

ANALISA AKURASI ALGORITMA LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE UNTUK PENYAKIT DIABETES

Fadly Kurniawan¹

Email: 2310031802120@sar.ac.id

Muhammad Iqbal Al Aksha²

Email: 2310031802026@sar.ac.id

Ega Ranaldi Pebriansyah³

Email: 2310031802132@sar.ac.id

Rahmaddeni⁴

Email: rahmaddeni@sar.ac.id

^{1,2,3,4}STMIK Amik Riau

ABSTRAK

Diabetes adalah suatu kondisi medis yang dapat menyerang siapa saja, mulai dari anak muda, remaja hingga orang dewasa. Dengan cara ini, diabetes adalah penyakit yang sangat menarik yang tidak dapat dikomunikasikan dan kini telah menjadi ancaman besar bagi bantuan pemerintah dunia. Investigasi ini dilakukan dengan menangani data tambahan dari indeks informasi bantuan pemerintah Diabetes Dataset yang didapat dari dataset Kaggle dan dapat diakses melalui <https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset>. Dimana informasi sebenarnya terdiri dari 769 catatan dengan sejumlah faktor indikator klinis (Kehamilan/Kehamilan, Glukosa/Glukosa, Tekanan Darah/Denyut Nadi, Ketebalan Kulit/Ketebalan Kulit, Insulin, BMI/Daftar Berat Badan, Fungsi Silsilah Diabetes/Keturunan, Usia/Umur dan Hasil/Hasil). Lalu informasi ini ditangani menggunakan strategi Support Vector Machine dan teknik Logistic Regression untuk menentukan ketepatan hasil temuan diabetes. Mengingat penemuan eksplorasi yang telah dilakukan, teknik Support Vector Machine memiliki nilai presisi yang jauh lebih baik dibandingkan dengan menggunakan Logistic Regression. Ketepatan insentif model teknik Support Vector Machine sebesar 83,11% dan ketepatan insentif Logistic Regression sebesar 77,27%. Mengingat nilai tersebut, selisih ketepatannya adalah 5,84%. Jadi ada anggapan bahwa penggunaan teknik Support Vector Machine dapat menghasilkan tingkat ketepatan yang lebih tinggi dalam kesimpulan diabetes dibandingkan dengan menggunakan Logistic Regression.

Kata Kunci: Diabetes, IDF, Database, Support Vector, Machine, Logistic Regression.

ABSTRACT

Diabetes is a medical condition that can attack anyone, from young people, teenagers to adults. In this way, diabetes is a very interesting disease that cannot be communicated and has now become a major threat to world government aid. This investigation was carried out by handling additional data from the Diabetes Dataset government aid information index which was taken from the Kaggle dataset and can be accessed via <https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset>. Where the actual information consists of 769 records with several clinical indicator factors (Pregnancy/Pregnancy, Glucose/Glucose, Blood Pressure/Pulse Rate, Skin Thickness/Skin Thickness, Insulin, BMI/Weight List, Diabetes/hereditary function, Age/Age and Results/Results). Then this information is handled using the Support Vector Machine strategy and Logistic Regression techniques to determine the accuracy of the diabetes findings. Considering the exploratory findings that have been carried out, the Support Vector Machine technique has a much better precision value compared to using Logistic Regression. The incentive accuracy of the Support Vector Machine technical model is 83.11% and the Logistic Regression incentive accuracy is 77.27%. Considering these values, the distinction in exactness is 5.84%. So there is an opinion that using the Support Vector Machine technique can produce a higher level of accuracy in diabetes conclusions compared to using Logistic Regression.

Keywords: *Diabetes, IDF, Database, Support Vector, Machine, Logistic, Regression.*

1. PENDAHULUAN

Diabetes menimbulkan kompleksitas tinggi banyak pada anggota tubuh dan umumnya menumbuhkan ancaman kematian yang tidak terduga. Potensi masalah yang dapat terjadi antara lain kegagalan ginjal, rusak saraf, aktuasi kaki, hilang penglihatan. Orang dewasa dengan diabetes juga memiliki risiko lebih besar terkena gangguan kardiovaskular dan stroke. Selama kehamilan, prevalensi diabetes akan meningkatkan risiko kematian janin dan komplikasi lainnya. (Sari & Putri, 2023)

Machine Learning adalah bagian dari pengetahuan buatan yang dapat menganalisis informasi dengan memanfaatkan dirinya sendiri. Machine Learning adalah model terukur untuk memperkirakan informasi menggunakan PC. (Anggoro & Novitaningrum, 2021). Salah satu perhitungan yang digunakan dalam Machine Learning adalah regresi strategis.

Support Vector Machine (SVM) merupakan suatu metode untuk membuat ekspektasi, baik dalam karakterisasi maupun kasus relaps. Support Vector Machine (SVM) memiliki standar dasar pengklasifikasi lurus, yaitu kasus karakterisasi spesifik yang dapat diisolasi secara langsung, akan tetapi Support Vector Machine (SVM) telah dimajukan sehingga dapat

menangani masalah tidak langsung dengan mengkonsolidasikan bagian-bagiannya (Cahyani et al., 2022). ide-ide di area kerja berlapis tinggi. di ruang berlapis tinggi, kita akan mencari hyperplane yang dapat memperluas batas antar kelas informasi.(Pratama et al., 2023)

Regresi Logistik dihitung memodelkan hubungan antara faktor reaksi yang tidak tanggung-tanggung dan kovariat. Secara khusus, terdapat kombinasi langsung antara faktor-faktor bebas dengan kemungkinan log-peluang suatu kejadian.(Apriyani & Kurniati, 2020)

2. KAJIAN PUSTAKA

Perbandingan Algoritma

Suatu strategi belajar atau penyelidikan dengan melakukan perbincangan antara sekurang-kurangnya dua objek kajian untuk menambah dan mengembangkan informasi tentang pokok bahasan yang sedang dipertimbangkan.(Maisyarah et al., 2022)

Penyakit Diabetes

Penyakit metabolik yang bisa menyerang siapa saja. Diabetes melitus (atau diabetes) adalah suatu kondisi yang menetap dan mengakar yang mempengaruhi kapasitas tubuh untuk memanfaatkan energi dari makanan.(Maulana et al., 2018)

Algoritma

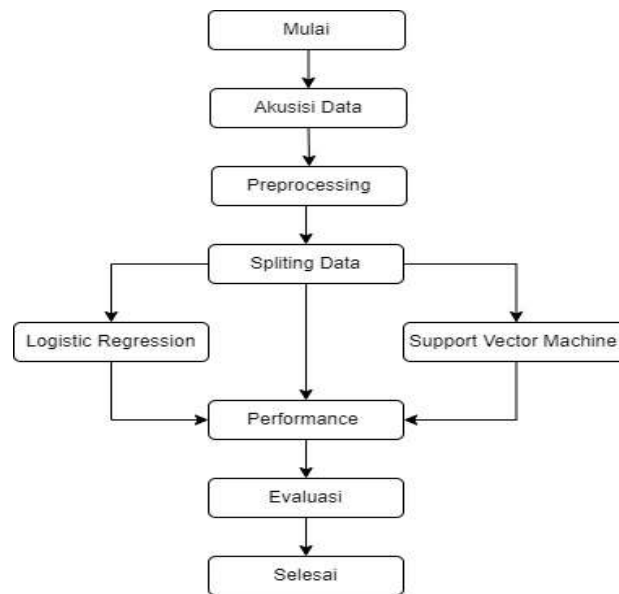
Sekumpulan arahan atau langkah yang disusun secara efisien dan digunakan untuk menangani masalah/masalah yang sah dan numerik dengan bantuan PC.(Hovi et al., 2022)

Klasifikasi

Mengumpulkan sesuatu dengan cara paling umum untuk mengenali dan mengedarkan "sesuatu" ke dalam berbagai pertemuan. Konsekuensi dari keteraturan dapat berupa bermacam-macam kelas.(Gde Agung Brahmana Suryanegara et al., 2021)

3. METODE PENELITIAN

Logistic Regression Algorithm dan Support Vector Machine Algorithm terhitung menghasilkan tingkat presisi yang unggul dalam kesimpulan diabetes pada total 769 informasi yang dapat diakses.

**Gambar 1.** Alur Penelitian

Akuisisi Data

Dataset diabetes yang dipakai ialah data yang berasal database kesehatan Diabetes Dataset yang bisa diakses melalui <https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset>. Terdapat 769 record menggunakan sebagian medis Pregnancies atau kehamilan, Glucose atau glukosa, BloodPressure atau tekanan darah tinggi, SkinThickness atau ketebalan kulit, Insulin, BMI atau indeks masa tubuh, Diabetes Pedigree Function atau keturunan, Age atau umur, serta Outcome atau hasil, yang diselesaikan menggunakan tools Python.

Tabel 1. Atribut Kumpulan Data

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Out come
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
769	1	93	70	31	0	30.4	0.315	23	0

Preprocessing

Pada tahap ini, merupakan metode penambangan informasi yang mendasari untuk

mengubah informasi kasar yang dikumpulkan dari berbagai sumber menjadi data yang lebih bersih dan dapat digunakan untuk penanganan lebih lanjut. Siklus ini juga dapat disebut sebagai langkah awal menuju pemulihan semua data yang sesuai dengan membersihkan, menyaring, dan menggabungkan informasi. 3 masalah umum yang ditangani dalam tahap prapemrosesan adalah pengelolaan kualitas yang hilang, informasi yang ramai, dan informasi yang bertentangan. Nilai yang hilang merupakan informasi yang salah dengan alasan data yang hilang menjadikan data yang terkandung di dalamnya tidak penting. Kualitas yang hilang seringkali terjadi ketika terdapat permasalahan dalam siklus pemilihan, seperti kesalahan pada bagian informasi atau permasalahan dalam pemanfaatan biometrik. Informasi keributan berisi informasi yang salah dan pengecualian yang dapat ditemukan dalam indeks informasi. Anomali dan informasi yang tidak akurat ini mengandung data yang sepele. Tujuan tertentu dari informasi yang gaduh adalah karena kesalahan manusia melalui kesalahan penandaan dan masalah lain selama pengumpulan informasi. Penyimpangan informasi terjadi ketika seseorang menyimpan catatan berisi informasi serupa di berbagai organisasi.

Splitting Data Algoritma

Pemeriksaan investigasi informasi bertujuan untuk membedah kumpulan data yang digunakan untuk merangkum atribut-atribut utama dari dataset tersebut dengan menggunakan bantuan faktual dan menyajikannya memakai strategi visual. Fase ini, data diperiksa mula bentuk dikumpulkan, sampai sebagian besar pengetahuan ekstrem diperoleh dari kumpulan data yang dapat dijangkau.(Purbolaksono et al., 2021)

Logistic Regression adalah jenis pemeriksaan regresi luar biasa yang menggunakan reaksi dan indikator ganda yang dapat berupa informasi nonstop, kelas, atau kombinasi keduanya. Pengujian ini tidak memerlukan kecurigaan terhadap dispersi multivariat biasa atau kewajaran kisi kovarians fluktuasi, dan juga dapat diterapkan pada skala informasi yang berbeda.(Soleh et al., 2021)

Support Vector Machine atau SVM adalah penghitungan AI teregulasi yang dapat dipakai untuk karakterisasi dan relaps. Cara kerjanya bergantung pada SRM atau Primary Gamble Minimization yang dimaksudkan untuk menangani informasi menjadi Hyperplane yang mencirikan ruang informasi menjadi dua kelas. Hipotesis SVM dimulai dengan pengumpulan kasus-kasus langsung yang dapat diisolasi dengan hyperplane dan dipisahkan berdasarkan kelas.

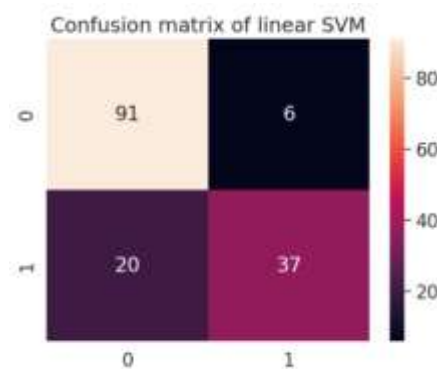
Evaluasi

Confusion Matrix ialah sebuah tabel yang berfungsi sebagaimana mestinya mengkontraskan kuantitas ekspektasi apa yang baik dan buruk di setiap kelas, sampai menyampaikan pemahaman terhadap kesalahan pada bentuk yang dihendak. Penilaian yang dipakai guna mengukur perhitungan/eksekusi bentuk adalah ketepatan, akurasi, review, dan skor F1 sebagai grid kekacauan yang telah banyak dimanfaatkan oleh berbagai spesialis.

4. HASIL DAN PEMBAHASAN

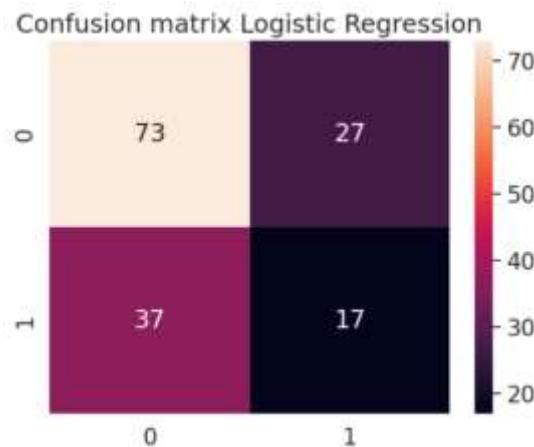
A. Hasil Penelitian

Pemeriksaan ini diharapkan mengetahui nilai presisi dan eksekusi mana yang terbaik antara teknik Support Vector Machine dan Logistic Regression dalam menyusun informasi ke makna bagian yang ditunjukkan sebelumnya. Pada tahap pendahuluan ini langkah untuk memperoleh perhitungan ketepatan adalah penyelidikan informasi, informasi yang tidak valid, membayangkan berbagai bagian di kelas dan memilih informasi dan model yang pas.



Gambar 2. Confusion Matrix Support Vector Machine

Perhitungan Disarray Lattice yang muncul karena strategi atau perhitungan Help Vector Machine, yaitu: Genuine Positive bernilai 91 untuk kelas Result, Genuine Negative bernilai 37 untuk kelas Result, Bogus Positive bernilai 6 yang untuk kelas Hasil, dan Misleading Negative bernilai 20 untuk kelas Hasil.



Gambar 3. Confusion Matrix Logistic Regression

Untuk nilai Disarray Lattice yang timbul karena strategi atau perhitungan Caled Relapse yaitu : Genuine Positive bernilai 73 dimana untuk kelas Result, Genuine Negative bernilai 17 dimana untuk kelas Result, Misleading Positive bernilai 27 untuk kelas Result dan Bogus Negative bernilai 37 untuk kelas Result.

B. Pembahasan

Analisis Logistic Regression Algorithm beserta Support Vector Machine Algorithm. Berdasarkan Penelitian yang telah dilakukan untuk menentukan keakuratan insentif diabetes adalah sebagai berikut:

```
[ ] from sklearn.metrics import accuracy_score
print("Train Accuracy of Logistic Regression ",lr.score(x_train,y_train)*100)
print("Accuracy(Test) score of Logistic Regression ",lr.score(x_test,y_test)*100)
print("Accuracy(Test) score of Logistic Regression",accuracy_score(y_test,lr_pred)*100)

Train Accuracy of Logistic Regression 77.36156351791531
Accuracy(Test) score of Logistic Regression 77.27272727272727
Accuracy(Test) score of Logistic Regression 77.27272727272727
```

Gambar 4. Akurasi Algoritma Logistic Regression

```
[ ] from sklearn.metrics import accuracy_score
print("Train Accuracy of Support Vector Machine",sv.score(x_train,y_train)*100)
print("Accuracy(Test) score of Support Vector Machine ",sv.score(x_test,y_test)*100)
print("Accuracy(Test) score of Support Vector Machine",accuracy_score(y_test,sv_pred)*100)

Train Accuracy of Support Vector Machine 81.92182410423453
Accuracy(Test) score of Support Vector Machine 83.11688311688312
Accuracy(Test) score of Support Vector Machine 83.11688311688312
```

Gambar 5. Akurasi Algoritma Support Vector Machine

C. Perbandingan Hasil Klasifikasi Algoritma**Tabel 2.** Hasil Klasifikasi Algoritma

Method	Accuracy	Recall	F1-Score
Logistic Regression	77.27%	0.73	0.74
Support Vector Machine	83.11%	0.79	0.81

5. KESIMPULAN DAN SARAN

Hasil pengkajian yang telah dikerjakan Support Vector Machine Algorithm memiliki nilai presisi yang lebih besar diperbandingkan pada Logistic Regression Algorithm. Nilai presisi untuk jenis algoritma Support Vector Machine sebesar 82,11 persen dan nilai akurasi algoritma Logistic Regression sebesar 77,27 persen. Mengikuti nilai tersebut didapat perbedaan kecermatan sebesar 5,84 persen. Maka disimpulkan maka implementasi algoritma Support Vector Machine bisa mewujudkan kualitas presisi diagnosis penyakit diabetes yang lebih bagus diperbandingkan pada algoritma Logistic Regression.

DAFTAR PUSTAKA

- Anggoro, D. A., & Novitaningrum, D. (2021). Comparison of accuracy level of support vector machine (SVM) and artificial neural network (ANN) algorithms in predicting diabetes mellitus disease. *ICIC Express Letters*, 15(1), 9–18. <https://doi.org/10.24507/icicel.15.01.9>
- Apriyani, H., & Kurniati, K. (2020). Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus. *Journal of Information Technology Ampera*, 1(3), 133–143. <https://doi.org/10.51519/journalita.volume1.issue3.year2020.page133-143>
- Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., Dwi, A., Putra, P., & Artikel, G. (2022). Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, 1(2), 2828–9099. <https://doi.org/10.55123/jomlai.v1i2.598>
- Gde Agung Brahmana Suryanegara, Adiwijaya, & Mahendra Dwifabri Purbolaksono. (2021). Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. *Jurnal RESTI (Rekayasa Sistem*

- Dan Teknologi Informasi*), 5(1), 114–122. <https://doi.org/10.29207/resti.v5i1.2880>
- Hovi, H. S. W., Id Hadiana, A., & Rakhmat Umbara, F. (2022). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM). *Informatics and Digital Expert (INDEX)*, 4(1), 40–45. <https://doi.org/10.36423/index.v4i1.895>
- Maisyarah, C., Haryatmi, E., Fajriatifah, R. Y., & Puspita, Y. H. (2022). Prediksi Penyakit Diabetes menggunakan Algoritma Artificial Neural Network. *Jurnal Data Science & Informatika*, 2(1), 46–52. <http://publikasi.bigdatascience.id/index.php/jdsi>
- Maulana, Y. I. R., Badriyah, T., & Syarif, I. (2018). Influence of Logistic Regression Models For Prediction and Analysis of Diabetes Risk Factors. *EMITTER International Journal of Engineering Technology*, 6(1), 151–167. <https://doi.org/10.24003/emitter.v6i1.258>
- Pratama, A., Nurcahyo, A. C., & Firgia, L. (2023). Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes. *Prosiding CORISINDO 2023*, 116–121.
<https://stmikpontianak.org/ojs/index.php/corisindo/article/view/30%0Ahttps://stmikpontianak.org/ojs/index.php/corisindo/article/download/30/22>
- Purbolaksono, M. D., Irvan Tantowi, M., Imam Hidayat, A., & Adiwijaya, A. (2021). Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(2), 393–399. <https://doi.org/10.29207/resti.v5i2.3008>
- Sari, A., & Putri, D. U. (2023). Penerapan Regresi Linear Berganda Untuk Memprediksi Diabetes Secara Dini. *Jurnal Teknisi*, 3(1), 32. <https://doi.org/10.54314/teknisi.v3i1.1255>
- Soleh, M., Ammar, N., & Sukmadi, I. (2021). Website-Based Application for Classification of Diabetes Using Logistic Regression Method. *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, 9(1), 23. <https://doi.org/10.24843/jim.2021.v09.i01.p03>