Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes

Aditya Pratama[1], Azriel Christian Nurcahyo[2], Listra Firgia[3]
[1] Program Studi Sistem Informasi, Universitas Nahdlatul Ulama Kalimantan Barat, Kubu Raya, Indonesia
[2,3] Program Studi Teknologi Informasi,Institut Shanti Buana, Bengkayang, Indonesia
e-mail: [1]adityapratamabadra@unukalbar.ac.id, [2]azriel@shantibuana.ac.id, [3]listrajsc@shantibuana.ac.id

Abstrak

Diabetes merupakan masalah kesehatan global yang signifikan, mempengaruhi kualitas hidup individu dan menimbulkan beban ekonomi dan sosial. Machine Learning (Pembelajaran Mesin) adalah cabang dari kecerdasan buatan yang berfokus pada pengembangan algoritma untuk komputer belajar dari data dan membuat keputusan atau melakukan tugas tanpa pemrograman eksplisit. Logistik Regresi adalah metode statistik yang digunakan untuk menganalisis hubungan antara variabel prediktor dan variabel respons yang bersifat biner. Penelitian ini bertujuan menerapkan algoritma logistik regresi dalam sistem cerdas untuk memprediksi risiko diabetes. Dalam penelitian ini, telah dibuat suatu model prediksi dengan menggunakan logistic regression di Python IDE (Jupyter Notebook) untuk tujuan deteksi dini, dengan fokus pada kemampuan untuk memprediksi apakah seseorang mungkin mengalami penyakit diabetes berdasarkan data awal yang disediakan. Eksperimen dilakukan menggunakan dataset dari Pima Indians Diabetes Database yang terdiri dari 768 data pasien dengan delapan variabel independen dan satu variabel dependen. Metode CRISP-DM digunakan dengan tahapan: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment. Hasil pengujian model logistik regresi menunjukkan tingkat akurasi sebesar 82%. Diharapkan sistem ini dapat membantu masyarakat dan tenaga medis untuk mengidentifikasi faktor risiko dan mencegah penyakit diabetes.
Kata kunci: *Diabetes,Logistik Regresi,CRISP-DM, Confusion Matrix.*

Abstract

Diabetes is a significant global health problem, affecting the quality of life of individuals and creating an economic and social burden. Machine Learning is a branch of artificial intelligence that focuses on developing algorithms for computers to learn from data and make decisions or perform tasks without any specified programming. Logistic Regression is a statistical method used to analyze the relationship between predictor variables and response variables that are binary in nature. This study aims to apply the logistic regression algorithm in an intelligent system to detect diabetes risk. In this research, a prediction model has been created using logistic regression in Python IDE (Jupyter Notebook) for early detection purposes, with a focus on the ability to predict whether a person is likely to have diabetes based on the initial data provided. The experiment was carried out using a dataset from the Pima Indians Diabetes Database which consisted of 768 patient data with eight independent variables and one dependent variable. The CRISP-DM method is used with the following stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The results of testing the logistic regression model show an accuracy rate of 82%. It is hoped that this system can help the public and medical personnel to identify risk factors and prevent diabetes.
Keywords*: Diabetes,Logistik Regresi,CRISP-DM, Confusion Matrix.*

1. Pendahuluan

Diabetes merupakan salah satu masalah kesehatan global yang signifikan di era modern. Penyakit ini menjadi perhatian serius karena dampaknya yang merugikan pada kualitas hidup individu, serta beban ekonomi dan sosial yang ditimbulkannya. Diabetes mellitus adalah kondisi kronis yang ditandai dengan tingginya kadar *glukosa* darah akibat masalah dalam produksi *insulin* atau resistensi insulin oleh sel tubuh. Diabetes dapat mengenai individu dari segala rentang usia, termasuk lansia, orang dewasa, dan anak-anak. Kondisi ini ditandai oleh peningkatan kadar gula (*glukosa*) dalam tubuh manusia[1].

Machine Learning (Pembelajaran Mesin) adalah cabang dari kecerdasan buatan (*Artificial Intelligence)* yang berfokus pada pengembangan sistem dan algoritma yang memungkinkan komputer

untuk belajar dari data, mengidentifikasi pola, dan mengambil keputusan atau melakukan tugas tanpa perlu diprogram secara eksplisit. Tujuan utama dari *Machine Learning* adalah membuat model prediktif atau deskriptif berdasarkan data yang diberikan, sehingga komputer dapat mengerti atau belajar dari data tersebut dan melakukan tugas-tugas tertentu dengan lebih baik seiring berjalannya waktu. Dan salah satu algoritma yang digunakan dalam *machine learning* adalah Logistik Regresi [2].

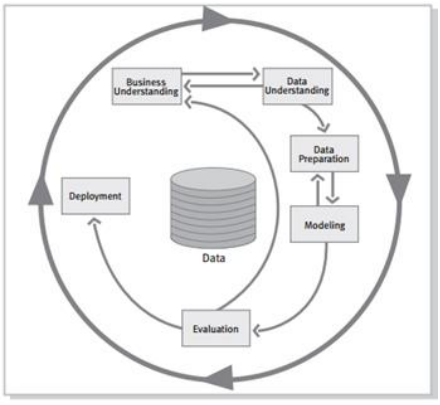
Logistik Regresi adalah metode statistik yang digunakan untuk menganalisis hubungan antara variabel prediktor (variabel independen) dan variabel respons (variabel dependen) yang bersifat biner atau kategorikal. Variabel respons dalam logistik regresi hanya memiliki dua nilai atau kategori, misalnya Ya dan Tidak ,Sukses dan Gagal ,atau Benar dan Salah. Logistik regresi cocok untuk mengatasi data yang bersifat dikotomus (hanya memiliki dua kategori) dan memberikan informasi tentang seberapa kuat pengaruh variabel prediktor terhadap kemungkinan kejadian pada variabel respons. Analisis ini banyak digunakan dalam berbagai bidang, termasuk ilmu sosial, kesehatan, ekonomi, sains, dan teknik, untuk memahami hubungan dan pengaruh faktor-faktor terhadap hasil tertentu [3].

Penelitian yang dilakukan oleh Sholeh M et al [6] membahas tentang penggunaan metode prediksi dalam data mining, khususnya regresi linear, untuk memprediksi hasil nilai ujian siswa berdasarkan atribut-atribut yang mempengaruhi prestasi akademik. Menggunakan metode *CRISP-DM*, penelitian ini melakukan proses pemodelan menggunakan aplikasi RapidMiner pada datasheet publik yang terdiri dari 395 data dan 33 atribut. Hasil evaluasi model regresi linear menunjukkan persamaan $y = 0{,}729 - (0{,}024 \times Medu) - (0{,}020 \times Fedu) + (0{,}053 \times failures) - (0{,}077 \times goout) - (0{,}012 \times absences) + (0{,}126 \times G1) + (0{,}862 \times G2)$ dengan nilai Root Mean Squared Error (RMSE) sebesar 0,675. Berdasarkan hasil evaluasi ini, model yang dihasilkan dapat direkomendasikan untuk digunakan dalam memprediksi nilai ujian siswa. Penelitian lain yang dilakukan oleh Cahyani et al [1] Hasilnya prediksi risiko diabetes menggunakan algoritma regresi logistik dengan normalisasi menghasilkan recall sebesar 55% sedangkan tanpa normalisasi sebesar 43%. Dengan demikian, normalisasi dapat meningkatkan kinerja prediksi risiko diabetes menggunakan algoritma regresi logistik

Berdasarkan penelitian-penelitian sebelumnya, penulis memutuskan untuk menerapkan algoritma regresi logistik dalam pembuatan sistem cerdas yang bertujuan untuk memprediksi apakah seseorang menderita diabetes atau tidak. Dengan sistem ini, diharapkan sistem dapat berfungsi sebagai panduan bagi dokter di rumah sakit dan masyarakat untuk mengenal cara menjaga pola hidup dan menghindari penyakit diabetes berdasarkan faktor-faktor yang mempengaruhi terjadinya penyakit tersebut

2. Metode Penelitian

Berdasarkan permasalahan penelitian tersebut dalam prediksi diabetes dengan logistik regresi , maka dari itu penulis menggunakan metode yaitu *CRISP-DM (Cross Industry Standard Process for Data Mining)*. Metode ini memiliki 6 tahapan dalam data mining, yaitu (1) *Business Understanding*; (2) *Data Understanding*; (3) *Data Preparation*; (4) *Modelling*; (5) *Evaluation*; (6) *Deployment*(Suhanda et al., 2020b).



Gambar 1 : Tahapan dalam CRISP-DM (Chapman et al. 2000)

1. Business Understanding

Ini adalah tahap pertama dalam *CRISP-DM* dan termasuk bagian yang cukup vital. Pada tahap ini membutuhkan pengetahuan dari objek bisnis, bagaimana membangun atau mendapatkan data, dan bagaimana untuk mencocokan tujuan pemodelan untuk tujuan bisnis sehingga model terbaik dapat dibangun. Kegiatan yang dilakukan antara lain: menentukan tujuan dan persyaratan dengan jelas secara

keseluruhan, menerjemahkan tujuan tersebut serta menentukan pembatasan dalam perumusan masalah data mining, dan selanjutnya mempersiapkan strategi awal untuk mencapai tujuan tersebut.

2. Data Understanding

Secara garis besar untuk memeriksa data, sehingga dapat mengidentifikasi masalah dalam data. Tahap ini memberikan fondasi analitik untuk sebuah penelitian dengan membuat ringkasan *(summary)* dan mengidentifikasi potensi masalah dalam data. Tahap ini juga harus dilakukan secara cermat dan tidak terburu-buru, seperti pada visualisasi data, yang terkadang *insight*-nya sangat sulit didapat jika dihubungkan dengan *summary* data nya. Jika ada masalah pada tahap ini yang belum terjawab, maka akan mengganggu pada tahap modeling. Ringkasan atau summary dari data dapat berguna untuk mengkonfirmasi apakah data terdistribusi seperti yang diharapkan, atau mengungkapkan penyimpangan tak terduga yang perlu ditangani pada tahap selanjutnya, yaitu Data *Preparation*. Masalah dalam data biasanya seperti nilai-nilai yang hilang, outlier, berdistribusi spike, berdistribusi bimodal harus diidentifikasi dan diukur sehingga dapat diperbaiki dalam Data *Preparation*.

3. Data Preparation

Secara garis besar untuk memperbaiki masalah dalam data, kemudian membuat variabel derived. Tahap ini jelas membutuhkan pemikiran yang cukup matang dan usaha yang cukup tinggi untuk memastikan data tepat untuk algoritma yang digunakan. Bukan berarti saat Data *Preparation* pertama kali dimana masalah-masalah pada data sudah diselesaikan, data sudah dapat digunakan hingga tahap terakhir. Tahap ini merupakan tahap yang sering ditinjau kembali saat menemukan masalah pada saat pembangunan model. Sehingga dilakukan iterasi sampai menemukan hal yang cocok dengan data. Tahap sampling dapat dilakukan disini dan data secara umum dibagi menjadi dua, data training dan data testing. Kegiatan yang dilakukan antara lain: memilih kasus dan parameter yang akan dianalisis (*Select Data*), melakukan transformasi terhadap parameter tertentu (*Transformation*), dan melakukan pembersihan data agar data siap untuk tahap *modeling* (*Cleaning*).

4. Modelling

Secara garis besar untuk membuat model prediktif atau deskriptif. Pada tahap ini dilakukan metode statistika dan *Machine Learning* untuk penentuan terhadap teknik data mining, alat bantu data mining, dan algoritma data mining yang akan diterapkan. Lalu selanjutnya adalah melakukan penerapan teknik dan algoritma data mining tersebut kepada data dengan bantuan alat bantu. Jika diperlukan penyesuaian data terhadap teknik data mining tertentu, dapat kembali ke tahap data preparation. Beberapa modeling yang biasa dilakukan adalah *classification, scoring, ranking, clustering, finding relation,* dan *characterization*.

5. Evaluation

Melakukan interpretasi terhadap hasil dari data yang dihasilkan dalam proses pemodelan pada tahap sebelumnya. Evaluasi dilakukan terhadap model yang diterapkan pada tahap sebelumnya dengan tujuan agar model yang ditentukan dapat sesuai dengan tujuan yang ingin dicapai dalam tahap pertama.

6. Deployment

Tahap deployment atau rencana penggunaan model adalah tahap yang paling dihargai dari proses CRISP-DM. Perencanaan untuk *Deployment* dimulai selama *Business Understanding* dan harus menggabungkan tidak hanya bagaimana untuk menghasilkan nilai model, tetapi juga bagaimana mengkonversi skor keputusan, dan bagaimana untuk menggabungkan keputusan dalam sistem operasional. Pada akhirnya, rencana sistem *Deployment* mengakui bahwa tidak ada model yang statis. Model tersebut dibangun dari data yang diwakili data pada waktu tertentu, sehingga perubahan waktu dapat menyebabkan berubahnya karakteristik data. Modelpun harus dipantau dan mungkin diganti dengan model yang sudah diperbaiki.

3. Hasil dan Pembahasan

Dengan menggunakan metode CRISP-DM, ada beberapa proses yang dilakukan oleh peneliti, diantaranya:

3.1 *Business Understanding*

Pada tahap ini, peneliti harus memahami secara menyeluruh tujuan dari studi kasus diabetes, masalah kesehatan terkait, dan pertanyaan penelitian yang ingin dijawab melalui analisis data. Serta mengembangkan model prediktif yang dapat membantu dalam diagnosis atau pengelolaan diabetes, serta implikasi praktis dari hasil penelitian untuk perbaikan perawatan dan kebijakan kesehatan.

3.2 *Data Understanding*

Pada penelitian ini dataset yang digunakan berasal dari *PIMA Indians Diabetes* dari "*Kaggle*". Pada dataset ini terdapat 768 baris data dan 9 Kolom atribut. Atribut tersebut terdiri dari *Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome*. pada umumnya kita cek apakah data kita baik, dimana tidak ada nilai yang kosong atau data yang tidak konsisten

untuk dilanjutkan ke langkah selanjutnya. Berikut dataset yang digunakan dalam penelitian ini yang terdiri dari 9 kolom yang digunakan sebagai model datanya.

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Gambar 2. Dataset

Adapun variabel yang digunakan dalam analisa prediksi diabetes adalah sebagai berikut:
Tabel 1 Variabel

No	Variabel	Deskripsi
1	*Pregnancies*	jumlah riwayat kehamilan pasien
2.	*Glucose*	Jumlah Konsentrasi Glukosa
3.	*BloodPressure*	Tekanan Darah (mmHg)
4.	*SkinThickness*	Ketebalan kulit trisep (mm)
5.	*Insulin*	Jumlah Insulin Serum
6.	*BMI*	Indeks Massa Tubuh (kg/m^2)
7.	*DiabetesPedigreeFunction*	Fungsi diabetes berdasarkan riwayat penyakit keluarga
8.	*Age*	Umur
9.	*Outcome*	Hasil yang menunjukkan 0 untuk nondiabetes 1 jika diabetes

3.3 Data Preparation

Pada langkah selanjutnya, penulis meninjau dataset setelah memahami dataset dan atributnya dimiliki jika nilai hilang, atau data tidak konsisten jika data hilang Pembersihan data kemudian dilakukan dan kolom kosong dapat diisi dengan kolom atau rata-rata Kosong dapat dihapus. Alasan mengapa penyiapan data antara dilakukan pada tahap ini Pemeriksaan kedua untuk nilai kosong untuk memastikan bahwa data yang diproses bersih dan diperlukan kesalahan dan menjaga konsistensi data (Cazacu & Titan, 2020), Adapun dataset yang digunakan dilakukan proses cleaning data adapun data terlihat gambar dibawah ini :

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Gambar 3. Data Sebelum di Cleaning

Setelah terlihat pada gambar 3 terdapat variabel *bloodPressure*, *SkinThickness*, *Insulin*, dan *BMI* , maka dilakukan proses cleaning dengan mengisi nilai yang kosong dengan nilai rata-rata setelah dilakukan proses cleaning akan tampak seperti dibawah ini

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.000000	79.799479	33.6	0.627	50	1
1	1	85.0	66.0	29.000000	79.799479	26.6	0.351	31	0
2	8	183.0	64.0	20.536458	79.799479	23.3	0.672	32	1
3	1	89.0	66.0	23.000000	94.000000	28.1	0.167	21	0
4	0	137.0	40.0	35.000000	168.000000	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101.0	76.0	48.000000	180.000000	32.9	0.171	63	0
764	2	122.0	70.0	27.000000	79.799479	36.8	0.340	27	0
765	5	121.0	72.0	23.000000	112.000000	26.2	0.245	30	0
766	1	126.0	60.0	20.536458	79.799479	30.1	0.349	47	1
767	1	93.0	70.0	31.000000	79.799479	30.4	0.315	23	0

768 rows × 9 columns

Gambar 4 data setelah Cleaning

Setelah proses tersebut dilakukan data tahapan selanjutnya adalah splitting dengan rasio 80% untuk data training data 20% untuk data testing. Adapun proses data splitting ini dilakukan secara acak agar terhindar dari biasnya data terhadap data training maupun data testing.

3.4 Modelling

Teknik pemodelan yang digunakan adalah algoritma Logistik Regresi, implementasi algoritma tersebut menggunakan *library scikit-learn* dengan bahasa pemrograman Python. Pada tahap ini dilakukan pemilihan desain uji untuk teknik pemodelan dan pada penelitian ini metode yang digunakan adalah *confusion matrix* untuk menghasilkan nilai dari akurasi, presisi, *recall* dan *F1-score*.

```
scaler= StandardScaler()
X =  pd.DataFrame(scaler.fit_transform(df.drop(["Outcome"],axis = 1)),
        columns=['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age'])
```

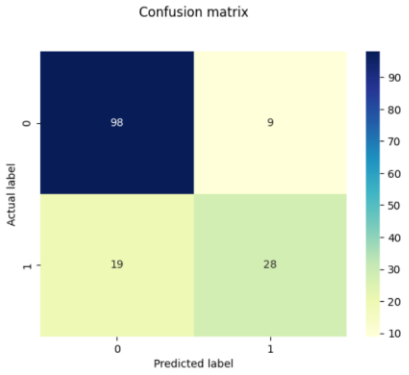
```
model = LogisticRegression(solver='newton-cg')
model.fit(X_train,y_train)
```

```
LogisticRegression
LogisticRegression(solver='newton-cg')
```

Gambar 5. Penerapan Pemodelan Logistik Regresi

3.5 Evaluasi(Pengujian)

Pada tahap ini penulis melakukan evaluasi peneliti menggunakan parameter confusion metrix terhadap model logistik regresi dengan nili dari akurasi, presisi, recall, f1-score. Berikut hasil dari evaluasi model logistik regresi



Gambar 6. Uji *Confusion Matrix*

Berdasarkan gambar tersebut diatas, adapun hasil dari pengujian model logistik regresi dengan menggunakan *tools jupyter notebook* dengan bahasa *python* menghasilkan nilai akurasi sebesar 0.82, *presisi* = 0,81 , *recall* = 0.82 , *f1-score* = 0.81 sehingga dari hasil pengujian tersebut didapatkan tingkat akurasi dengan algoritma logistik regresi sebesar 82%.

4. Kesimpulan

Penelitian ini membahas tentang penerapan algoritma logistik regresi dengan metode *CRISP-DM* terhadap penyakit diabetes. Dengan jumlah data sebanyak 786 data dengan memiliki 8 atribut independen dan 1 atribut dependen. Berdasarkan pengujian yang dilakukan penelitian ini memiliki tingkat akurasi sebesar 80%. Diharapkan di kemudian hari penelitian ini dapat dilakukan dengan jumlah data yang lebih dari sekarang sehingga menghasilkan model dengan tingkat akurasi yang lebih baik.

Daftar Pustaka

[1] Cahyani Q,Finandi M,Rianti J et al.,"Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm " JOMLAI: Journal of Machine Learning and Artificial Intelligence, (2022), 2828-9099, 1(2)

[2] Diksa I,Fithriasari K, "Analisis Faktor Resiko Penyebab Diabetes Mellitus dengan Regresi Logistik Biner," Inferensi, (2020), 69, 4(1).

[3] Wayan I et al., "Implementasi *Logistic Regression* dalam Sistem Diagnosa Penyakit Diabetes dengan KNN," Elektronik Jurnal, pp. 743-749, 11(4)

[4] Galih M, Dina Atika P, "Prediksi Penjualan Menggunakan Algoritma Regresi Linear pada Koperasi Karyawan Usaha Bersama", *Journal of Informaton and Information Security(JIFORTY)*,(2023),193-202,3(2)

[5] Maisyarah C,Haryatmi E,Fajriatifah R et al."Prediksi Penjualan Menggunakan Algoritma Regresi Linear pada Koperasi Karyawan Usaha Bersama", Jurnal Data *Science* dan Informatika,(2022),46-52,2(1)

[6] Karyadiputra E,Setiawan A, "Penerapan Data Mining Untuk Prediksi Awal Kemungkinan Terindikasi Diabetes",Tekno Sains,(2022),221-232,16(2)

[7] Karyadiputra E,Setiawan A, "Penerapan Data Mining dengan Metode Regresi Linear untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner",Jurnal Informatika Sunan Kalijaga (JISKA),(2023),10-21,8(1)