Jurnal Elektronik Ilmu Komputer Udayana
Volume 11, No 4. Mei 2023

p-ISSN: 2301-5373
e-ISSN: 2654-5101

# Implementasi Logistic Regression dalam Sistem Diagnosa Penyakit Diabetes dengan KNN

I Wayan Trisna Wahyudi[a1], I Gusti Agung Gede Arya Kadyanan[a2],

[a]Informatika, Universitas Udayana
Badung, Indonesia

[1]wayantrisna79@gmail.com
[2]gungde@unud.ac.id

***Abstract***

*Diabetes is a serious chronic disease that occurs when the pancreas does not produce enough insulin. The number of Indonesians suffering from diabetes is estimated to reach 8.2 million in 2020. The existing method for the detection of diabetes is to use laboratory tests. Logistic regression is a statistical tool that can be used in classification modeling about the presence or absence of diabetes. The aim of this study is to predict diagnostically whether a patient has diabetes or not. The results obtained are relatively low predictions because the ranges of values of several factors that cause it are very far apart so normalization is carried out so that the ranges of values are close together. A system can be developed to predict the disease using the principle of classification.*

***Keywords:*** *Exploratory Data Analysis, Logic Regression, Statistical Analysis, Health Care, Diabetes, K-Means*

## 1. Pendahuluan

Diabetes adalah penyakit kronis serius yang terjadi karena pankreas tidak menghasilkan cukup insulin (hormon yang mengatur gula darah atau glukosa), atau ketika tubuh tidak dapat secara efektif menggunakan insulin yang dihasilkannya. International Diabetic Federation (IDF) mengestimasi jumlah penduduk Indonesia usia 20 tahun ke atas, menderita diabetes sebanyak 5,6 juta orang pada tahun 2001, dan meningkat menjadi 8,2 juta orang pada tahun 2020 [1].

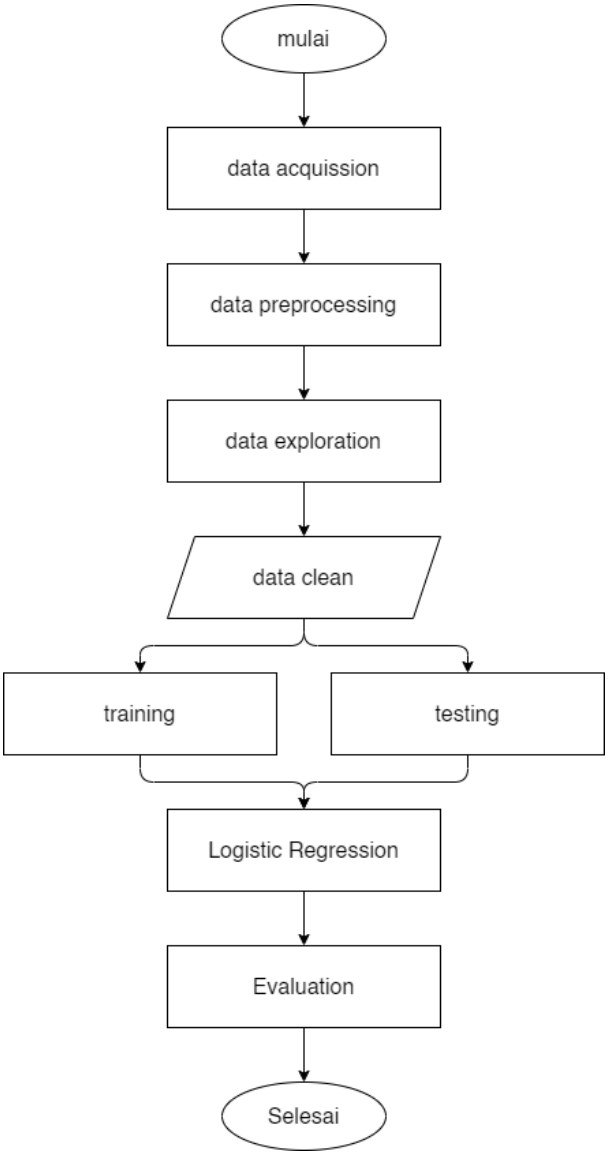
Banyak masyarakat yang awalnya tidak tahu bahwa mereka menderita penyakit diabetes karena tidak mempunyai pengetahuan dasar mengenai penyakit diabetes serta mengalami keterbatasan waktu untuk melakukan konsultasi kepada dokter [2]. Metode yang ada untuk deteksi diabetes adalah dengan menggunakan tes laboratorium seperti glukosa darah dan toleransi glukosa oral. Namun, metode ini memakan waktu lama [3].

Untuk melakukan deteksi dini penyakit diabetes, dapat dikembangkan suatu sistem untuk memprediksi penyakit dengan memanfaatkan berbagai metode. Salah satu metode yang dapat digunakan yaitu metode data mining dengan prinsip klasifikasi [4]. Pada penelitian-penelitian terdahulu, sudah dilakukan penelitian klasifikasi di bidang kesehatan dengan menggunakan teknik atau algoritma data mining dengan studi kasus penyakit diabetes di antaranya menggunakan Algoritma Klasifikasi Decision Tree, Naïve Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN), C4.5, dan penggunaan Logistic Regression Statistical Model yang data nya diperoleh dataset publik [5].

Pada penelitian ini digunakan data hasil survey pasien untuk memprediksi penyakit diabetes. Data tersebut berisi berbagai faktor yang memungkinkan seseorang terkena diabetes, seperti kehamilan, kadar gula darah, tekanan darah, usia, kadar insulin, dan lain-lain. Data-data tersebut kemudian akan diolah untuk mengklasifikasikan pasien apakah terkena diabetes atau tidak. Algoritma untuk mengolah data tersebut pada penelitian ini adalah algoritma logistic regression.

## 2. Metode Penelitian

Pada penelitian ini terdapat beberapa langkah, antara lain: data acquisition, data exploration, modelling, dan evaluation. Metode penelitian dapat dilihat pada gambar 1.




**Gambar 1.** Langkah-Langkah Penelitian

### 2.1. Data Acquission

Data acquisition adalah tahap di mana dilakukan pengumpulan data apa yang diperlukan. Data yang digunakan pada penelitian ini berupa dataset yang berasal dari Institut Nasional Diabetes dan Penyakit Pencernaan dan Ginjal dengan format .csv yang diperoleh melalui situs kaggle. Mengenai karakteristik atribut atau variabel pada dataset dapat dilihat pada tabel 1.

| Variabel | Deskripsi |
|---|---|
| Pregnancies | Jumlah kehamilan pada wanita |
| Glucose | Diukur menggunakan tes toleransi glukosa oral dalam 2 jam |

p-ISSN: 2301-5373
e-ISSN: 2654-5101

| | |
|---|---|
| BloodPressure | Tekanan darah diastolic (mm Hg) |
| SkinThicness | Ketebalan lipatan kulit triceps (mm) |
| Insulin | Serum insulin dalam 2 jam (mu U/ml) |
| BMI | Index masa tubuh (kg/m$^2$) |
| DiabetesPedigreeFunction | Diabetes pedigree function |
| Age | Umur (tahun) |
| Outcome | Class variable (0 or 1) |

**Tabel 1.** Karakteristik Dataset

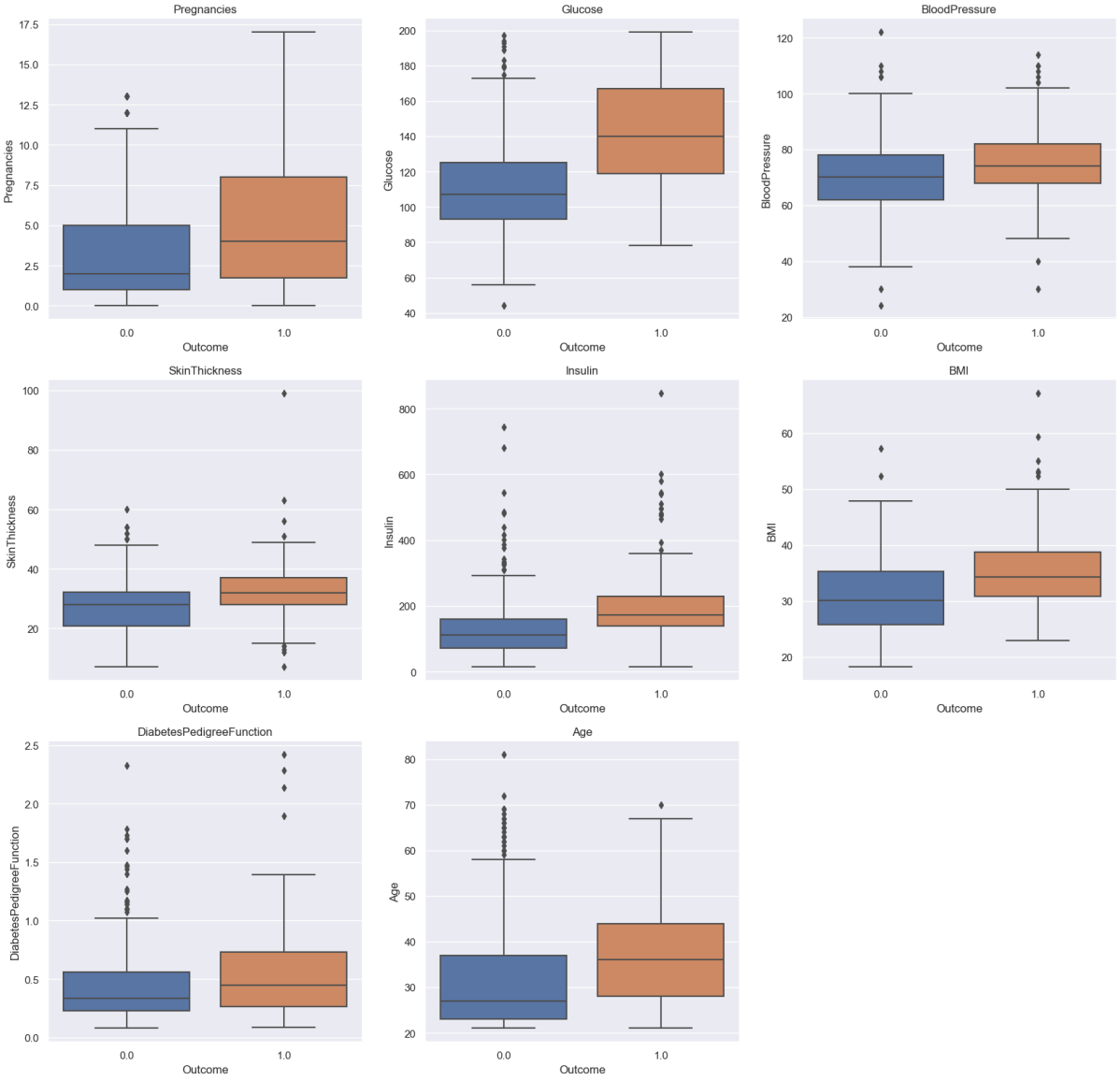
### 2.2. Data Exploration

Setelah tahap data acquisition, proses selanjutnya adalah data exploration. Data exploration adalah tahap yang bertujuan untuk memahami data. Pada proses eksplorasi ini kumpulan dataset yang telah didapatkan melalui situs kaggle, dilakukan preprocessing dengan melihat data duplikat dan memeriksa missing value. Tabel 2 menunjukkan nilai-nilai yang hilang.

| **Variabel** | **Missing Value** |
|---|---|
| Pregnancies | 14.453125 |
| Glucose | 0.651042 |
| BloodPressure | 4.557292 |
| SkinThicness | 29.557292 |
| Insulin | 48.697917 |
| BMI | 1.432292 |
| DiabetesPedigreeFunction | 0.000000 |
| Age | 0.000000 |
| Outcome | 65.104167 |

**Tabel 2.** Missing Value Data

Setelah mengecek data duplikat dan missing value, tahap preprocessing selanjutnya adalah melakukan pengecekan outlier. Gambar 2 menunjukkan outliers pada variabel. Outliers dihapus dengan menggunakan Z-Score. Selanjutnya dilakukan analisis korelasi antar variabelnya. Analisis korelasi variabel digunakan untuk modelling kemudian evaluation. Untuk model ini terdiri dari dua kasus yaitu adanya normalisasi (data clean) dan tanpa normalisasi sebelum modelling. Normalisasi digunakan agar nilai berada pada rentang yang berdekatan sehingga meningkatkan kinerja prediksi.




**Gambar 2.** Outliers Variable

### 2.3. Modelling dan Evaluation

Modelling merupakan tahap dalam pembuatan model dari sistem klasifikasi yang dibuat. Pada penelitian ini menggunakan algoritma logistic regression. Lib linear adalah algoritma yang baik digunakan dalam masalah optimasi logistic regression untuk kumpulan data kecil. Parameter ini mendukung logistic regression dan linear support vector machine.

Lib linear sangat efisien pada kumpulan data yang kecil, besar, dan jarang. Pemilihan algoritma ini didasarkan pada dataset yang dimiliki peneliti memiliki jumlah data yang berkategori dan data numerik sehingga cocok menggunakan algoritma tersebut, dengan demikian dapat diketahui jumlah prediksi dan jumlah sebenarnya dari penderita diabetes. Setelah melakukan training dengan logistic regression, selanjutnya melakukan hasil data testing dan evaluation model.

Evaluation dilakukan dengan memilih satu metrik diantara metrik akurasi, presisi, recall, atau f1-score yang berdasarkan perhitungan nilai True Positive, True Negative, False Positive, dan False Negative pada confusion matrix [15]. Nilai-nilai tersebut dapat digunakan sebagai perbandingan untuk pemilihan acuan metrik pada algoritma untuk model klasifikasi diabetes.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

## 3. Hasil dan Pembahasan

### 3.1. Modelling dan Evaluation

Setelah melihat karakteristik variabel pada Tabel 1, lakukan analisis terhadap nilai-nilai pada setiap variabel. 8 variabel dependen tersebut adalah pregnancies, glucose, blood pressure, skin thickness, insulin, BMI (body mass index), diabetes pedigree function, dan age. Sedangkan 1 variabel dependen adalah outcome. Setiap variabel memiliki rentang nilai yang berbeda-beda. Rentang nilai tiap variabel dapat dilihat pada Tabel 3.

| | **Pregnancies** | Glucose | BloodPressure | SkinThicness | Insulin | BMI | Diabetes Pedigree Function | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| Count | 768.000000 | 763.000000 | 733.000000 | 541.000000 | 394.000000 | 757.000000 | 768.000000 | 768.000000 | 768.000000 |
| Min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.000000 |
| Max | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

**Tabel 3.** Rentang Nilai Tiap Variabel

### 3.2. Preprocessing

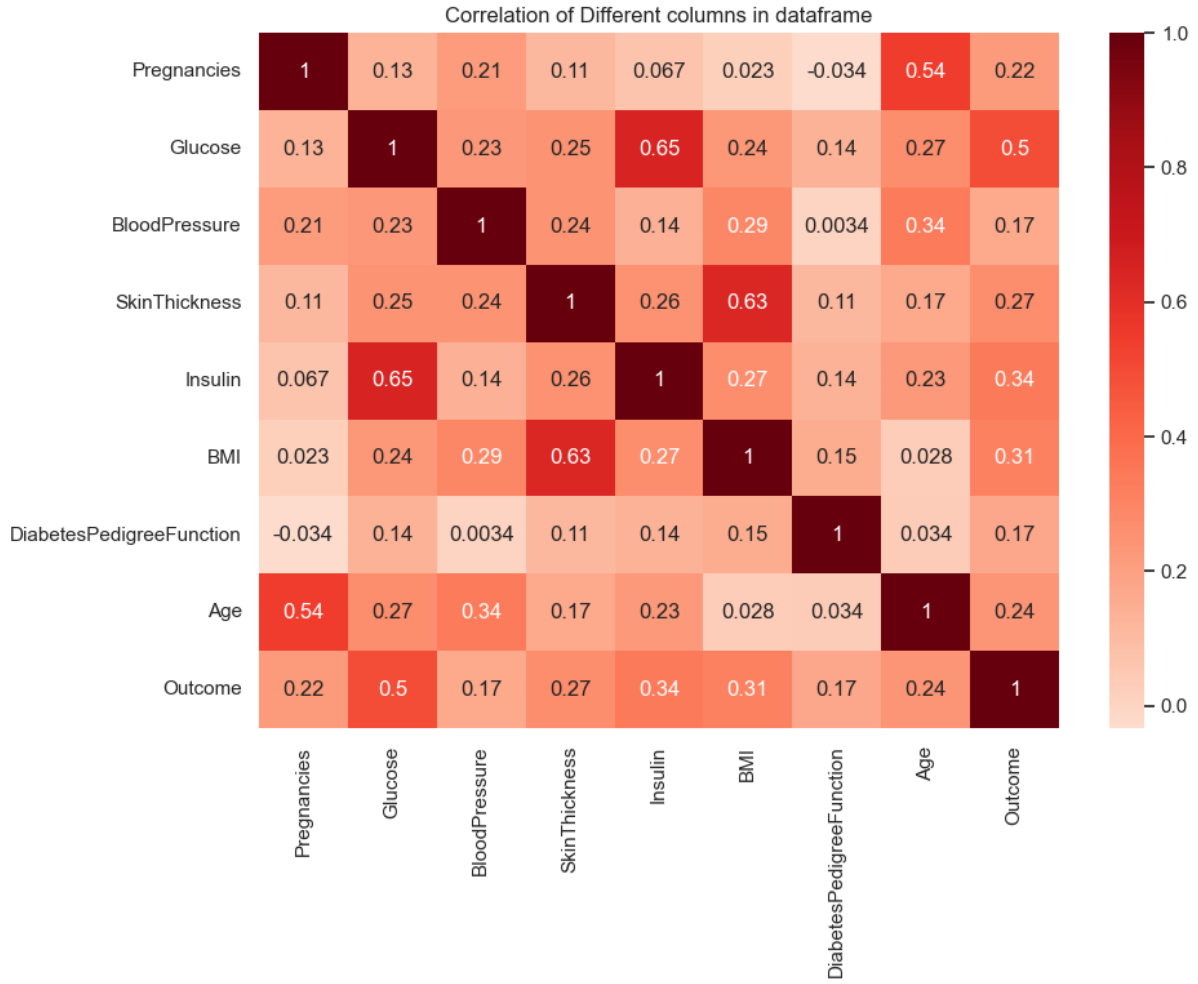
Data yang telah didapatkan dari situs kaggle perlu dibersihkan terlebih dahulu dengan pengecekan data duplikat, missing value, dan outlier. Pada 768 data ini tidak terdapat data duplikat serta tidak terdapat missing value, hanya saja terdapat banyak data yang bernilai nol (0) pada variabel glucose, blood pressure, skin thickness, insulin, dan BMI (Body Mass Index) sehingga termasuk pada nilai yang hilang. Nilai yang hilang ini kemudian diganti dengan mengisi nilai tersebut dengan nilai rata-rata seperti yang terlihat pada Tabel 3.

| | **Pregnancies** | Glucose | BloodPressure | SkinThicness | Insulin | BMI | Diabetes Pedigree Function | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| Count | 768.000000 | 763.000000 | 733.000000 | 541.000000 | 394.000000 | 757.000000 | 768.000000 | 768.000000 | 768.000000 |
| Min | 0.000000 | 44.000000 | 24.000000 | 7.000000 | 14.000000 | 18.200000 | 0.078000 | 21.000000 | 0.000000 |
| Max | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

**Tabel 4.** Missing Value Diganti dengan Nilai Rata-rata

### 3.3. Data Exploration

Data yang sudah bersih kemudian dilihat korelasi (hubungan) antar variabel. Hubungan antar variabel berguna untuk menentukan variabel apa saja yang digunakan untuk modelling. Berikut peta korelasi antar variabel yang ditunjukkan oleh Gambar 3.



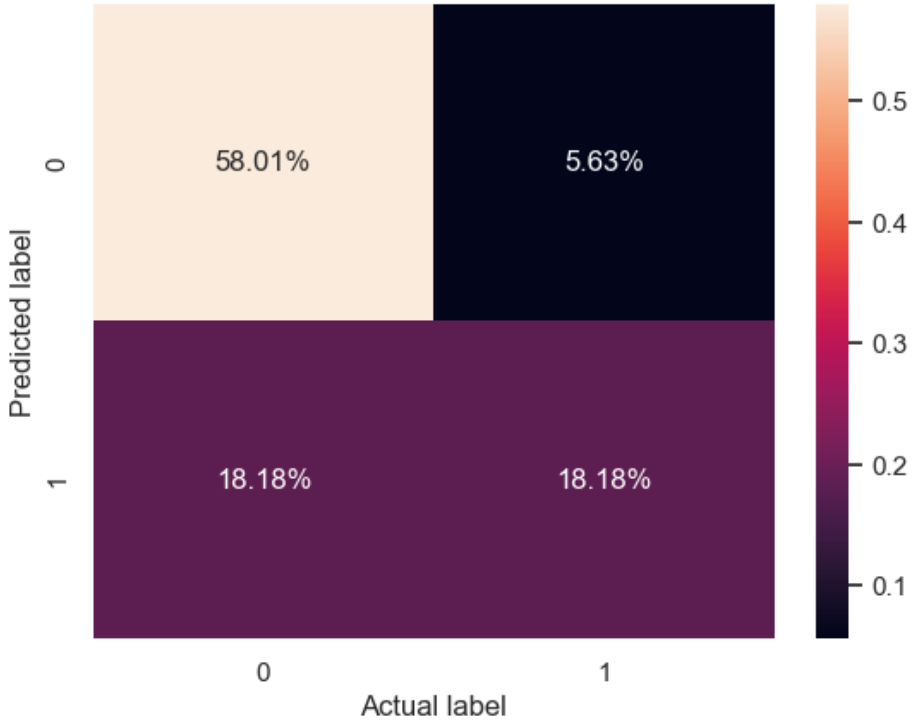

**Gambar 3.** Korelasi antar variabel pada data

Jika nilai korelasi > 0 maka terdapat korelasi positif. Sementara nilai satu variabel meningkat, nilai variabel lainnya juga meningkat. Jika persamaan korelasi = 0 maka tidak ada korelasi. Jika korelasi < 0 maka ada korelasi negatif. Sementara satu variabel meningkat, variabel lainnya menurun. Ketika korelasi diperiksa, ada 2 variabel yang bertindak sebagai korelasi positif terhadap variabel dependen outcome, variabel tersebut adalah glucose. Seiring peningkatan ini, variabel dependen juga meningkat. Dengan demikian, semua variabel digunakan untuk modelling karena korelasinya berdekatan.

### 3.4. Modelling dan Evaluation

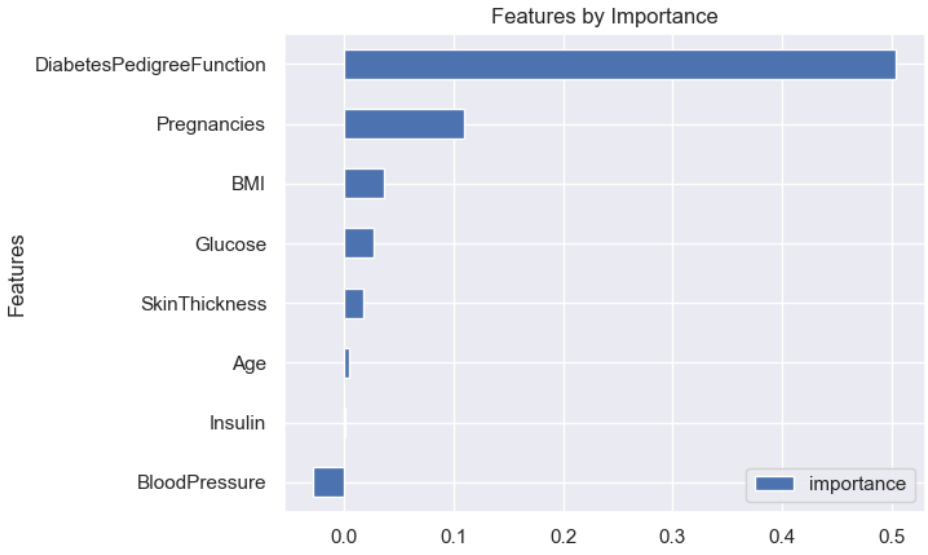
Modeling dilakukan pada data testing, data dipisahkan (split) menjadi data training dan testing dengan rasio 70:30 sehingga dari keseluruhan data berjumlah 768, jumlah data training sebanyak 537 dan testing yang digunakan untuk modelling sebanyak 231 data. Model ini menggunakan semua variabel independen karena hampir semua variabel memiliki korelasi yang mendekati 1. Adapun Confussion matrix dari hasil pengujian model dapat dilihat pada gambar 4.

p-ISSN: 2301-5373
e-ISSN: 2654-5101




**Gambar 4.** Confusion Matrix

Berdasarkan hasil evaluation maka matrix yang paling cocok digunakan dalam sistem ini adalah recall, metrik recall digunakan sebagai acuan pemilihan algoritma terbaik untuk model klasifikasi diabetes karena lebih baik terjadi banyak kesalahan prediksi positif diabetes namun sebenarnya tidak diabetes daripada kesalahan prediksi negatif namun sebenarnya positif diabetes atau lebih baik sedikit jumlah error type II daripada type I di mana semakin besar error type semakin membahayakan untuk kasus prediksi diabetes atau tidak. dari pengujian di atas, terlihat prediksi sistem bernilai 58,01% (dengan normalisasi). Kemudian dari model tersebut dapat dibuat sebuah tabel yang menggambarkan variabel yang paling berpengaruh dalam prediksi penyakit diabetes.




**Gambar 5.** Variable Importance Table

## 4. Conclusion

Prediksi risiko diabetes menggunakan algoritma regresi logistik menggunakan liblinear dengan normalisasi menghasilkan recall sebesar 58%. Model ini diharapkan dapat menjadi acuan untuk pengobatan penderita diabetes bagi dokter di rumah sakit dan di masyarakat untuk mengetahui cara menjaga pola hidup dan cara menghindari penyakit diabetes dilihat dari variabel yang mempengaruhi terjadinya penyakit. Selain itu, disarankan untuk melakukan penelitian tentang prediksi risiko diabetes menggunakan algoritma lain agar mendapatkan kinerja model yang lebih tinggi.

## References

[1] D. Y. Utami, E. Nurlelah, and F. N. Hasan, “Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to predict diabetes,” *J. Inform. Telecommun. Eng.*, vol. 5, no. 1, pp. 53–64, 2021.

[2] Y. B. Widodo, S. A. Anggraeini, and T. Sutabri, “Perancangan Sistem Pakar Diagnosis Penyakit Diabetes Berbasis Web Menggunakan Algoritma Naive Bayes,” *J. Teknol. Inform. Dan Komput. MH Thamrin*, vol. 7, no. 1, pp. 112–123, 2021.

[3] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, “Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest,” *Sist. J. Sist. Inf.*, vol. 10, no. 1, pp. 163–171, 2021.

[4] M. S. Efendi and H. A. Wibawa, “Prediksi Penyakit Diabetes Menggunakan Algoritma ID3 dengan Pemilihan Atribut Terbaik,” *JUITA J. Inform.*, vol. 6, no. 1, pp. 29–35, 2018.

[5] H. Hairani, G. S. Nugraha, M. N. Abdillah, and M. Innuddin, “Komparasi akurasi metode correlated naive Bayes classifier dan naive Bayes classifier untuk diagnosis penyakit diabetes,” *InfoTekJar J. Nas. Inform. Dan Teknol. Jar.*, vol. 3, no. 1, pp. 6–11, 2018.