

MENGUKUR TINGKAT AKURASI 6 MODEL REGRESI DALAM MACHINE LEARNING UNTUK ESTIMASI PENYAKIT DIABETES

¹Feby Wulandari Sembiring

Sistem Informasi, STMIK Royal Kisaran

Email: febywulandari@gmail.com

Abstract: This research aims to develop a companion application that uses the regression method as a warning system for the public in maintaining blood sugar levels to minimize the risk of diabetes based on machine learning. Diabetes is one of the diseases with the highest death rate in Indonesia, but the public's low awareness of checking blood sugar levels regularly is due to the high cost of controlling it at hospitals or clinics. This has an impact on increasing the prevalence of diabetes and state losses. This research utilizes the linear regression method with 6 regression models and 3 testing methods to find the best and most accurate model in predicting diabetes levels. Data was processed using Jupyter Notebook tools with the Python programming language. It is hoped that the results of this research can contribute to increasing public awareness of the importance of maintaining healthy blood sugar to prevent diabetes.

Keywords: machine learning; regression; diabetes; warning system.

Abstrak: Penelitian ini bertujuan untuk mengembangkan sebuah aplikasi pendamping yang menggunakan metode regresi sebagai warning system bagi masyarakat dalam menjaga kadar gula darah untuk meminimalisir risiko diabetes berbasis machine learning. Diabetes merupakan salah satu penyakit dengan angka kematian tertinggi di Indonesia, namun rendahnya kesadaran masyarakat dalam memeriksa kadar gula darah secara berkala disebabkan oleh mahalnya biaya pengontrolan di rumah sakit atau klinik. Hal ini berdampak pada meningkatnya prevalensi diabetes dan kerugian negara. Penelitian ini memanfaatkan metode regresi linier dengan 6 model regresi dan 3 metode pengujian untuk mencari model yang paling baik dan akurat dalam memprediksi kadar diabetes. Data diproses menggunakan tools Jupyter Notebook dengan bahasa pemrograman Python. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam meningkatkan kesadaran masyarakat akan pentingnya menjaga kesehatan gula darah untuk mencegah diabetes.

Kata Kunci: machine learning; regresi; diabetes; warning system.

PENDAHULUAN

Machine learning merupakan salah satu kecerdasan buatan (*artificial intelligence*) dan penerapan ilmu komputer yang mencakup algoritma dan pemodelan yang digunakan untuk pengolahan data dan telah digunakan dalam berbagai disiplin ilmu [1]. *Machine learning* dapat kita temui dalam berbagai disiplin ilmu tidak terkecuali ilmu kesehatan untuk memecahkan masalah – masalah yang ada. Selain itu, adanya penggunaan *machine learning* bertujuan untuk membantu meminimalisir kemungkinan kesalahan diagnosis oleh para ahli medis serta membantu untuk mengestimasi atau prediksi hasil dari perkiraan objek tertentu [2]. Untuk itu, penerapan *machine learning* dapat digunakan dalam mendapatkan hasil estimasi dari sebuah objek dari salah satu masalah kesehatan yang menjadi perhatian Indonesia yaitu diabetes.

Diabetes merupakan salah satu penyakit yang menyebabkan angka kematian tertinggi di Indonesia. Berdasarkan data dari Kementerian Kesehatan Republik Indonesia (Kemenkes) pada tahun 2020 bahwa diabetes merupakan penyakit tidak menular (PTM) yang dapat menyebabkan kematian penyakit serta juga dapat menyebabkan kerugian bagi penderita maupun orang disekeliling penderita yang berakibat pada hilangnya produktifitas bagi mereka. Diabetes dapat terjadi akibat adanya gangguan dalam pengelolaan gula darah berlebihan dalam tubuh.

Tingginya kadar gula darah dalam tubuh sering tidak disadari dan tidak hanya menyerang masyarakat dengan usia lanjut namun diabetes juga dapat menyerang masyarakat yang berusia ≥ 15 tahun sehingga hal tersebut menjadi salah satu alasan mengapa dikatakannya diabetes menjadi salah satu penyakit dengan penderita terbanyak di Indonesia. Untuk itu, diperlukan kesadaran untuk menjaga kesehatan dan menjaga kadar gula darah sejak dini agar tetap normal dengan melakukan pengecekan secara berkala yang bertujuan agar masyarakat menjadi lebih berhati – hati terhadap resiko diabetes dan masyarakat dapat mengambil tindakan secara dini terhadap pencegahan diabetes dalam tubuh.

Namun permasalahannya adalah mahalny biaya yang harus dikeluarkan

masyarakat setiap pengecekan gula darah di rumah sakit atau klinik sehingga berdampak pada menurunnya kesadaran masyarakat yang menjadi tidak peduli akan pentingnya memeriksa kadar gula darah secara berkala untuk mengantisipasi diabetes dalam tubuh. Selain itu permasalahan tersebut dapat menghambat masyarakat untuk hidup sehat bebas dari diabetes.

Berikut ini merupakan beberapa dampak negatif dari kelalaian masyarakat yang enggan untuk melakukan pengontrolan gula darah diantaranya : (1) beresiko memiliki resiko diabetes lebih tinggi, (2) apabila telah terjangkit diabetes dan jika dibiarkan akan menyebabkan terjadinya penyakit komplikasi, (3) beresiko mengalami infeksi paru hal tersebut terjadi akibat adanya *hiperglikemia* dimana apabila kadar gula darah tinggi dan kemampuan sel untuk fagosit menurun maka terjadilah infeksi. Tidak hanya itu, meningkatnya prevalensi diabetes berdampak negatif pula pada negara meliputi kerugian finansial dan meningkatnya angka sakit dan angka kematian di Indonesia [3]

Dilansir dari Pusdatin Kemkes RI menyatakan bahwa di Indonesia terdapat prevalensi diabetes berdasarkan diagnosis dokter pada umur 15 tahun sebesar 1,5% dari hasil Riskesdas 2013 dan meningkat menjadi 2% dari hasil Riskesdas 2018. Prevalensi diabetes berdasarkan hasil pemeriksaan gula darah juga turut mengalami peningkatan yaitu sebesar 6,9% dari hasil Riskesdas 2013 dan meningkat menjadi 8,5% dari hasil Riskesdas 2018.

Selain itu, menurut WHO peningkatan yang sangat signifikan yang terjadi pada penderita diabetes yaitu sebesar 108 juta jiwa pada 1980 menjadi 422 juta jiwa pada tahun 2014. Dari pengamatan data diatas, dikhawatirkan terjadinya peningkatan penderita diabetes yang semakin besar setiap tahunnya, maka tidak menutup kemungkinan bahwa akan meningkatnya pula tingkat kematian yang disebabkan oleh diabetes di Indonesia.

Berdasarkan permasalahan mengenai mahalny biaya untuk melakukan pengontrolan gula darah secara berkala yang berdampak pada rendahnya kesadaran masyarakat dari menjaga kesehatan dan gula darah dari bahaya diabetes dan meningkatnya

prevalensi diabetes yang menyebabkan kerugian negara, maka dibutuhkan sebuah aplikasi pendamping untuk membantu mengestimasi kadar diabetes dalam tubuh dengan metode regresi linier sebagai aplikasi yang menjadi *warning system* bagi masyarakat untuk menjaga kadar gula darah dalam upaya meminimalisir diabetes berbasis *machine learning*.

Pada riset sebelumnya yang dilakukan oleh Haryadi, Atmaja, Hakim dan Suwaryo (2021) dengan judul riset “Identifikasi Tingkat Resiko Penyakit Stroke Menggunakan Algoritma Regresi Linear Berganda” riset tersebut untuk tahap proses pengolahan datanya menggunakan *tools RapidMiner*, tahap pengujiannya hanya menggunakan 1 metode saja yaitu *root mean square error* (RMSE) dimana riset tersebut bertujuan untuk mengetahui bagaimana variabel dependen dapat diprediksi melalui variabel independennya.

Sedangkan pada riset ini, sistem berbasis *machine learning* yang akan dikembangkan adalah sistem yang menggunakan tujuh metode regresi yang dimana proses pengolahan data akan dilakukan menggunakan *tools jupyter notebook* pada *software anaconda navigator* dengan bahasa pemrograman python dan tahap pengolahan menggunakan 6 model regresi berganda diantaranya : (1) *Linear Regression*, (2) *Support Vector Regression*, (3) *Ridge Regression*, (4) *Decision Tree Regression*, (5) *Lasso Regression*, (6) *ElasticNet Regression* dan untuk tahap pengujiannya menggunakan 4 metode yaitu MAE, MSE dan RMSE. Tujuannya menggunakan metode tersebut dalam penelitian ini adalah untuk mencari model regresi yang paling baik dan paling akurat yang akan diimplementasikan ke dalam sistem sehingga hasil estimasi dari sistem yang akan digunakan merupakan hasil terbaik.

METODE

Machine Learning

Machine Learning merupakan sebuah pendekatan dari *Artificial Intelligence* (AI) yang digunakan untuk menyelesaikan sebuah permasalahan dengan meniru tingkah laku manusia [4]

Linear Regression

Suatu cara prediksi dengan garis lurus yang menggambarkan hubungan 2 variabel atau lebih dimana hubungannya secara linier antara variabel dependen (Y) dengan variabel independen (X) [5]. Adapun persamaan yang digunakan dalam metode *linear regression* sebagai berikut:

$$\hat{y} = b_0 + b_1x \quad (1)$$

$\hat{y}$: nilai dependen

b_0 : *intercept* (bias)

b_1 : koefisien

x : variabel independen

Support Vector Regression

Metode memodelkan variabel x dan variabel y dengan konsep svm, tujuan metode ini untuk menemukan fungsi yang paling sesuai dengan data yang diberikan dan meminimalkan kesalahan prediksi. Formulanya sebagai berikut [6]:

$$f(x) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(x_1, x) + b \quad (2)$$

$f(x)$: nilai yang diprediksi model

$\alpha_i - \alpha_i^*$: koefisien lagrange

$K(x_1, x)$: fungsi kernel yang digunakan untuk menghitung kedekatan antara 2 titik

b : bias

Ridge Regression

Digunakan untuk mengatasi masalah multikolinearitas (korelasi tinggi antara variabel independen) dan overfitting dalam model regresi linear.

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y \quad (3)$$

$\hat{\beta}^{ridge}$: vektor koefisien regresi ridge

X : Matrix desain yang berisi semua observasi dari variabel independen

Y : vektor variabel dependen

λ (lambda) : parameter penalti (*hyperparameter*)

I : Matriks identitas

X^T : matriks transpos dari X

Decision Tree Regression

Nilai prediksi untuk setiap titik data diperoleh dengan menghitung rata-rata nilai target dari semua titik data yang jatuh ke dalam setiap daun keputusan [7] [8].

Lasso Regression

Atau *least Absolute Shrinkage and selection Operator Regression* Yaitu metode regresi yang digunakan untuk mereduksi dimensi dan melakukan regularisasi dalam model regresi linier [9], dimana tujuannya untuk meningkatkan prediksi dan interpretasi mode dengan mengurangi koefisien regresi yang tidak signifikan

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

$\hat{\beta}^{lasso}$: vektor koefisien regresi lasso

y_i : nilai variabel dependen untuk observasi ke - i

λ (lambda) : parameter penalti yang diatur sebelumnya

P : jumlah variabel independen

Elastic Net Regression

Model regresi yang menggabungkan elemen dari *ridge and lasso regression*. Tujuannya untuk mengatasi masalah multikolinearitas dengan memperkenalkan dua jenis regulasi secara bersamaan [10].

HASIL DAN PEMBAHASAN

Berdasarkan teknik analisis kuantitatif yang sejalan dengan riset ini, proses penafsiran dan penyimpulan hasil riset ini dilakukan berdasarkan proses kerja secara kuantitatif yang mencakup pengumpulan dataset, pengolahan dengan 6 model regresi dan MAE, MSE dan RMSE untuk nilai evaluasinya. Pengolahan dan penganalisisan data dikumpulkan untuk mencari hubungan antar variabel riset. Teknik skala likert juga difokuskan melakukan generalisasi untuk menjelaskan efektivitas sistem perkiraan kadar diabetes dengan metode regresi linier berbasis *machine learning* terhadap kadar diabetes pada masyarakat.

Model	MAE	MSE	RMSE
Linear Regression	0.33325048	0.15983873	0.39979836
Ridge Regression	0.33328463	0.15981181	0.3997647
Lasso Regression	0.37585989	0.17966188	0.4238654
ElasticNet Regression	0.28881838	0.17138319	0.41398453
SVR	0.10156525	0.16912406	0.41124696
Decision Tree	0.16744992	0.1015625	0.31868872

Pada proses akhir, 6 model regresi diukur akurasi dengan menggunakan 3 nilai error. MAE, MSE dan RMSE dikatakan nilainya semakin baik apabila nilainya semakin kecil dan semakin mendekati angka nol. Berdasarkan tabel hasil diatas, terdapat 2 metode yang memiliki nilai paling kecil ketika diuji dengan nilai error yaitu SVR nilai paling kecil pada pengukuran nilai error MAE dan *decision tree* pada pengukuran nilai error MSE dan RMSE. Sehingga dapat diketahui model *decision tree* paling bagus karena terdapat 2 nilai error paling kecil dengan menggunakan model tersebut.

SIMPULAN

Penelitian ini memperoleh hasil model regresi paling baik yang digunakan untuk meng estimasi penyakit diabetes yaitu model *decision tree*, sehingga model tersebut dapat dikembangkan untuk estimasi menggunakan sistem yang berbasis *website* untuk membantu estimasi penyakit diabetes sehingga dapat menjadi *warning system* bagi masyarakat agar terhindar dari penyakit diabetes.

DAFTAR PUSTAKA

- [1] A. Raup, W. Ridwan, Y. Khoeriyah, S. Supiana, and Q. Y. Zaqiah, "Deep Learning dan Penerapannya dalam Pembelajaran," *JiIP - J. Ilm. Ilmu Pendidik.*, vol. 5, no. 9, pp. 3258–3267, 2022, doi: 10.54371/jiip.v5i9.805.
- [2] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Search pada Algoritma Logistic Regression," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, p. 280, 2020, doi: 10.26418/jp.v6i3.40718.
- [3] C. W. Imam, R. Ariyanti, and V. D. P. Putri, "Edukasi Kesehatan dalam Upaya Pencegahan Penyakit Diabetes di Era Pandemi COVID-19," *JPKMI (Jurnal Pengabdian Kpd. Masy. Indones.)*, vol. 2, no. 3, pp. 238–242, 2021, doi: 10.36596/jpkmi.v2i3.187.
- [4] Gunawan, M.I., Sugiarto, D. and Mardianto, I., 2020. Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Search pada Algoritma Logistic Regression. *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, 6(3), pp.280-284.
- [5] Imam, C.W., Ariyanti, R. and Putri, V.D.P., 2021. Edukasi Kesehatan dalam Upaya Pencegahan Penyakit Diabetes di Era Pandemi COVID-19. *JPKMI (Jurnal Pengabdian Kepada Masyarakat Indonesia)*, 2(3), pp.238-
- [6] Haryadi, D., Atmaja, D.M.U., Hakim, A.R. and Suwaryo, N., 2021, December. Tik-50 Identifikasi Tingkat Resiko Penyakit Stroke Menggunakan Algoritma Regresi Linear Berganda. In *Prosiding Seminar Nasional Teknologi Energi dan Mineral* (Vol. 1, No. 2, pp. 1198-1207).
- [7] Kaur, H. and Kumari, V., 2020. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*.
- [8] Hayuningtyas, R.Y. and Sari, R., 2022. Implementasi Data Mining Dengan Algoritma Multiple Linear Regression Untuk Memprediksi Penyakit Diabetes. *Jurnal Teknik Komputer*, 8(1), pp.40-44.
- [9] Siregar, A.M., 2022. Model Analisis Kasus Covid-19 Di Indonesia Menggunakan Algoritma Regresi Linier Dan Random Forest. *Petir*, 15(1).
- [10] Ridwan, D.I., Setianingsih, C. and Murti, M.A., 2021. Prediksi Penggunaan Energi Listrik Menggunakan Metode Support Vector Regression. *eProceedings of Engineering*, 8(6). Asrianda, A., Aidilof, H.A.K.