Terbit online pada laman web jurnal: **http://ejurnal.unim.ac.id/index.php/submit**

SUBMIT
(Jurnal Ilmiah Teknologi Informasi dan Sains)
Vol. 4 No. 1(2024)19-23 ISSN Media Elektronik: 2798-6861

PREDIKSI PASIEN TERINDIKASI PENYAKIT JANTUNG MENGGUNAKAN METODE LOGISTIC REGRESSION

Gita Rohma Utami Asyafiiyah[*1], Ronny Makhfuddin Akbar [2]

[1,2] Universitas Islam Majapahit
Email: [1]gitarohma7@gmail.com, [2]ronnyma.ft@unim.ac.id

(Naskah masuk: 20 Juni 2024, diterima untuk diterbitkan: 30 Juni 2024)

Abstrak

Penyakit serangan jantung (*Heart Attack*) telah terbukti sebagai salah satu penyakit berbahaya di dunia. Penyakit serangan jantung adalah kondisi dimana tersumbatnya arteri yang disebabkan oleh timbunan lemak. Deteksi penyakit jantung perlu dilakukan sejak dini karena gejala awalnya sering kali tidak jelas dan mudah terabaikan sehingga banyak orang tidak menyadari bahwa mereka sedang mengalami kondisi yang berakibat fatal nantinya. Untuk mengatasi masalah tersebut teknologi machine learning dapat digunakan untuk membantu mendeteksi penyakit jantung dengan menggunakan data historis pasien dengan berbagai metode yang ada. Pada penelitian ini metode yang digunakan adalah *logistic regression* dimana metode tersebut memodelkan probabilitas kejadian dari suatu peristiwa untuk menghasilkan nilai binary, yaitu nol dan satu sebagai penentuan klasifikasinya. Model yang diterapkan memberikan hasil akurasi pada data *training* sebesar 86% dan pada data *testing* sebesar 88%. Berdasakan hasil confusion matrix, model mampu memprediksi sampel yang benar-benar negatif (TN) dengan baik, ditunjukkan dari persentase yang cukup bagus mencapai nilai 85,39%. Kemudian proporsi false negative (FN) juga cukup rendah 9,52%. Nilai true positif (TP) mencapai 90,48% dan ROC curve menunjukkan nilai AUC 0,95 (mendekati 1) yang berarti model memiliki 95% area dibawah curva. Hal ini menunjukkan bahwa model yang dibangun memberikan performa yang baik. Jika dibandingkan dengan beberapa metode machine learning lainnya, metode *logistic regression* lebih unggul dari tingkat akurasinya.

Kata kunci: *penyakit jantung, regresi logistik, prediksi*

PREDICTION OF PATIENTS INDICATED WITH HEART DISEASE USING LOGISTIC REGRESSION METHOD

Abstract

Heart attacks have been proven to be one of the most dangerous diseases in the world. A heart attack is a condition where the arteries are blocked due to fat deposits. Detection of heart disease needs to be done early because the initial symptoms are often unclear and easily overlooked so many people do not realize that they are experiencing a condition that can be fatal later. To overcome this problem, machine learning technology can be used to help detect heart disease using historical patient data using various existing methods. In this research, the method used is logistic regression, where the method models the probability of occurrence of an event to produce binary values, namely zero and one to determine the classification. The model applied provides accuracy results on training data of 86% and on testing data of 88%. Based on the results of the confusion matrix, the model can predict truly negative (TN) samples well, as shown by a fairly good percentage reaching a value of 85.39%. Then the proportion of false negatives (FN) is also quite low, 9.52%. The true positive (TP)

value reached 90.48% and the ROC curve showed an AUC value of 0.95 (close to 1) which means the model has a 95% area under the curve. This shows that the model built provides good performance. When compared with several other machine learning methods, the logistic regression method is superior in terms of accuracy.

Keywords: *heart disease, logistic regression, prediction*

1. PENDAHULUAN

Penyakit serangan jantung (*Heart Attack*) telah terbukti sebagai salah satu penyakit berbahaya di dunia. Menurut data melalui *World Health Organization* (WHO), penyakit serangan jantung adalah kondisi dimana tersumbatnya arteri yang disebabkan oleh timbunan lemak **(Cardiovascular diseases (cvds), 2021)**. Penyumbatan ini mengurangi atau bahkan menghentikan aliran darah bagian tertentu dari otot jantung dan menyebabkan jaringan tersebut kekurangan oksigen dan nutrisi penting. Penyakit ini juga menyebabkan beberapa gejala seperti sesak napas dan nyeri dada. Data dari WHO menyebutkan bahwa lebih dari 17,8 juta orang di dunia meninggal akibat penyakit jantung dan pembuluh darah, sedangkan di Indonesia sendiri kematian akibat penyakit jantung mencapai 651.481 penduduk per tahun.

Salah satu tantangan besar yang sedang dihadapi di bidang kesehataan saat ini adalah sulitnya mendeteksi penyakit jantung sejak dini, dimana gejala awalnya sering kali tidak jelas dan mudah terabaikan. Kurangnya deteksi dini ini dapat mengakibatkan penanganan yang terlambat, sehingga meningkatnya risiko komplikasi serius dan menurunkan peluang kesembuhan. Oleh sebab itu, penting untuk mencari cara yang lebih efektif untuk mengidentifikasi penyakit jantung agar dapat memberikan intervensi medis yang tepat waktu. Solusi potensial untuk mengatasi masalah ini adalah dengan memanfaatkan penggunaan machine learning yang memiliki kemampuan untuk menganalisis gejala yang dianggap berhubungann dengan penyakit jantung. Data historis pasien dapat digunakan untuk melatih algoritma, sehingga mampu mengenali pola-pola yang mungkin tidak terdeteksi oleh metode konvensional.

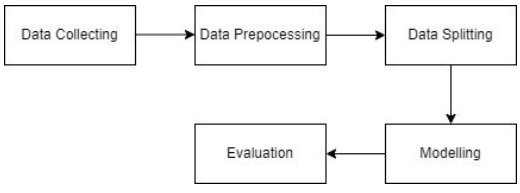
Ada banyak metode pemodelan yang dapat digunakan dalam penerapan *machine learning*, namun pada penelitian kali ini penulis menggunakan metode pemodelan *Logistic Regression*. Dalam kasus klasifikasi biner seperti ini, *logistic regression* menghasilkan fungsi logistik yang dapat memprediksi probabilitas kejadian target berdasarkan nilai-nilai variabel independen dengan dua kemungkinan hasil yaitu 0 dan 1.

Dari penelitian sebelumnya dengan judul " Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network dalam Prediksi Penyakit Jantung" oleh **(Handayani, F., 2021)** telah dilakukan perbandingan antara metode SVM, *Logistic Regression* dan ANN untuk mendeteksi penyakit jantung pada pasien menggunakan data yang sama. Jika dilihat dari nilai akurasi model menunjukkan bahwa metode *logistic regression* lebih unggul dibanding dengan metode lainnya. Pada penelitian lain dengan judul "Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor dan Logistic Regression" oleh **(Sitanggang, D. et al., 2022)** juga menunjukkan hasil akurasi bahwa metode logistic regression lebih unggul dibanding dengan metode *K-Nearest Neighbor* (KNN).

Tujuan dari penelitian ini adalah mengembangkan model prendiksi yang akurat dan andal yang dapat membantu tenaga medis dalam mengidentifikasi pasien yang berisiko tinggi terkena penyakit jantung. Hasil penelitian ini diharapkan dapat memberikan alat bantu yang efektif bagi tenaga medis dalam mendetiksi dan mengelola risiko penyakit jantung secara lebih dini dan akurat, sehingga dapat meningkatkan kualitas perawatan kesehatan dan menurunkan angka pasien yang terkena penyakit jantung.

2. METODE PENELITIAN

Pada Gambar 1 menunjukkan tahapan dari penelitian ini yang dimana terbagi menjadi 5 tahap yaitu data *collecting* proses pengambilan data, kemudian data *prepocessing* yaitu proses pembersihan data dari *missing value*, *duplicate value* dan data *outlier*, kemudian data *splitting* membagi data menjadi data *train* dan data *test*. Tahap selanjutnya yaitu *modelling* menggunakan algoritma *logistic regression* dan terakhir yaitu evaluasi terhadap model yang digunakan.



Gambar 1. Tahap Metode Penelitian

2.1. Data Collecting

Data yang digunakan dalam penelitian ini merupakan data pasien penyakit jantung yang diperoleh secara *online* melalui *website kaggle*. Fitur-fitur yang terdapat dalam dataset yaitu:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

2.2. Data Prepocessing

Tahap *prepocessing* pada penelitian ini dilakukan secara 3 tahap yaitu *detect missing value*, *detect* dan *handling duplicat value*, kemudian *detect* dan *handling* data *outlier*. Pada dataset yang digunakan tidak terdapat data yang hilang dan tidak ada duplikat data. Namun, terdapat beberapa data *outlier*, untuk menangani masalah tersebut, data *outlier* yang nilainya melebihi batas ditangani dengan cara dihapus dari data set.

2.3. Data Splitting

Pada tahap data *splitting,* dataset yang digunakan dalam penelitian ini dibagi menjadi dua yaitu data pelatihan (*training set*) dan data pengujian (*testing set*). Dalam proses pembentukan model perbandingan pembagian data *training* dan data *testing* adalah 80:20.

2.4. Modelling

Pembangunan model prediksi pada penelitian ini menggunakan model *logistic regression* dan beberapa *library* pendukung untuk machine learning yaitu *pandas, numpy, matplotlib, seaborn* dan *sklearn*.

Model *logistic regression* bekerja dengan memodelkan probabilitas kejadian dari suatu peristiwa biner sebagai fungsi dari variabel independen. Terdapat 13 variabel independen yang digunakan pada pemodelan ini dan 1 variabel dependent yang direpresentasikan dengan (1 = terkena penyakit jantung, 0 = tidak terkena penyakit jantung). Fungsi logistik yang digunakan dalam model digambarkan pada persamaan 1 dibawah ini:

$$P(Y=1|X) = \frac{1}{1+e-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)} \quad (1)$$

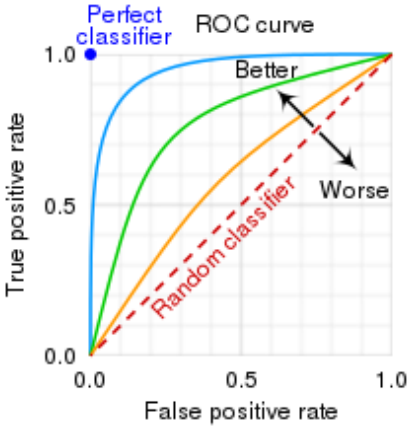
Dalam persamaan diatas :

$P(Y=1|X)$ adalah probabilitas pasien yang terindikasi terkena penyakit jantung berdasarkan variabel-variabel independen *X*. β_0 adalah konstanta intercept. $\beta_1, \beta_2, \ldots, \beta_{13}$ adalah koefisisen yang mengukur dampak dari masing-masing variabel independen.

Fungsi logistik mengonversi kombinasi linear dari variabel independen menjadi probabilitas antara 0 dan 1, yang kemudian dapat digunakan untuk menentukan klasifikasi biner. Jika probabilitas yang dihasilkan lebih besar dari 0,5 maka model akan mengklasifikasikan pengamatan sebagai kelas positif 1 dan jika tidak diklasifikasikan sebagai kelas negatif 0.

2.5. Evaluation Model

Setelah model berhasil dibangun, evaluasi dilakukan menggunakan *learning curve, confusion matrix*, dan ROC *curve*. *Learning curve* merupakan kurva yang menunjukkan bagaimana performa model berubah seiring bertambahnya jumlah data pelatihan yang digunakan dalam melatih model dan berguna untuk mengidentifikasi apakah model mengalami *overfitting* atau *underfitting* .*Confusion matrix* digunakan untuk memberi gambaran yang lebih rinci mengenai performa model dengan menghitung jumlah *true positives* (TP), *true negatives* (TN), *false positives* (FP) dan *false negtives* (FN). Sedangkan ROC *curve* digunakan untuk mengevaluasi model dalam membedakan *true positive rate* (TPR) dan *false positive rate* (FPR). Dimana jika kurva semakin mendekat ke sudut kiri atas (*perfect classifier*) menunjukkan kinerja yang semaik baik.



Gambar 2. ROC Curve

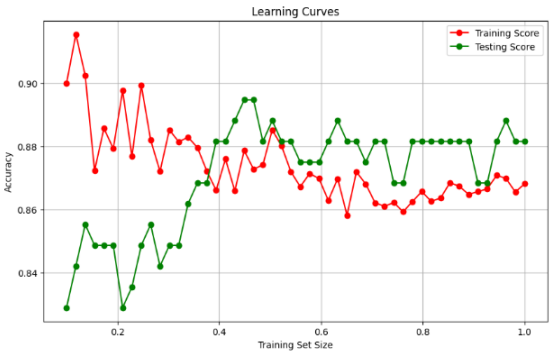
3. HASIL DAN PEMBAHASAN

Pembangunan model dilakukan menggunakan aplikasi Google Colab dengan detail spesifikasi yang dapat dilihat pada Tabel 1.

Tabel 1. Spesifikasi Sistem

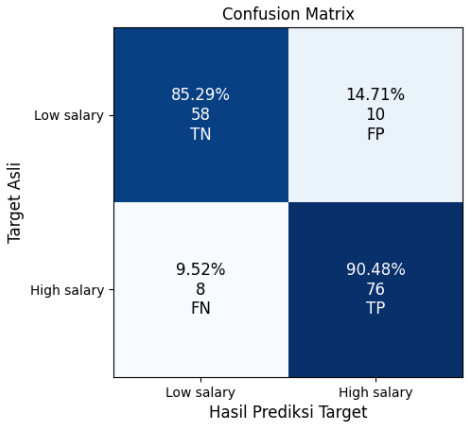
No	Jenis Kebutuhan	Spesifikasi
1	Sistem Operasi	Windows 10
2	Bahasa Pemrograman	Python 3.10
3	Library Pendukung	Pandas 2.0.3 Numpy 1.25.2 Matplotlib 3.7.1 Seaborn 0.13.1 Scikit-learn 1.2.2

Setelah dilakukan proses training dan testing pada model *Logistic Regression* yang telah dibuat diperoleh nilai akurasi pada data *training* sebesar 86% dan pada data *testing* sebesar 88%. Kemudian jika dilihat pada hasil *learning curve* pada Gambar 3, pada awalnya, grafik menunjukkan bahwa akurasi data *training* cenderung lebih tinggi daripada data *testing*. Namun seiring bertambahnya ukuran set data *training* akurasi pengujian meningkat dan menjadi lebih stabil.



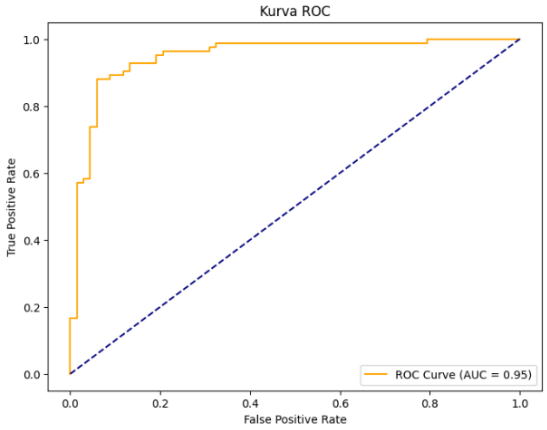
Gambar 3. Learning Curve

Evaluasi lain juga dilakukan menggunakan *confusion matrix* dan ROC *curve*.



Gambar 4. Confusion Matrix

Dilihat dari hasil confusion matrix pada Gambar 4, model mampu memprediksi sampel yang benar-benar negatif (TN) dengan benar dengan persentase yang cukup bagus mencapai nilai 85,39%. Kemudian proporsi false negative (FN) juga cukup rendah 9,52%. Nilai true positif (TP) mencapai 90,48% yang berarti dari semua sampel yang benar-benar positif diprediksi dengan benar sebagai positif oleh model. Dan nilai false positif juga cukup rendah yaitu 14,71%. Sedangkan dari hasil ROC curve pada Gambar 5 menunjukkan bahwa model yang dibuat memberikan performa yang cukup baik dengan nilai AUC 0,95 (mendekati 1) yang artinya model memiliki 95% area dibawah curva.



Gambar 5. Hasil ROC Curve

4. KESIMPULAN

Hasil analisa dari prediksi penyakit jantung menggunakan metode logistic regression dapat diambil kesimpulan sebagai berikut:

- Model yang telah dibuat menunjukkan performa yang cukup baik dan konsisten, dengan nilai akurasi sebesar 86% pada data pelatihan dan 88% pada data pengujian. Hal tersebut menunjukkan bahwa model tidak mengalami *overvitting* atau *underfitting* secara signifikan.
- Model *Logistic Regression* terbukti menjadi yang terbaik dalam memprediksi penyakit jantung dibandingkan metode *machine learning* yang lain. Namun hasil tersebut masih perlu dievaluasi lebih lanjut karena setiap metode memiliki keunggulan masing-masing. Beberapa metode yang dapat digunakan adalah decision tree dan random forest, karena kedua metode tersebut memberikan pemisahan fitur yang lebih rinci sehingga dapat membantu mesin mempelajari data dengan lebih baik.

DAFTAR PUSTAKA

Handayani, F. (2021) 'Komparasi Support Vector Machine, logistic regression Dan Artificial Neural Network dalam prediksi penyakit jantung', *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, 7(3), p. 329. doi:10.26418/jp.v7i3.48053.

Pangaribuan, J.J., Tanjaya, H. and Kenichi (2021) 'MENDETEKSI PENYAKIT JANTUNG MENGGUNAKAN MACHINE LEARNING DENGAN ALGORITMA LOGISTIC REGRESSION', Journal Information System Development (ISD) , 6(2).

Sitanggang, D. et al. (2022) 'Implementasi data mining untuk Memprediksi Penyakit Jantung Menggunakan metode k-nearest neighbor Dan logistic regression', Jurnal Teknik Informasi dan Komputer (Tekinkom), 5(2), p. 493. doi:10.37600/tekinkom.v5i2.698.

Lapp, D. (2019) Heart disease dataset, Kaggle. Available at: https://www.kaggle.com/datasets/johnsmith88/heart-disease dataset/code?datasetId=216167&outputs=Visualization (Accessed: 12 June 2024).

Apa Itu Regresi Logistik? (2024) *IBM*. Available at: https://www.ibm.com/id-id/topics/logistic-regression (Accessed: 12 June 2024).

Cardiovascular diseases (cvds) (2021) *World Health Organization*. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (Accessed: 12 June 2024).

Muralidhar, K. (2023) *Learning curve to identify overfitting and underfitting in machine learning*, *Medium*. Available at: https://towardsdatascience.com/learning-curve-to-identify-overfitting-underfitting-problems-133177f38df5 (Accessed: 12 June 2024).