ISSN: 2808-7771 online

**Prosiding SENAM 2024:**
**Sistem Informasi & Informatika Vol. 4**
pp. 48-57, 2024

# Penerapan Model CRISP-DM untuk Prediksi Penyakit Diabetes Menggunakan Metode K-Nearest Neighbor dan Logistic regression

**Angel Aprilia Putri Lo[1] dan Vincentia Jennifer Evelyn Tjioe[2]**

[1,2] Program Studi Sistem Informasi, Universitas Ciputra Surabaya
CitraLand CBD Boulevard, Surabaya, Indonesia, 60219

**Korespondensi:** Angel Aprilia Putri Lo (aaprilia01@student.ciputra.ac.id)

*Received*: 24 Juli 2024 – *Revised*: 31 Agustus 2024 - *Accepted*: 05 Sept 2024 - *Published*: 10 Sept 2024

**Abstrak.** Penyakit diabetes merupakan tantangan kesehatan global yang semakin meningkat. Penelitian ini menggunakan model CRISP-DM, metode K-Nearest Neighbor (K-NN) dan regresi logistik, untuk memprediksi kemungkinan seseorang menderita diabetes berdasarkan data medis. Dataset yang digunakan berasal dari The National Institute of Diabetes and Digestive and Kidney Diseases, yang tersedia di Kaggle. Metode K-NN mengklasifikasikan data baru berdasarkan kedekatan dengan data dalam dataset menggunakan jarak Euclidean, sedangkan regresi logistik memprediksi probabilitas kejadian biner (diabetes atau tidak) berdasarkan hubungan linier antara variabel independen dan dependen. Hasil analisis menunjukkan bahwa metode regresi logistik memiliki performa yang lebih baik dalam klasifikasi diabetes, dengan regresi logistik menunjukkan hasil yang lebih superior dalam akurasi. Akurasi model regresi logistik mencapai 79%, sementara K-NN mencapai 74%.

**Kata kunci:** : Diabetes, K-NN, Regresi Logistik, Pembelajaran Mesin, Klasifikasi, CRISP-DM

**Citation Format:** Lo, A.A.P., & Tjioe, V.J.E. (2024). Penerapan Model CRISP-DM untuk Prediksi Penyakit Diabetes Menggunakan Metode K-Nearest Neighbor dan Logistic regression. *Prosiding SENAM 2024: Seminar Nasional Sistem Informasi & Informatika Universitas Ma Chung*. 4, 48-57. Malang: Ma Chung Press.

## PENDAHULUAN

Diabetes adalah penyakit metabolik yang ditandai dengan peningkatan kadar gula darah. Glukosa merupakan sumber energi utama bagi sel tubuh manusia. Akan tetapi, pada penderita diabetes, glukosa tersebut tidak dapat digunakan oleh tubuh. Kadar gula (glukosa) dalam darah dikendalikan oleh hormon insulin yang diproduksi pankreas. Insulin merupakan hormon yang mengatur metabolisme karbohidrat, protein, dan lemak. Termasuk mengendalikan kadar gula darah dalam tubuh agar tetap normal. Namun, pada penderita diabetes, pankreas tidak mampu memproduksi insulin sesuai kebutuhan tubuh. Tanpa insulin, sel-sel tubuh tidak dapat menyerap dan mengolah glukosa menjadi energi. "*Tingkat cepat diabetes berkembang tak hanya mengkhawatirkan, tapi juga menantang*

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 48-57, 2024
ISSN: 2808-7771 online

*sistem kesehatan global, terutama bagaimana penyakit ini meningkatkan risiko penyakit jantung iskemik dan stroke*" kata Liane Ong, ahli kesehatan di IHME Fakultas Kedokteran Universitas Washington.

Data dari *International Diabetes Federation* (IDF) menunjukan jumlah penderita diabetes di dunia pada tahun 2021 mencapai 537 juta. Angka ini diprediksi akan terus meningkat mencapai 643 juta di tahun 2030 dan 783 juta pada tahun 2045. Perhitungan terbaru dan terlengkap menunjukkan tingkat prevalensi global diabetes saat ini mencapai 6,1 persen. Hal ini menjadikan diabetes sebagai salah satu dari 10 penyebab utama kematian dan kecacatan.

Dari kasus tersebut untuk membantu memprediksi penyakit diabetes, maka akan dianalisa menggunakan model CRISP-DM dan dua metode yaitu *K-Nearest Neighbor* dan logistic regression yang merupakan algoritma *supervised learning.* Keduanya digunakan untuk membangun model prediktif berdasarkan data historis yang telah diberi label, sehingga dapat memprediksi apakah seorang pasien mungkin menderita diabetes berdasarkan variabel tertentu seperti kadar glukosa darah, level insulin, BMI, usia, dan sebagainya. K-NN mengklasifikasikan data baru berdasarkan kedekatan dengan data yang sudah ada dalam dataset. Sedangkan, logistic regression memprediksi probabilitas kejadian biner (misalnya, 1 atau 0 apakah pasien mengalami diabetes atau tidak). Bahasa pemrograman yang digunakan untuk menganalisis dan memvisualisasikan *dataset* tersebut adalah Python. Python merupakan sebuah bahasa pemrograman yang dikembangkan oleh Guido van Rossum pada akhir tahun 1980an (Kurnia, 2022). Python adalah bahasa umum yang secara luas digunakan oleh administrator sistem, pengembang web sebagai alat untuk membuat situs web dinamis, dan oleh ahli bahasa untuk tugas pemrosesan bahasa alami.

*Dataset* yang digunakan pada penelitian ini berasal dari The National Institute of Diabetes and Digestive and Kidney Diseases yang dapat diakses melalui *Kaggle*. *Dataset* ini berisi 768 data dengan 8 atribut kondisi medis pasien yang berbeda.

## MASALAH

Masalah yang dihadapi adalah meningkatnya jumlah penderita diabetes di seluruh dunia yang diperkirakan akan terus bertambah, sebagaimana diprediksi oleh International Diabetes Federation (IDF). Hal ini menciptakan tantangan signifikan dalam bidang kesehatan masyarakat global. Untuk mengatasi tantangan ini, diperlukan upaya yang lebih efektif dalam memprediksi dan mencegah diabetes. Salah satu cara yang dapat dilakukan

adalah dengan menggunakan teknologi *machine learning* untuk membangun model prediksi yang akurat. Model ini dapat membantu mengidentifikasi individu yang berisiko tinggi menderita diabetes sehingga langkah-langkah pencegahan dapat diambil lebih awal. Dengan demikian, diharapkan angka prevalensi diabetes dapat ditekan dan kualitas hidup penderita diabetes dapat ditingkatkan.

## KAJIAN PUSTAKA

### Studi Terdahulu

Penelitian yang dilakukan oleh Utami (2021) membahas perbandingan tiga metode yaitu algoritma neural network, naive bayes, dan logistic regression untuk memprediksi penyakit diabetes dengan melihat metode mana yang memiliki *accuracy* tertinggi. Hasil studi menyatakan bahwa algoritma logistic regression memiliki *accuracy* paling tinggi yaitu 75.78%, diikuti naive bayes yaitu 74.87% dan neural network yaitu 69.27%. Dengan demikian, metode logistic regression merupakan metode yang baik untuk memprediksi secara dini diagnosis penyakit diabetes.

Penelitian yang dilakukan oleh Widodo (2021) membahas metode mana yang memiliki tingkat akurasi tertinggi diantara K-NN, J48, Naive Bayes, dan logistic regression untuk memprediksi apakah individu merupakan penderita diabetes atau tidak. Hasil studi menyatakan bahwa K-NN memiliki performa yang paling baik dalam menentukan diagnosis penyakit diabetes yaitu dengan *accuracy* sebesar 98% dibandingkan algoritma lainnya.

### Metode Analisis

K-NN adalah algoritma non-parametrik yang digunakan untuk klasifikasi dan regresi. Penelitian oleh Argina (2020) menyatakan metode K-NN memiliki kelebihan dan kekurangan. Kelebihan dari K-NN dapat dilihat dari bagaimana metode tersebut mengolah objek *dataset*. Logistic regression merupakan salah satu algoritma *machine learning* populer yang digunakan untuk masalah klasifikasi. Penelitian oleh Nusinovici (2020) menyatakan logistic regression memberikan performa yang setara dengan model *ML* untuk memprediksi risiko penyakit kronis utama dalam sebuah studi epidemiologi dengan ukuran sampel sedang yang merupakan ciri khas dari banyak studi dengan jumlah kejadian insiden yang terbatas dan sejumlah prediktor klinis sederhana. Di antara berbagai model yang berbeda, logistic regression memiliki performa terbaik untuk prediksi risiko *CKD* dan *DM*. *Cross Industry Standard Process for Data Mining* (CRISP-DM) yang dikembangkan tahun

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 48-57, 2024
ISSN: 2808-7771 online

1996 oleh analisis dari beberapa industri seperti standarisasi Daimler Chrysler (Daimler-Benz), SPSS, NCR. CRISP-DM menyediakan standar proses data *mining* sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian (Larose, 2006). CRISP-DM memiliki 6 tahapan yaitu *Business Understanding*, Data *Understanding*, Data *Preparation*, *Modeling*, *Evaluation*, dan *Deployment. Confusion Matrix* digunakan sebagai pengukur kinerja setelah mengolah data *mining* dengan model klasifikasi. Pada dasarnya, pengukuran dengan *confusion matrix* digunakan untuk memberikan informasi perbandingan dari hasil klasifikasi yang dilakukan oleh algoritma yang digunakan dengan hasil klasifikasi sebenarnya.

## METODE PELAKSANAAN

Untuk memprediksi penyakit diabetes pada total 768 data yang tersedia di *dataset* digunakan penerapan model CRISP-DM dan dua algoritma *machine learning* yaitu *K-Nearest Neighbor (*K-NN*)* dan logistic regression. *Dataset* ini bertujuan untuk memprediksi berdasarkan pengukuran diagnostik apakah seorang pasien menderita diabetes. Langkah-langkah yang harus dilakukan untuk menganalisis data adalah sebagai berikut:

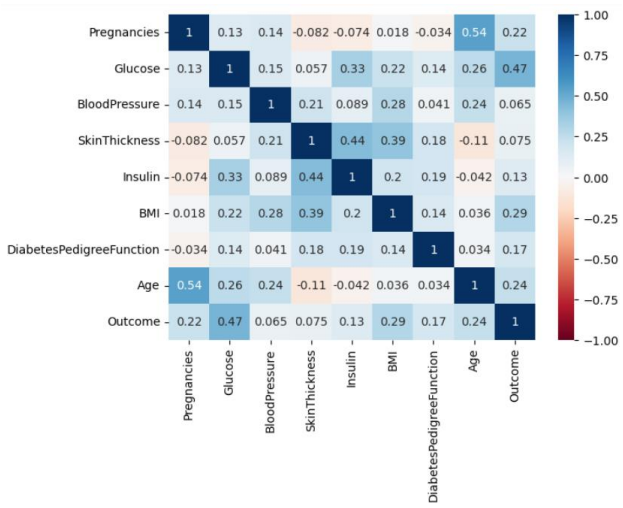
### *Business Understanding*

Tahap pertama dalam proses analisis data adalah memahami konteks bisnis dan tujuan analisis. Pemahaman ini membantu menentukan jenis data yang diperlukan dan bagaimana data tersebut akan digunakan untuk menjawab pertanyaan bisnis, membuat prediksi, atau mengembangkan model. *Dataset* merupakan kumpulan data yang telah disusun berasal dari informasi masa lalu dan siap untuk dikelola menjadi sebuah informasi baru. Contoh *repository public* yang menyediakan *dataset* untuk dianalisa oleh para peneliti adalah *Kaggle*, *UCI Machine Learning Repository, data.gov, Zdataset*, dan lain-lain. Flach (2012) mengatakan kualitas kumpulan data sangat penting untuk keberhasilan *machine learning* atau ilmu data apa pun. *Dataset* dapat digunakan untuk menjawab pertanyaan, membuat prediksi, dan membuat model. *Dataset* yang dipilih berasal dari *Kaggle* yang merupakan *multivariate dataset. Multivariate dataset* adalah *dataset* yang memiliki tiga atau lebih variabel.

### *Data Understanding*

### *Correlation Matrix*

Sebelum menentukan variabel yang akan dianalisis, diperlukan *correlation matrix* untuk melihat korelasi antar variabel dalam *dataset*. Variabel yang dipilih adalah variabel yang mendekati angka 1 atau -1 yang menunjukkan hubungan kuat antar variabel. *Correlation matrix* dapat dilihat di bawah ini:

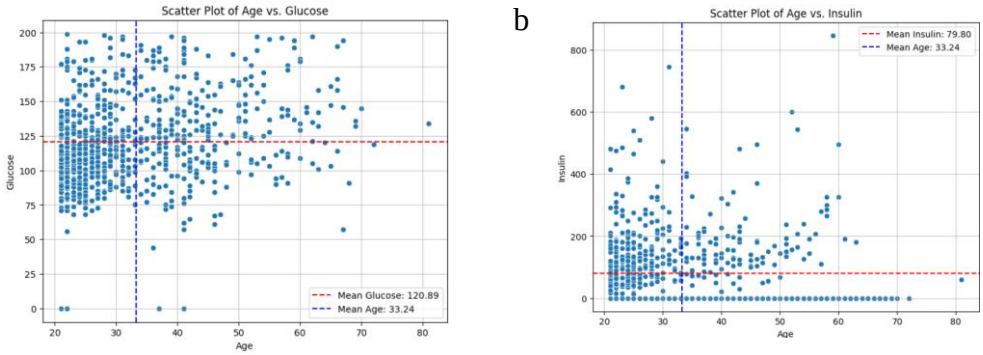



**Gambar 1.** *Correlation Matrix*

### Visualisasi Data

Proses visualisasi data diperlukan agar *dataset* lebih mudah untuk dipahami. Penggunaan jenis visualisasi mempengaruhi informasi yang ingin diketahui dari *dataset*. *Scatter plot* merupakan salah satu jenis visualisasi data yang dapat menunjukkan hubungan antar variabel. Untuk *dataset* ini, semua variabel akan dibandingkan dengan usia. Usia merupakan faktor penting dalam kesehatan karena berkaitan dengan penyakit dan kondisi medis. Berikut visualisasi menggunakan *scatter plot*:

a b




**Gambar 2.** (a) *Scatter plot of Age vs Glucose* ; (b) *Scatter plot of Age vs Insulin*

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 48-57, 2024
ISSN: 2808-7771 online

Dari kedua visualisasi di atas, dapat dilihat rata-rata (*mean*) dari masing-masing variabel yang mana akan membantu dalam memahami karakteristik variabel yang diteliti. Sebagian besar individu dalam *dataset* ini memiliki usia dibawah 50 tahun. Terdapat variabilitas yang signifikan dalam tingkat glukosa dan level insulin di berbagai usia. Visualisasi ini juga dapat digunakan untuk mengidentifikasi pola atau anomali dalam data diabetes terkait usia.

***Data Preparation***

*Dataset* yang diunduh dari *Kaggle* akan didapatkan dalam bentuk .CSV lalu akan di upload ke *google sheets* agar dapat diakses oleh *Jupyter Notebook* dan *Google Colab*. Terdapat 768 data pasien dengan 9 variabel. Hasil dari *load dataset* akan menampilkan 5 data awal dengan menggunakan *df.head()*. Nilai yang hilang dapat mempengaruhi hasil analisis dan kinerja model *machine learning*. Untuk mengetahui letak nilai yang hilang, diperlukan pengecekan kelengkapan semua data. Ada beberapa langkah untuk menghilangkan nilai yang hilang seperti menghapus baris atau kolom atau mengganti nilai yang hilang dengan rata-rata dari data. Untuk *scaling* digunakan *Min-Max scaling* yaitu mengubah nilai sehingga berada antara 0 dan 1. Karena semua variabel sudah berupa angka maka tidak diperlukan lagi transformasi data. Langkah selanjutnya adalah menentukan variabel independen dan dependen dari variabel yang digunakan. *Dataset* ini terdiri dari delapan variabel independen dan satu variabel target (dependen). Variabel yang digunakan akan dijelaskan tabel berikut:

**Table 1.** Variabel yang digunakan dalam penelitian

| Variabel | Keterangan | Tipe data |
|---|---|---|
| *Glucose* | Konsentrasi glukosa plasma selama dua jam dalam tes toleransi glukosa oral (mg/dL) | *Ratio* |
| Insulin | Insulin serum dua jam (IU/mL) | *Ratio* |
| BMI | Indeks massa tubuh (kg/m2) | *Ratio* |
| *DiabetesPedigreeFunction* | Fungsi yang menilai kemungkinan diabetes berdasarkan riwayat keluarga | *Ratio* |
| *Age* | Usia | *Ratio* |
| *Outcome*/Target | Hasil berupa : 0 jika tidak diabetes, 1 jika diabetes | Nominal |

Dengan memisahkan *training* data dan *test* data, kita dapat mendapatkan estimasi yang lebih realistis. Untuk penelitian ini, 20% dari total data akan digunakan sebagai *test*

data dan 80% akan digunakan sebagai *train* data. Random state ditetapkan 0 agar setiap kali kode dijalankan akan didapatkan pembagian data yang sama.

***Modeling***

Setelah dataset dibagi menjadi *train* data dan *test* data, langkah berikutnya adalah membuat model. Penelitian ini menganalisis *dataset* dengan dua model yang harus dibuat yaitu K-NN dan logistic regression dengan menggunakan *library scikit-learn* yang sudah di *import*. Melakukan perbandingan hasil antara *testing* data dan hasil prediksi baik menggunakan model yang sudah dilatih menggunakan algoritma *machine learning* K-NN dan logistic regression untuk memprediksi target pada *testing dataset*.

***Evaluation***

*Classification report* memberikan ringkasan evaluasi model klasifikasi data yang diprediksi. *Confusion matrix* adalah alat evaluasi kinerja klasifikasi dengan memberikan gambaran menyeluruh tentang hasil prediksi model. *Confusion matrix* membantu analisis data, sejauh mana model mengenali dan mengklasifikasikan dengan benar setiap kategori dalam data. Langkah terakhir adalah membandingkan antara logistic regression dan K-NN untuk menentukan model yang lebih superior.

***Deployment***

Hasil penelitian menunjukkan bahwa metode logistic regression adalah metode yang lebih baik dalam memprediksi kemungkinan seseorang menderita diabetes. Hal ini bermanfaat untuk memprediksi apakah seseorang menderita diabetes atau tidak. Misalnya seorang wanita berumur 33 tahun dengan kadar gula darah 121 mg/dL, insulin 80 µU/ml, BMI 32 kg/m², dengan *diabetes pedigree function* sebesar 0.472. Dengan memanfaatkan hasil logistic regression, diperoleh bahwa orang tersebut tidak menderita diabetes.

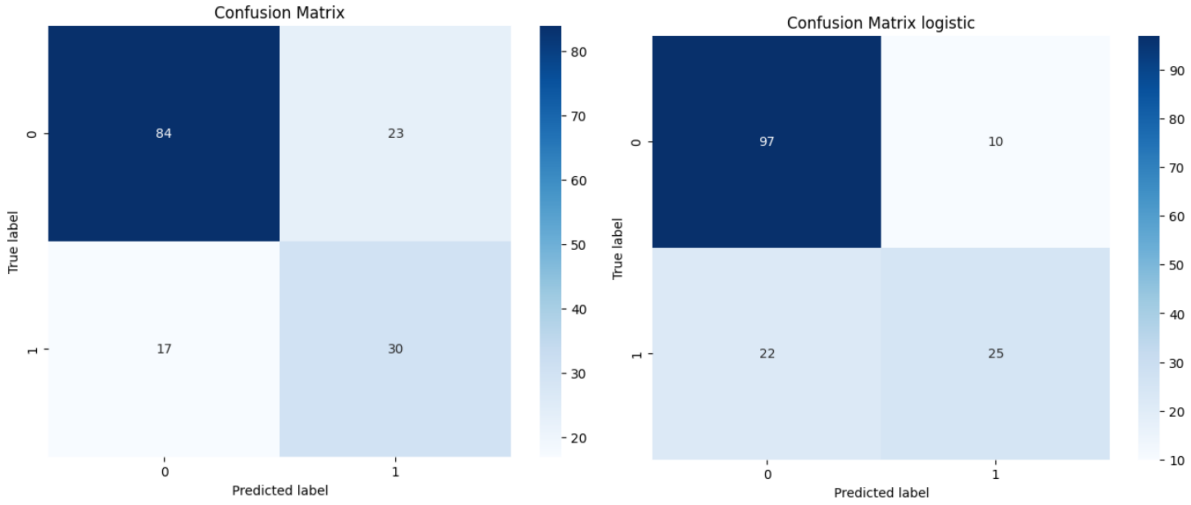
## HASIL DAN PEMBAHASAN

Hasil dari evaluasi model ditampilkan dalam bentuk *confusion matrix* dan *evaluation metrics*, yang memberikan gambaran tentang performa model klasifikasi. Hasil *confusion matrix* dapat dilihat di gambar bawah ini:

***Confusion Matrix***

*Confusion matrix* akan menunjukkan seberapa baik model klasifikasi melakukan prediksi. *Confusion matrix* terdiri dari empat sel, yang mewakili kemungkinan hasil prediksi dari proses klasifikasi diatas. Berikut hasil *confusion matrix* dari kedua metode:

Prosiding SENAM 2024:
Sistem Informasi & Informatika Vol. 4
pp. 48-57, 2024
ISSN: 2808-7771 online




a b

**Gambar 3.** (a) *Confusion matrix* logistic regression ; (b) *Confusion matrix* K-NN

Model logistic regression memiliki performa yang lebih baik dibandingkan K-NN berdasarkan *confusion matrix*. Dengan jumlah *True Positive* (TP) yang lebih tinggi logistic regression lebih akurat dalam mengidentifikasi kasus positif sebenarnya dan mengurangi kesalahan dalam mengklasifikasikan kasus positif sebagai negatif. Hal ini mengindikasikan bahwa logistic regression memberikan prediksi yang lebih konsisten dan tepat dibandingkan dengan model K-NN. *Evaluation metrics* dapat dilihat di bawah ini:

**Tabel 2.** Perbandingan *Evaluation metrics*

| | **Logistic regression** | ***K-Nearest Neighbor (KNN)*** |
|---|---|---|
| *Accuracy* | 0.79 | 0.74 |
| *Precision* | 0.78 | 0.75 |
| *Recall* | 0.79 | 0.74 |
| F1-*Score* | 0.78 | 0.74 |

Dalam perbandingan hasil *evaluation metrics* antara model logistic regression dan *K-Nearest Neighbor (KNN)*, dapat dilihat bahwa logistic regression menunjukkan performa yang lebih baik dalam memprediksi kasus diabetes berdasarkan *dataset* yang digunakan. Logistic regression mencapai akurasi sebesar 0.79, sedangkan K-NN memiliki akurasi yang lebih rendah, yaitu 0.74. Hal ini menandakan bahwa logistic regression mampu memprediksi dengan benar sekitar 79% dari total kasus, sedangkan K-NN mencapai sekitar 74%. Selanjutnya, logistic regression juga menunjukkan *precision* sebesar 0.78, sedangkan K-NN memiliki *precision* 0.75. *Precision* yang tinggi menunjukkan bahwa dari semua prediksi positifnya, logistic regression cenderung lebih tepat daripada K-NN dalam mengidentifikasi kasus yang sebenarnya positif. *Recall* juga menunjukkan hasil yang lebih

baik untuk logistic regression dengan nilai 0.79, dibandingkan dengan K-NN yang memiliki nilai 0.74. *Recall* mengukur seberapa baik model dalam menangkap semua kasus positif yang sebenarnya. *F1-score*, yang merupakan *harmonic mean* dari *precision* dan *recall*, juga menunjukkan bahwa logistic regression memiliki nilai 0.78, sedangkan K-NN memiliki nilai 0.74. Secara keseluruhan, hasil ini menunjukkan bahwa logistic regression memiliki performa yang lebih unggul dalam memodelkan hubungan antara variabel input dengan prediksi diabetes dibandingkan dengan K-NN dalam konteks *dataset* yang digunakan.

## KESIMPULAN

Penelitian ini mengimplementasikan algoritma machine learning logistic regression dan K-NN untuk memprediksi penyakit diabetes dan membandingkan model klasifikasi untuk mendapatkan model dengan hasil yang lebih superior. Proses klasifikasi ini melalui proses pemisahan dataset, penyesuaian nilai yang hilang, penskalaan, dan transformasi data, pembuatan model dari training data, prediksi berdasarkan testing dataset, evaluasi model menggunakan accuracy, precision, recall, dan F1-score, serta visualisasi berupa scatter plot. Accuracy model logistic regression mencapai 79%, sementara K-NN mencapai 74%. Precision untuk logistic regression adalah 78%, sedangkan K-NN 75%. Nilai recall logistic regression juga lebih tinggi yaitu 79%, dibandingkan K-NN yang mencapai 74%. Nilai F1-score untuk logistic regression adalah 78%, sementara K-NN mencapai 74%. Dari hasil evaluasi, model logistic regression memiliki accuracy yang lebih tinggi dibandingkan dengan K-NN, sehingga disimpulkan sebagai model yang lebih superior untuk memprediksi penyakit diabetes berdasarkan dataset yang digunakan. Selain variabel yang sudah ada, dapat menjadi pertimbangan untuk menambah variabel lain yang berhubungan dengan diabetes. Sehingga, dataset menjadi lebih lengkap dan memberikan informasi lebih jelas lagi. Akan lebih bagus juga, apabila data yang dikumpulkan berasal dari kondisi diri atau lingkungan sekitar agar prediksi yang didapatkan lebih realistis. Selain itu, mempertimbangkan penggunaan metode machine learning lainnya juga bisa menjadi saran yang baik untuk meningkatkan akurasi prediksi.

## DAFTAR PUSTAKA

Argina, A. M. (2020, July 31). Penerapan metode klasifikasi K-Nearest Neighbor pada dataset penderita penyakit diabetes. *Indonesian Journal of Data and Science, 1*(2), 29. https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/11/14

ISSN: 2808-7771 online

Kurnia, R. P. (2022). Analisis rekomendasi film dari data IMDb menggunakan Python. *Journal of Information System, Computer Science and Information Technology, 3*(2), 25. https://doi.org/10.46576/device.v3i2.2698

Larose, D. T. (2014). *Discovering knowledge in data: An introduction to data mining* (2nd ed.). IEEE Computer Society. https://doc.lagout.org/Others/Data%20Mining/Discovering%20Knowledge%20in%20Data_%20An%20Introduction%20to%20Data%20Mining%20%282nd%20ed.%29%20%5BLarose%20%26%20Larose%202014-06-30%5D.pdf

Nusinovici, S. (2020, June). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology, 122*, 56–59. https://www.sciencedirect.com/science/article/abs/pii/S0895435619310194

Utami, D. Y. (2021, July 5). Comparison of neural network algorithms, Naive Bayes, and logistic regression to find the highest accuracy in diabetes. *JITE (Journal of Informatics and Telecommunication Engineering*, *5*(1), 53–64. https://doi.org/10.31289/jite.v5i1.5201

Widodo, A. M. (2021, September 25). Performansi K-NN, J48, Naive Bayes, dan regresi logistik sebagai algoritma pengklasifikasi diabetes. *Prosiding Seminar Nasional Sistem Informasi dan Teknologi (SISFOTEK)*, *5*(1), 27–33. https://www.seminar.iaii.or.id/index.php/SISFOTEK/article/view/253/223