]:

- Terdapat 346 juta penduduk yang menderita penyakit diabetes mellitus
- Pada tahun 2004, estimasi 3.4 juta penduduk meninggal dikarenakan konsekuensi dari tingginya gula darah.
- Lebih dari 80% penderita diabetes mellitus yang meninggal merupakan warga dengan pendapatan menengah kebawah.

Di Indonesia, berdasarkan data yang dikeluarkan oleh WHO pada tahun 2016, kematian yang disebabkan oleh diabetes sebanyak 99.400 jiwa, dimana 48.300 kasus kematian terjadi pada usia produktif [3]. Sedangkan berdasarkan data yang diambil oleh Riset Kesehatan Dasar (Riskesdas) pada tahun 2013 terdapat 2.650.340 jiwa yang diketahui menderita penyakit diabetes [4]. Dari angka yang diberikan oleh WHO dan Riskesdas dapat disimpulkan bahwa diabetes merupakan salah satu penyakit yang mematikan.

Penggunaan sistem cerdas dapat membantu tenaga medis untuk melakukan klasifikasi terhadap penyakit diabetes berdasarkan data kesehatan pasien. Proses pengklasifikasian diharapkan dapat memberikan pilihan terhadap tenaga medis sehingga tidak terjadi kesalahan dalam pengobatan yang dapat mengakibatkan kematian. Hal inilah yang menyebabkan perkembangan penelitian untuk meningkatkan kemampuan komputer untuk melakukan klasifikasi dengan data pasien[5].

Tujuan dari penelitian ini adalah untuk melakukan proses klasifikasi diabetes menggunakan metode linear regression dengan *activation function* menggunakan *Heaviside methods* kemudian metode akan dibandingkan dengan metode *Logistic Regression*.

Jurnal ini terdiri atas 5 bagian. Pada Bagian I akan menjelaskan mengenai pendahuluan, pada Bagian II akan menjelaskan mengenai penelitian sebelumnya, Bagian III akan menjelaskan mengenai eksperimen yang akan dilakukan, Bagian IV akan memberikan hasil dan diskusi dari penelitian dan Bagian V akan menjelaskan mengenai kesimpulan yang didapatkan dari penelitian ini.

II. PENELITIAN SEBELUMNYA

Beberapa penelitian yang mencoba menyelesaikan permasalahan mengenai klasifikasi diabetes. Heydari mencoba membandingkan diabetes *type* 2 di Iran dengan menggunakan *Support Vector Machine*, *Artificial Neural Network*, *Decision Tree*, *Bayesian Network* dan *Nearest Neighbour*. Data yang digunakan untuk penelitian ini adalah dataset dari *Iranian Ministy Health, Food and Drug Administration* untuk diabetes *type* 2, hasil terbaik didapatkan dengan menggunakan *Artificial Neural Network* dengan tingkat akurasi 0.9744[5] .

Jaafar mencoba menangani permasalahan diabetes melitus dengan menggunakan *Back-propagation Neural Network*. Data yang digunakan adalah data Pima Indian dimana fitur diambil adalah sebanyak delapan fitur. Penelitian ini menyatakan bahwa penggunaan neural network sangat bergantung pada data apabila data yang

digunakan lebih banyak dapat memberikan hasil yang lebih baik dibandingkan dengan saat percobaan dilakukan[6].

Nai-arun mencoba melakukan klasifikasi diabetes dengan menggunakan *Decision Tree, Artificial Neural Network, Random Forest* dan *Naïve Bayes* dan setiap metode akan dilakukan proses *ensemble* dengan menggunakan metode *Bagging* dan *Boosting*. Data didapatkan dari *Sawanpracharak Regional Hospital* antara tahun 2012-2013, kemudian diambil sebanyak 11 fitur yang digunakan. Hasil akurasi terbaik didapatkan metode *Random Forest* dengan akurasi sebesar 85.558%. Hasil penelitian ini kemudian diaplikasikan ke web yang telah dibuat[7].

Penelitian lainnya dilakukan oleh Chandgude, metode yang digunakan adalah *Fuzzy Inference System*. Hasil yang didapatkan bahwa hasil menggunakan FIS didapatkan *precision* sebesar 0.98 dengan 900 *training dataset*. Data yang digunakan adalah *dataset Pima Indian*.

Kumari menggunakan *Back Propagation Neural Network* untuk mencoba mengatasi klasifikasi terhadap orang penderita diabetes mellitus. Data yang digunakan adalah data yang diambil dengan melakukan proses survei. Fitur yang digunakan adalah jenis kelamin, umur, berat, tinggi badan, kekurangan berat badan, keinginan untuk minum, keinginan untuk makan, nafsu makan, pusing, muntah, infeksi dan pandangan berkabur. Hasil yang didapatkan menggunakan *Back Propagation Neural Network* adalah 92.8%[2].

III. EKSPERIMEN

A. Data

Data yang digunakan berasal dari data *Pima Indians Diabetes dataset* yang berasal dari *University of California Irvine Machine Learning Respository* (https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes). Data pima indian sendiri berasal dari *National Institute of Diabetes and Digestive and Kidney Disease* dengan mengambil sampel terhadap 768 orang perempuan keturunan pima indian antara umur 21 tahun ke atas[8].

Data pima indian memiliki 9 attribute yang digunakan:

1. *Pregnancies*: berapa kali sampel mengandung
2. *Glucose*: Konsentrasi *plasma glucose*, 2 jam setelah melakukan *oral glucose tolerance test*
3. *Blood Pressure*: tekanan darah diastolik (mm Hg)
4. *Skin Thickness*: ketebalan kulit trisep
5. Insulin: 2 jam serum insulin
6. BMI: *body mass index*
7. *Diabetes Pedigree Function*: Silsilah diabetes
8. Umur
9. Hasil: diabetes atau tidak diabetes (0 atau 1)

Data pima indian sendiri memiliki permasalahan utama dimana terdapat *missing value* pada beberapa fitur data. Pada penelitian ini nilai *missing value* hanya diubah menjadi 0.

B. Pengaturan Eksperimen

Data yang ada dibagi menjadi dua bagian yaitu untuk proses pelatihan dan proses pengujian. Metode yang digunakan adalah pembagian acak terhadap data utama. Pembagian dilakukan dengan aturan 75% untuk data pelatihan dan 25% untuk data pengujian.

C. Proses Eksperimen

Pada tahapan penelitian data yang sudah dibagi akan dilakukan pencarian model untuk mendapatkan model regresi yang diinginkan. Dari data tersebut dipisah fitur yang digunakan dengan target yang ingin dicapai. Pada penelitian ini fitur yang digunakan adalah *pregnancies, glucose, skin thickness, insulin, BMI, diabetes pedigree function* dan umur. Sedangkan atribut hasil adalah target yang diinginkan.

Pada penelitian mencoba membandingkan metode regresi linear dengan *Heaviside Activation* function dengan metode *Logistic Regression*. Regresi linear merupakan salah satu metode yang digunakan untuk melakukan prediksi nilai sehingga untuk melakukan suatu proses klasifikasi dibutuhkan sebuah *activation function* untuk memberikan klasifikasi terhadap hasil prediksi nilai yang dibuat oleh regresi linear. Fungsi Regresi linear dapat dilihat pada persamaan (1)[9]:

$$\hat{y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \quad (1)$$

Dimana ŷ adalah hasil prediksi, $b_0, b_1, b_2, \ldots, b_n$ adalah koefisien regresi dan $X_1, X_2, \ldots, X_n$ adalah Fitur yang digunakan untuk melakukan klasifikasi.

Proses mencari koefisien regresi yang digunakan dalam penelitian ini dapat dilihat dari persamaan (2):

$$b = (X_i^T . X_i)^{-1} X_i^T Y \quad (2)$$

Dimana Y adalah target yang diinginkan X_i adalah fitur yang digunakan. Hasil yang telah didapatkan dari mencari koefisien regresi akan dimasukkan ke dalam persamaan (1) menjadi sebuah persamaan untuk proses klasifikasi menggunakan metode *Heaviside activation function*. *Heaviside activation function* merupakan metode sederhana dengan menggunakan *threshold* untuk membagi antara nilai 0 dan 1 atau -1 dan 1. Persamaan *Heaviside activation function* adalah[10]:

$$f(x) = \begin{cases} 0 & if\ x < \theta \\ 1 & if\ x \geq \theta \end{cases} \quad (3)$$

Dimana θ adalah *threshold*, dan $f(x)$ adalah nilai hipotesis target yang akan didapatkan.

Selain itu hasil dari proses klasifikasi akan dibandingkan dengan *Logistic Regression*. *Logistic Regression* merupakan salah satu metode untuk melakukan klasifikasi yang baik. *Logistic Regression* itu sendiri adalah sebuah metode untuk pembuatan model yang outputnya berbentuk *crisp* Persamaan untuk *Logistic Regression*[7]:

$$\hat{y} = \frac{e^{(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n)}}{1+e^{(b_0+b_1X_1+b_2X_2+\cdots+b_nX_n)}} \quad (4)$$

Dimana **e** adalah *Euler's equations*

D. Evaluasi

Pengevaluasian hasil terhadap penelitian ini akan menggunakan *accuracy*, *F1-score*, *precision* dan *recall*. *Precision* digunakan untuk mengukur probabilitas terhadap *classifier exactness*, sedangkan recall digunakan untuk mengukur probabilitas terhadap *classifier completeness*. Berbeda dengan *precision* dan *recall*, *F1-score* mencoba membandingkan keseimbangan antara *precision* dan *recall*[11][12].

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} * 100\% \quad (5)$$

$$precision = \frac{tp}{tp+fp} * 100\% \quad (6)$$

$$recall = \frac{tp}{tp+fn} * 100\% \quad (7)$$

$$F1-Score = 2 * \left(\frac{recall * precision}{recall + precision}\right) * 100\% \quad (8)$$

dimana ***tp*** dan ***tn*** merupakan jumlah prediksi yang benar.

IV. HASIL EKSPERIMEN

Pada penelitian ini akan dilakukan pencarian nilai menggunakan regresi linear yang dibantu dengan metode *Heaviside activation function*. Pada penelitian ini *threshold* (θ) yang digunakan adalah 1 sehingga apabila θ lebih kecil dari 1 akan diberi nilai 0 dan sebaliknya maka akan diberi nilai 1. Selain itu penelitian ini akan membandingkan dengan menggunakan *logistic regression*. Hasil untuk *accuracy*, *pression*, *recall* dan *F-Score* dapat dilihat pada Tabel 1 dan Tabel 2. sedangkan pada Tabel 3 dan Tabel 4 akan diberikan data *confusion matrix*.

Tabel 1. Hasil Akurasi, *Precision*, *Recall* dan *F1-score* dari Metode *Linear Regression* menggunakan *Heaviside activation Function*

	Result(%)
Accuracy	74.03
Precision	72.5
Recall	50
F1-Score	59.18

Tabel 2. Hasil Akurasi, *Precision*, *Recall* dan *F1-score* dari Metode *Logistic Regression*

	Result(%)
Accuracy	75.97
Precision	76.92
Recall	51.72
F1-Score	61.86

Tabel 3. *Confusion Matrix* dari Metode *Linear Regression* menggunakan *Heaviside activation Function*

	Diabetes	Tidak Diabetes
Diabetes	85	11
Tidak Diabetes	29	29

Tabel 4. *Confusion Matrix* dari Metode *Logistic Regression*

	Diabetes	Tidak Diabetes
Diabetes	87	9
Tidak Diabetes	28	30

Berdasarkan tabel 1, dan tabel 2 dapat dikatakan bahwa *logistic regression* memberikan hasil yang lebih baik dibandingkan dengan metode linear regression dengan menggunakan *Heaviside activation function*. Pada tabel *confusion matrix* dari kedua metode dapat dilihat adanya ketimpangan data pada data diabetes. Hal ini dapat membuat hasil klasifikasi lebih condong untuk memilih seseorang diabetes.

V. KESIMPULAN DAN PENELITIAN SELANJUTNYA

Pada penelitian ini dapat dilihat bahwa berdasarkan nilai akurasi, *precision*, *recall* dan *F-Score*, metode *Logistic Regression* memberikan hasil yang lebih baik dibandingkan dengan metode linear regresi dengan *Heaviside activation function*. Tetapi perbedaan akurasi kedua metode tersebut tidak terlalu signifikan sehingga kedua metode ini layak untuk digunakan pada penelitian lainnya.

Beberapa kendala yang belum ditangani pada data pima indian adalah ketidakseimbangan antara kelas diabetes dengan tidak diabetes. Kendala pada data ini perlu dikembangkan seperti penambahan data seperti pengambil data lebih banyak untuk menyetarakan ketimpangan pada data. Selain itu kendala pada *missing value* dapat dilakukan proses pencarian nilai *missing value* pada penelitian berikutnya.

DAFTAR PUSTAKA

[1] WebMD, "WebMD Diabetes Center: Types, Causes, Symptoms, Tests, and Treatments." [Online]. Available: https://www.webmd.com/diabetes/default.htm. [Accessed: 02-Feb-2018].

[2] S. Kumari and A. Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus," in *International Conference on Intelligent Systems and Control*, 2013, pp. 373–375.

[3] World Health Organization, "Diabetes country profile 2016," 2016.

[4] Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia, "Situasi dan Analisis Diabetes," Jakarta, 2014.

[5] M. Heydari, M. Teimouri, and Z. Heshmati, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran," *Int. J. Diabetes Dev. Ctries.*, vol. 36, no. 2, pp. 167–173, 2015.

[6] S. F. B. Jaafar and D. M. Ali, "Diabetes Mellitus Forecast Using Artificial Neural Network (ANN)," in *2005 Asian Conference on Sensor and The International Conference on New Techniques in Pharmaceutical and Biomedical Research Proceedings*, 2005, pp. 135–139.

[7] N. Nai-arun and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," in *7th International Conference on Advances in Information Technology*, 2015, pp. 132–142.

[8] R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," *Johns Hopkins APL Tech. Dig.*, vol. 10, pp. 262–266, 1988.

[9] I. Naseem, R. Togneri, S. Member, and M. Bennamoun, "Linear Regression for Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, 2010.

[10] J. Wang, "Analysis and Design of a k -Winners-Take-All Model With a Single State Variable and the Heaviside Step Activation Function," *IEEE Trans. Neural Networks*, vol. 21, no. 9, pp. 1496–1506, 2010.

[11] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision , Recall and F -Score , with Implication for Evaluation," in *European Conference on Infromation Retrieval*, 2005, pp. 345–359.

[12] J. Brownlee, "Classification Accuracy is not Enough Measures You Can See," 2014. [Online]. Available: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/. [Accessed: 06-Apr-2018].