Jurnal Sistem dan Teknologi Informasi (JSTI)

https://journalpedia.com/1/index.php/jsti Vol. 06, No. 3 Agustus 2024

# KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION

**Zaenal Mutaqin**[1]
Email: 2010631170130@student.unsika.ac.id

**Chaerur Rozikin**[2]
Email: chaerur.rozikin@staff.unsika.ac.id

**Yusrizal Anastya Tomo**[3]
Email: 2010631170129@student.unsika.ac.id

[1,2,3]Universitas Singaperbangsa Karawang

## ABSTRAK

Diabetes merupakan suatu penyakit tidak menular yang cukup serius dimana pankreas tidak dapat memproduksi insulin secara maksimal. Diabetes dapat menyerang siapa saja tanpa mengenal usia, baik lansia, orang dewasa, maupun anak-anak. Jumlah penderita penyakit diabetes meningkat dari tahun ke tahun, baik dari jumlah kasus maupun prevalensi. Pada tahun 2019, jumlah penderita diabetes di dunia sudah mencapai 463 juta orang dan diprediksi akan terus bertambah mencapai angka 700 juta orang pada tahun 2045. Sebanyak 1,6 juta kematian langsung disebabkan diabetes tiap tahunnya. Hal ini menjadikan diabetes sebagai satu dari sepuluh penyakit penyebab kematian di seluruh dunia. Metode yang digunakan dalam penelitian ini adalah menggunakan Metode kuantitatif dengan klasifikasi regresi logistik. Evaluasi model klasifikasi dilakukan dengan melihat nilai Confusion Matrix dan Classification Report dari ketiga model yang dibuat. Model pertama memiliki nilai TP: 96, FP: 9, FN: 21, dan TN: 28. Model kedua memiliki nilai TP: 146, FP: 12, FN: 34, dan TN: 39. Terakhir, model ketiga memiliki nilai TP: 170, FP: 31, FN: 46, dan TN: 61. Sementara untuk Classification report dari ketiga model klasifikasi tersebut menunjukkan bahwa model pertama memiliki akurasi sebesar 81%, model kedua sebesar 80%, dan model ketiga sebesar 75%.

**Kata Kunci:** Diabetes, Penyakit, Algoritma, Logistik, Regresi

## *ABSTRACT*

*Diabetes is a serious non-contagious disease in which the pancreas cannot produce insulin optimally. Diabetes can strike anyone regardless of age, both the elderly, adults and children. The number of people with diabetes is increasing from year to year, both in terms of the number of cases and the prevalence. In 2019, the number of people with diabetes in the world has reached 463 million people and is predicted to continue to increase to reach 700 million people in 2045. As many as 1.6 million deaths are directly caused by diabetes each year. This makes diabetes one of the top ten causes of death worldwide. The method used in this study is to use a quantitative method with logistic regression classification. Evaluation of the classification*

*model is carried out by looking at the Confusion Matrix and Classification Report values of the three models created. The first model has TP values: 96, FP: 9, FN: 21, and TN: 28. The second model has TP values: 146, FP: 12, FN: 34, and TN: 39. Finally, the third model has TP values: 170, FP: 31, FN: 46, and TN: 61. Meanwhile, the Classification report of the three classification models shows that the first model has an accuracy of 81%, the second model is 80%, and the third model is 75%.*

***Keywords:*** *Diabetes, Ilness, Algorithm, Logistic, Regression.*

## 1. PENDAHULUAN

Diabetes merupakan suatu penyakit tidak menular yang cukup serius di mana pankreas tidak dapat memproduksi insulin secara maksimal. Diabetes dapat menyerang siapa saja tanpa mengenal usia baik lansia, orang dewasa, maupun yang terjadi pada anak-anak yang ditandai dengan meningkatnya kadar gula (glukosa) darah dalam tubuh manusia (Reva Cahyani, 2022).

Jumlah penderita penyakit diabetes meningkat dari tahun ke tahun, baik dari jumlah kasus maupun prevalensi. Pada tahun 2019, jumlah penderita diabetes di dunia sudah mencapai 463 juta orang dan diprediksi akan terus bertambah mencapai angka 700 juta orang pada tahun 2045. Penderita diabetes mayoritas tinggal di negara berpenghasilan rendah dan menengah dan 1,6 juta kematian langsung disebabkan diabetes setiap tahunnya. Hal ini menjadikan diabetes sebagai salah satu dari sepuluh penyakit yang termasuk kedalam suatu kategori penyakit yang berbahaya dan juga menjadi penyebab yang paling utama dalam kejadian yang dinamakan dengan kematian di seluruh dunia (Erlin, 2022).

Pada tahun 2019 silam, Negara Indonesia sendiri menempati posisi ke-7 di dunia, menduduki posisi tersebut setelah Negara China, Negara India, Negara Amerika Serikat, Negara Pakistan, Negara Brazil, dan juga dengan Negara Meksiko. Sebagai negara dengan jumlah penderita diabetes yang paling tinggi, dengan jumlah penderitnya yang ternyata sebesar 10,7 juta orang. Pada tahun 2020 silam pula, angka ini meningkat menjadi sebesar 10,8 juta dengan angka prevalensi pasien pengidap diabetes mencapai 6,2% dan diperkirakan jumlah penderita diabetes di Indonesia meningkat menjadi 16,7 juta pada tahun 2045 (Erlin, 2022).

Machine learning merupakan bagian dari kecerdasan buatan yang mampu mempelajari data dengan sendirinya. Machine learning adalah suatu model statistik untuk memprediksi data menggunakan komputer. Salah satu algoritma yang sering digunakan dalam machine learning adalah algoritma yang berupa regresi logistik (Reva Cahyani, 2022).

Algoritma regresi logistik adalah salah satu algoritma yang dapat digunakan dalam

proses machine learning untuk melakukan tugas klasifikasi. Regresi logistik merupakan bentuk khusus analisis regresi dengan menggunakan respon biner dan prediktor yang dapat terdiri dari data kontinu, kategori, atau campuran antara keduanya. Analisis ini tidak memerlukan asumsi distribusi multivariat normal ataupun pada suatu kesamaan matrik varian kovarian, serta dapat juga diterapkan dalam berbagai skala data (Reva Cahyani, 2022).

Diabetes Mellitus merupakan salah satu penyakit terbanyak penderitanya di Indonesia, untuk melakukan prediksi bisa saja terhambat karena bertambahnya data setiap harinya. Dengan berkembangnya suatu teknologi internet maka berbagai sumber data yang berkaitan dengan data diabetes mellitus tersebut dapat dengan mudah diperoleh secara realtime pada waktu tersebut. Namun perkembangan varian dan volume data yang makin berkembang berakibat terhadap kebutuhan sistem komputasi yang baik. Permasalahan ini dapat diatasi dengan menggunakan Cloud Computing yang dalam hal ini tentu saja dengan menggunakan Algoritma Logistic Regression (Muhamad Ichsan Gunawan, 2020).

Berdasarkan dari adanya penjabaran latar belakang yang demikian, maka penulis menggunakan algoritma regresi logistik untuk membuat model yang dapat melakukan klasifikasi penyakit diabetes, sehingga dapat digunakan sebagai acuan untuk pengobatan penderita diabetes bagi dokter di rumah sakit dan masyarakat berdasarkan variabel yang mempengaruhi terjadinya penyakit

## 2. METODE PENELITIAN

Bagian ini menjelaskan tentang metode penelitian yang digunakan. Metode yang digunakan dalam penelitian ini adalah menggunakan Metode kuantitatif dengan klasifikasi regresi logistik. Metode penelitian yang digunakan dalam penelitian ini terdiri dari lima (5) tahapan yang juga dapat untuk dijalankan secara berurutan (Sugiyono, 2016).

Tahapan-tahapannya tersebut terdiri dari pemilihan data, pra-pemrosesan data, normalisasi data, klasifikasi menggunakan regresi logistik, dan evaluasi pada hasil klasifikasi (Arikunto, 2006).

### A. Pemilihan Data

Dataset yang digunakan dalam penelitian ini merupakan Dataset National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) yang diperoleh dari situs Kaggle. Dataset ini berisi informasi tentang pasien wanita yang berusia 21 tahun atau lebih dan diambil dari Pima Indian Heritage. Dataset ini terdiri dari 768 entri dengan 9 variabel yang mencakup

beberapa informasi seperti umur, jumlah kehamilan, konsentrasi glukosa plasma, tekanan darah, ketebalan kulit lipatan trisep, kadar insulin, indeks yang berupa indeks masa tubuh, selain itu dengan indikator riwayat diabetes dalam suatu keluarga, dan juga dengan adanya sebuah atribut label yang dapat dibedakan menjadi dua yaitu tidak menderita diabetes yang dilambangkan dengan 0 dan menderita diabetes yang dilambangkan dengan menggunakan angka 1.

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 3 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 4 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 5 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 6 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 7 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 8 | 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 |
| 9 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 10 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |

Gambar 1. Dataset NIDDK

## B. Pra-Pemrosesan Data

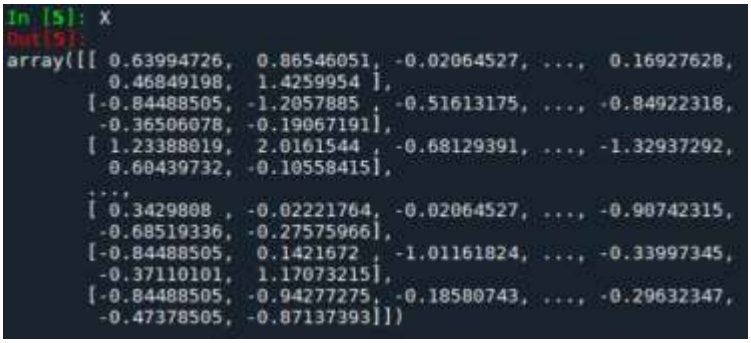
Pra-pemrosesan data dalam penelitian ini dilakukan dengan melakukan imputasi terhadap nilai-nilai nol yang terdapat di dalam beberapa atribut yang tidak boleh diisi dengan nilai nol. Atribut-atribut yang dimaksud diantara lain adalah Glucose, BloodPressure, SkinThickness, Insulin, dan juga dengan BMI. Apabila terdapat nilai nol dalam atribut-atributnya tersebut maka nilainya akan diganti dengan menggunakan rata-rata (mean) dari masing-masing atributnya yang dimaksudkan tersebut.

```
In [4]: df.min()
Out[4]:
Pregnancies                 0.000
Glucose                    44.000
BloodPressure              24.000
SkinThickness               7.000
Insulin                    14.000
BMI                        18.200
DiabetesPedigreeFunction    0.078
Age                        21.000
Outcome                     0.000
dtype: float64
```

Gambar 2. Hasil Proses Imputasi Mean

## C. Normalisasi Data

Normalisasi data adalah suatu strategi atau teknik yang juga dapat digunakan untuk mengubah data ke dalam suatu skala yang lebih standar atau yang biasa disebut dengan normal. Dalam penelitian ini, StandardScaler digunakan untuk mengubah skala data dari fitur-fitur prediktornya sehingga data memiliki rata-rata nol dan standar deviasi satu.




Gambar 3. Hasil Normalisasi Data

## D. Klasifikasi dengan Regresi Logistik

Regresi logistik merupakan salah satu algoritma dari pada proses data mining yang juga dimanfaatkan untuk melakukan suatu identifikasi dan juga dengan analisis data yang mendeskripsikan antara variabel dependen dengan satu atau lebih variabel predictor (Bimantara, 2019). Dalam suatu regresi logistic tersebut, adanya suatu variabel yang dapat dikatakan sebagai suatu variabel yang dependen dan juga yang bersifat dikotomi, yaitu yang bernilai 0 (tidak) atau 1 (ya), sehingga hasil dari pada prediksinya tersebut justru akan selalu diantara 0 atau pun pada angka 1 (Balogun, 2013). Model nilai dari pada huruf w ini juga dapat dilihat pada persamaan (1) berikut ini.

$$W = b + w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n \quad (1)$$

Nilai w yang diperoleh kemudian dipetakan menggunakan fungsi logistic regression.

$$P = \frac{1}{1+e^{-(w)}} (2)$$

Persamaan yang ke (2) yang tentu saja merupakan persamaan umum yang dapat diubah menjadi:

$$P = \frac{1}{1+e^{-(b+w_1x_1+w_2x_2+w_3x_3+\cdots+w_nx_n)}} \quad (3)$$

## E. Evaluasi

Hasil yang diperoleh dari adanya suatu proses yang dalam klasifikasi akan ditampilkan dalam bentuk confusion matrix, akurasi, presisi, recall, dan juga dengan f1-score. Confusion matrix yang mana akan menampilkan hasil klasifikasi dalam 4 bagian, yaitu True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN) dalam penelitian yang dituliskan oleh (Muhamad Ichsan Gunawan, 2020).

Kemudian, untuk akurasi, presisi, dan recall diperoleh dengan perhitungan sebagai

berikut:

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \quad (4)$$

$$Presisi = \frac{TP}{TP+FP} \times 100\% \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (6)$$

Dalam rumus-rumus di atas, TP merupakan jumlah prediksi yang benar untuk kelas positif, TN adalah jumlah prediksi yang benar untuk kelas negatif, FP adalah jumlah prediksi yang salah untuk kelas positif, dan FN adalah jumlah prediksi yang salah untuk kelas negatif

## 3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan dataset yang berbentuk dataset NIDDK, yang mana NDKK ini memiliki 768 entri dalam 9 variabel untuk mendapatkan model yang dapat digunakan untuk melakukan adanya suatu klasifikasi penyakit diabetes yang tentu saja berdasarkan beberapa parameter tertentu. Metode klasifikasi yang digunakan adalah regresi logistik. Dalam penelitian ini nantinya akan ada 3 model dari pada klasifikasi diabetes yang dibedakan berdasarkan perbandingan data latih dan juga dengan data ujinya.

Model pertama memiliki persentase data uji sebesar 20%, model kedua memiliki persentase data uji sebesar 30%, dan model ketiga memiliki presentase data juga dengan uji sebesar 40%.

```
In [55]: confusion_matrix(y_test_1, y_predictions_1)
Out[55]:
array([[96,  9],
       [21, 28]], dtype=int64)

In [56]: confusion_matrix(y_test_2, y_predictions_2)
Out[56]:
array([[146,  12],
       [ 34,  39]], dtype=int64)

In [57]: confusion_matrix(y_test_3, y_predictions_3)
Out[57]:
array([[170,  31],
       [ 46,  61]], dtype=int64)
```

Gambar 4. Hasil *Confusion Matrix*

Gambar 4 tersebut menunjukkan adanya suatu hasil dari pada perhitungan confusion matrix dari pada ketiga model yang sudah dianalisis menggunakan algoritma regresi logistics tersebut. Model yanag pertama memiliki nilai TP sebanyak 96, FP sebanyak 9, FN sebanyak 21, dan TN sebanyak 28. Model kedua memiliki nilai TP sebanyak 146, FP sebanyak 12, FN

sebanyak 34, dan TN sebanyak 39. Terakhir, model ketiga memiliki nilai TP sebanyak 170, FP sebanyak 31, FN sebanyak 46, dan TN sebanyak 61.

```
In [58]: print(report_1)
              precision    recall  f1-score

           0       0.82      0.91      0.86
           1       0.76      0.57      0.65

    accuracy                           0.81
```

Gambar 5. *Classification Report* Model Pertama

```
In [59]: print(report_2)
              precision    recall  f1-score

           0       0.81      0.92      0.86
           1       0.76      0.53      0.63

    accuracy                           0.80
```

Gambar 6. *Classification Report* Model Kedua

```
In [60]: print(report_3)
              precision    recall  f1-score

           0       0.79      0.85      0.82
           1       0.66      0.57      0.61

    accuracy                           0.75
```

Gambar 7. *Classification Report* Model Ketiga

Gambar 5, 6, dan 7 tersebut menunjukkan adanya suatu classification report dari pada ketiga model klasifikasi regresi logistik. Nilai 0 tersebut juga yang menandakan hasil untuk yang tidak menderita diabetes dan nilai 0 mendandakan hasil untuk yang menderita diabetes. Model pertama memiliki akurasi sebesar 81%, model kedua memiliki akurasi sebesar 80%, dan model ketiga tersebut memiliki akurasi yang sebesar 75%.

## 4. KESIMPULAN

Diabetes merupakan suatu penyakit tidak menular yang cukup serius di mana pankreas tidak dapat memproduksi insulin secara maksimal. Diabetes dapat menyerang siapa saja tanpa mengenal usia baik lansia, orang dewasa, maupun anak-anak yang ditandai dengan meningkatnya kadar gula (glukosa) darah dalam tubuh manusia. Jumlah penderita penyakit diabetes meningkat dari tahun ke tahun, baik dari jumlah kasus maupun prevalensi. Pada tahun

2019, jumlah penderita diabetes di dunia sudah mencapai 463 juta orang dan diprediksi akan terus bertambah mencapai angka 700 juta orang pada tahun 2045. Penderita diabetes mayoritas tinggal di negara berpenghasilan rendah dan menengah dan 1,6 juta kematian yang secara langsung disebabkan karena diabetes pada tiap tahunnya.

Confusion matrix dari ketiga model. Model pertama memiliki nilai TP sebanyak 96, FP sebanyak 9, FN sebanyak 21, dan TN sebanyak 28. Model kedua memiliki nilai TP sebanyak 146, FP sebanyak 12, FN sebanyak 34, dan TN sebanyak 39. Terakhir, model ketiga memiliki nilai TP sebanyak 170, FP sebanyak 31, FN sebanyak 46, dan TN sebanyak 61. Sementara untuk Classification report dari ketiga model klasifikasi regresi logistic tersebut menunjukkan bahwa nilai dari nol (0) tersebut yang menandakan bahwa hasil dari pada yang tidak menderita suatu penyakit diabetes dan juga dengan nilai nol (0) yang juga mendandakan hasil untuk yang menderita penyakit diabetes. Model pertama memiliki akurasi sebesar 81%, dan untuk model yang kedua yang tentu saja memiliki tingkat akurasi yang sebesar 80%, dan model ketiga memiliki akurasi sebesar 75%..

## DAFTAR PUSTAKA

Arikunto, S. (2006). Prosedur Penelitian Suatu Pendekatan Praktik. Jakarta: PT. Rineka Cipta.

Balogun, O. S. (2013). Evaluation of logistic regression in classification of drug data in Kwara State. . International Journal of Computational Engineering Research, 3(3), 54-58.

Bimantara, A. &. (2019). Klasifikasi Web Berbahaya Menggunakan Metode Logistic Regression. Annual Research Seminar (ARS), 4(1), 173-177.

Erlin, S. B. (2022). Analisis Diabetes Menggunakan Algoritma Regresi. Jurnal Kesehatan Umum, 1 - 10.

Muhamad Ichsan Gunawan. (2020). Information Technology Articles Comparison of Rice Price Forecasting Using the ARIMA Method on Amazon Forecast and Sagemaker. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) Vol 4 No 3 (2020): Juni 2020 , 1 - 15.

Reva Cahyani. (2022). Klasifikasi Penyakit Diabetes Menggunakan Algoritma Logistic Regression. Jurnal Kesehatan, 1 - 17.

Sugiyono. (2016). Cara Mudah Menyusun: Skripsi, Tesis, dan DisertasI. Bandung: Alfabeta.

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., ... & Makaroff, L. E. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice,

138, 271-281.

World Health Organization (WHO). (2016). Global report on diabetes. Geneva, Switzerland: WHO.

Riskesdas (Riset Kesehatan Dasar) Indonesia 2018. Kementerian Kesehatan RI, Badan Penelitian dan Pengembangan Kesehatan.

Soewondo, P., Ferrario, A., Tahapary, D. L., & Sugiarto, A. (2013). Economic costs of diabetes in Indonesia. Diabetes Research and Clinical Practice, 99(3), 377-384.

Susanto, A., & Wibowo, B. (2020). "Analisis Regresi Logistik untuk Prediksi Diabetes." Jurnal Kesehatan Masyarakat, 5(2), 85-92.

Fitriana, R., & Prasetyo, A. B. (2018). "Pemodelan Klasifikasi Regresi Logistik dalam Mendeteksi Risiko Diabetes Tipe 2 pada Penderita Obesitas." Jurnal Ilmiah Teknologi Informasi Asia, 12(3), 215-222. DOI: 10.21512/asia.v12i3.4721

Suryanto, B., & Wulandari, D. (2017). "Penggunaan Metode Regresi Logistik untuk prediksi Penyakit Diabetes Mellitus pada Pasien Rawat Jalan." Jurnal Ilmu Kesehatan, 5(1), 12-20.

Gunawan, M. I. (2020). "Analisis Klasifikasi untuk Prediksi Diabetes menggunakan Regresi Logistik." Jurnal Penelitian Kesehatan, 7(2), 45-54.

Lestari, D. W., & Herawati, D. (2017). Pemodelan Klasifikasi Diabetes Melitus Menggunakan Metode Support Vector Machine (SVM). Jurnal Gaussian, 6(1), 105-114.

Dewi, N. M., & Budiarto, R. (2018). Klasifikasi Diabetes Melitus Menggunakan Metode K-Nearest Neighbor (KNN). Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI), 7(2), 171-178.

Pradana, A. W., & Rachmawati, D. (2019). Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5. Jurnal Teknologi Informasi dan Komputer (JTIK), 3(1), 38-44.

Suhartono, D., & Purnomo, H. (2020). Implementasi Algoritma Naive Bayes untuk Klasifikasi Pasien Diabetes. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 7(4), 419-428.

Sugiarto, A., & Ariani, N. A. (2021). Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest. Jurnal Teknologi Informasi dan Komunikasi (JTIK), 11(1), 8-16.

Winarno, R., & Hartati, S. (2021). Klasifikasi Penyakit Diabetes Menggunakan Algoritma K-Nearest Neighbor dengan Pemilihan Fitur Berbasis Kombinasi Mutual Information dan Chi-Squared. Jurnal Nasional Teknologi Elektro dan Teknologi Informasi (JNTETI),

10(1), 25-31.

International Diabetes Federation (IDF). (2019). IDF Diabetes Atlas, 9th edition. Brussels, Belgium: IDF.

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (n.d.). National Institute of Diabetes and Digestive and Kidney Diseases Data Sets. Diambil dari situs Kaggle: [https://www.kaggle.com/mathchi/diabetes-data-set].