Memprediksi Gagal Jantung Menggunakan Metode Klasifikasi Machine Learning

Billy Renatasiva[1*], Jehuda Rivaldo Soetiyono[2],Putri Novi Ramadayati[3]

[1*]Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur
[2] Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur
[3] Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur
Email: [1*]s160419131@student.ubaya.ac.id, [2] s160419133@student.ubaya.ac.id, [3] s160419148@student.ubaya.ac.id

(Naskah masuk: dd mmm yyyy, direvisi: dd mmm yyyy, diterima: dd mmm yyyy)

Abstrak

Jantung adalah organ yang sangat penting pada tubuh manusia. Jantung berfungsi untuk memompa darah keseluruh tubuh, jantung juga memiliki peranan penting untuk mengantarkan oksigen keseluruh tubuh agar tubuh berfungsi dengan baik dan tidak terjadi kesalahan yang fatal. Penelitian yang kami lakukan dalam hal ini mungkin bisa bermanfaat untuk hidup pasien yang menderita masalah jantung. dengan menggunakan metode klasifikasi yang ada di machine learning, data pasien dapat diklasifikasikan. metode yang digunakan dalam penelitian ini yaitu KNN, ANN, Naive Bayes, dan Logistic Regression, dari beberapa metode di atas, penelitian ini mencoba untuk membandingkan metode mana yang menghasilkan akurasi tertinggi. dalam penelitian ini juga membandingkan PCA dan non PCA untuk menemukan model mana yang menghasilkan akurasi yang tettinggi. Hasil akhir dari penelitian ini menunjukan bahwa model dengan menggunakan klasifikasi Artificial Neural Network dengan PCA menghasilkan akurasi yang paling tinggi dibandingkan metode lainnya.

Kata Kunci: Jantung, Machine Learning, klasifikasi, memprediksi

Predicting Heart Failure Using Classification Method in Machine learning

Abstract

The heart is a very important organ in the human body. The heart functions to pump blood throughout the body, it has an important role to deliver oxygen throughout the body so that the body can function properly. The research we do may be beneficial to the lives of patients suffering from heart problems. By using the classification method in machine learning, data that are collected from patients can be processed further to predict life expectancy. The methods used in this research are K-Nearest Neighbors, Artificial Neural Network, Gaussian Naive Bayes, and Logistic Regression. From the above methods, this research tries to compare which method will produce highest accuracy. This research also uses PCA to see which model produce highest accuracy. The final result of this study shows that the model using the Artificial Neural Network classification with PCA produces the highest accuracy compared to other methods.

Keywords: *Heart, Machine Learning, Classification, predict*

I. PENDAHULUAN

Penyakit jantung merupakan masalah terbesar yang sedang dihadapi dunia dalam hal kesehatan. Penyakit jantung sendiri merupakan salah satu penyakit yang mempunyai kasus kematian tertinggi didunia. WHO sendiri mengeklaim bahwa 85% kematian di dunia terjadi karena penyakit gagal janutng[1]. Di Indonesia serangan jantung juga termasuk dalam penyakit yang sangat sering ditemui, tetapi masih banyak pasien-pasien yang jantungnya bermasalah telat ditangani. Tidak sedikit juga yang mengalami serangan jantung/gagal jantung secara mendadak yang disebut dengan jantung koroner, jantung koroner ini sendiri biasanya bisa tidak bisa sembuh tetapi bisa dikurangi gejalanya dengan mengkonsumsi obat-obatan yang dianjurkan oleh dokter.

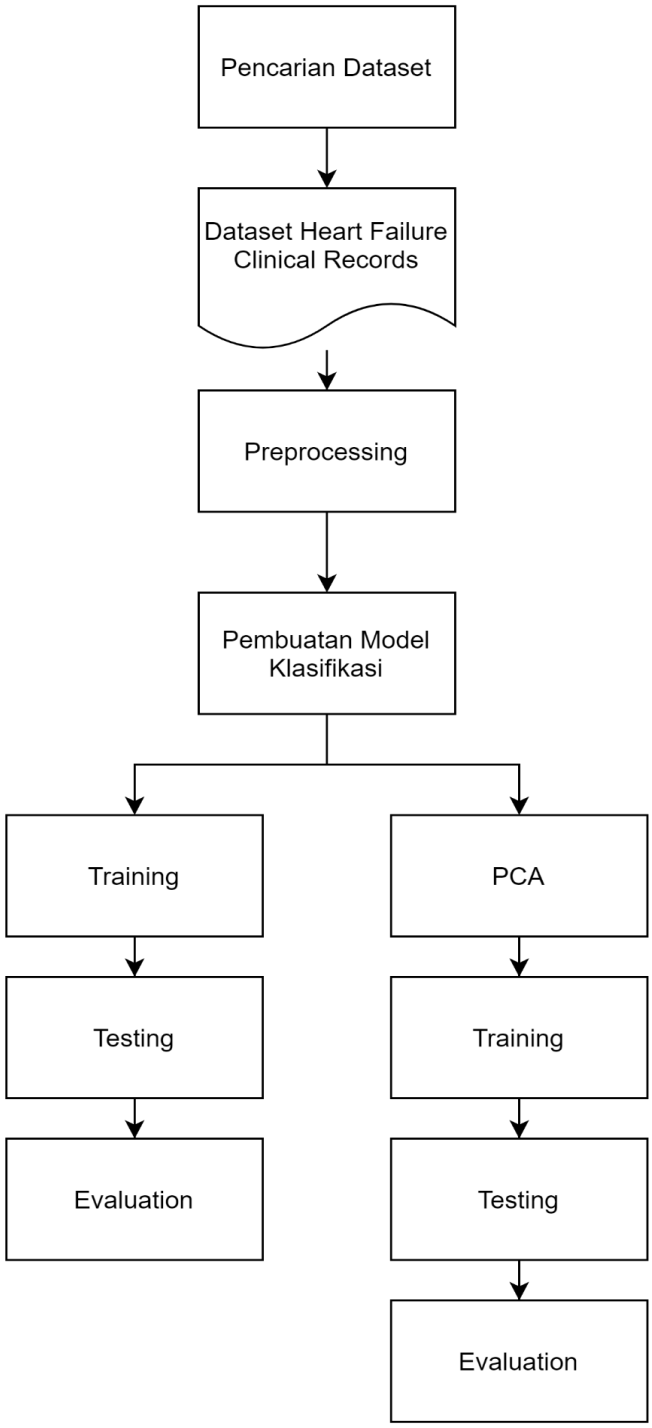
Banyak di Indonesia serangan gagal jantung ini yang tidak terdeteksi, jadi banyak pasien penyakit jantung yang sebelumnya tidak terdeteksi mengalami serangan jantung / gagal jantung mendadak dan tidak sedikit pula yang mengalami kematian secara mendadak[2]. Ada beberapa pengobatan yang disarankan oleh dokter untuk menerapi penyakit jantung ini contohnya seperti pemberian obat untuk meringankan gejala dan mengobati jantungnya, melakukan operasi bypass / menggunakan memasang ring, dan jika ada case yang buruk bisa sampai memasang semacam alat di jantung pasien. Dari beberapa terapi yang digunakan tadi kita harus melakukan pencegahan sejak dini dengan pola hidup sehat, mengurangi konsumsi makanan yang tidak baik untuk jantung, dan melakukan analisa gejala untuk mengurangi terjadinya penyakit ini. Dalam analisa penyakit jantung dapat menggunakan metode-metode yang terdapat di machine learning untuk menganalisa apa saja yang dapat di cegah dalam penanganan penyakit jantung ini.[3]

Pada saat ini pemakaian metode-metode yang ada di machine learning sangat membantu manusia dalam menganalisa dan memprediksi suatu masalah itu terjadi. Pada kasus ini machine learning diharapkan dapat membantu menganalisa dan memprediksi data pasien untuk memprediksi adanya gejala atau kelainan yang ada di jantung pasien agar cepat ditangani. Metode yang digunakan dalam kasus ini adalah Logistic Regression, KNN, ANN, Gaussian. Dalam penelitian ini dataset yang digunakan berasal dari kaggle.[4]

Penelitian kasus ini bertujuan untuk mencoba menanggulangi masalah penyakit gagal jantung ini. Dari metode-metode yang sudah dijelaskan akan diambil tingkat akurasi tertinggi untuk menghasilkan output yang mendekati akurasi sempurna. Saat metode-metode yang sudah disebutkan menemukan tingkat akurasi yang tinggi maka kemungkinan penelitian ini dapat membantu untuk melihat seberapa banyak orang yang dibilang akan terkena penyakit gagal jantung ini.

II. METODOLOGI PENELITIAN

Metodologi penelitian merupakan sebuah cara untuk mengetahui hasil dari sebuah permasalahan yang spesifik, dimana permasalahan tersebut disebut juga dengan permasalahan penelitian [5]. Pada metodologi penelitian kami menggunakan beberapa tahapan yaitu, dimulai dari mencari dataset yang sesuai dengan kasus penelitian, lalu melakukan tahap preprocessing. Setelah melakukan tahap preprocessing data akan masuk ke tahap pembuatan model klasifikasi. dalam proses pembuatan model klasifikasi dalam penelitian ini menggunakan Logistic Regression, ANN, Naive Bayes dan KNN. Setelah melakukan pembuatan model klasifikasi data akan masuk ke tahap training, setelah melakukan training data selanjutnya akan dilakukan testing pada model klasifikasi yang sudah di training. Tahap terakhir adalah melakukan evaluation model.



Gambar 1. Metodologi Penelitian

A. Pencarian Dataset

Dataset yang digunakan dalam penelitian ini adalah adalah *heart failure clinical records dataset* [6]. Dataset yang kami gunakan memiliki 12 atribut yang dapat digunakan untuk memprediksi kematian akibat gagal jantung. Berikut adalah tabel data atribut dari *heart failure clinical records dataset* pada Tabel 1

Tabel 1. Deskripsi Data

Nama Atribut	Tipe Data	Deskripsi
Age	Kontinu	Umur pasien (Tahun)
Anemia	Biner	Apakah pasien mengidap Anemia (0:Tidak, 1:Ya)
creatinine_phosphokinase	Kontinu	Level ensim CPK di dalam darah pasien (mcg/L)
diabetes	Biner	Apakah pasien mengidap Diabetes (0:Tidak, 1:Ya)
ejection_fraction	Kontinu	Persentase darah meninggalkan jantung tiap-tiap kontraksi (%)
high_blood_pressure	Biner	Apakah pasien mengidap hipertensi (0:Tidak, 1:Ya)
platelets	Kontinu	Jumlah trombosit dalam darah (kiloplatelets/mL)
serum_creatinine	Kontinu	Level kreatinin serum dalam darah pasien (mg/dL)
serum_sodium	Kontinu	Level sodium serum dalam darah pasien (mEq/L)
sex	Biner	Kelamin pasien (0:Tidak, 1:Ya)
smoking	Biner	Apakah pasien merokok atau tidak (0:Tidak, 1:Ya)
time	Kontinu	Interval waktu tindak lanjut (hari)
DEATH_EVENT	Biner	Jika pasien meninggal selama masa tindak lanjut (0:Tidak, 1:Ya)

B. Preprocessing

Data *preprocessing* adalah teknik yang digunakan untuk mengubah data mentah dalam format yang berguna dan efisien [7]. Pada penelitian ini kami menggunakan teknik *preprocessing* min max scaling. Pada Gambar 1 adalah rumus yang kami gunakan pada tahap *preprocessing*,

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Gambar 2. Rumus MinMaxScaling

x = nilai tertentu dalam data
min(x) = nilai minimum data
max(x) = nilai maximum dat

C. PCA

PCA adalah sebuah metode yang mereduksi dimensi dengan menggunakan *principal components.* Tujuan dilakukan PCA pada penelitian ini untuk memudahkan dalam menginterpretasikan data dan melihat pembagian data.

D. Pembuatan Model Klasifikasi

1) Logistic regression

Metode Logistic Regression adalah metode untuk memprediksi suatu masalah/peristiwa yang terjadi biasanya metode ini cocok untuk melakukan prediksi untuk memprediksi bencana alam, memprediksi angka kenaikan harga barang, dll. Dalam kasus ini logistic regression diharapkan mampu untuk memprediksi pasien yang diduga memiliki penyakit jantung.

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1 X$$

Gambar 3. Rumus Logistic Regression

keterangan :
ln : Logaritma Natural (B0 + B1X)

2) ANN

Metode ANN adalah metode yang diperlihatkan seperti sistem saraf manusia yang memecahkan masalah dengan mengalir melalui jaringan-jaringan saraf tersebut. Metode ANN disini diharapkan dapat memproses data *heart failure clinical records dataset* untuk memecahkan masalah penelitian ini.

3) Naive Bayes

Naïve bayes adalah metode yang menghitung pengalaman probabilitas dari masa lalu, disini kami menggunakan metode Naïve bayes yang metode Gaussian.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Gambar 4. Rumus Gaussian Naïve Bayes

Keterangan:
P(A|B) : Probabilitas posterior kelas (A, target) yang diberikan prediktor (B, atribut).
P(A) : Probabilitas kelas sebelumnya.
P(B|A): Probabilitas kelas yang diberikan oleh prediktor.
P(B): Probabilitas sebelumnya(*prior probability*) dari prediktor.
Posterior Probability = (Conditional Probability x Prior probability)/ Evidence

4) KNN

Metode KNN adalah metode yang melakukan klasifikasi terhadap objek berdasarkan data yang jaraknya dekat dengan objeknya tersebut. Metode KNN sendiri biasanya digunakan untuk mengklasifikasikan suatu data, yang data tersebut masih belum diklasifikasikan dengan data terdekatnya. Dalam penelitian ini metode KNN diharapkan untuk bisa mengklasifikasikan pasien yang diduga mengidap penyakit jantung. Berikut adalah rumus untuk menghitung jarak dengan KNN.

Euclidean Distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

Manhattan Distance:

rumus *manhattan distance* untuk mencari jarak hanya dengan menjumlahkan semua selisih dari jarak *xi* dan *yi*

$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i|$$

Minkowsky Distance:

$$d(x,y) = (\sum_{i=1}^{m} |x_i - y_i|^r)^{1/r}$$

Chebychev Distance:

Rumus *Chebychev Distance bertujuan untuk mencari jarak yang terbesar antara x_i dan y_i*

$$d(x,y) = \max_{i=1}^{n} |x_i - y_i|$$

E. Training

Pada proses training akan digunakan membuat model data yang akan digunakan untuk proses selanjutnya yaitu proses testing. Training dilakukan menggunakan algoritma Logistic Regression , ANN, Gaussian Naive Bayes dan KNN.

F. Testing

Testing adalah tahap perbandingan kinerja model yang telah divalidasi dengan data prediksi, lalu mengaplikasikan data yang sudah ditraining dengan menciptakan prediksi data baru. Testing dilakukan dengan algoritma logistic regression, ANN, Gaussian Naive Bayes, KNN.

G. Evaluasi model

Tahapan evaluasi model adalah tahap lanjutan dari testing. Pada penelitian ini menggunakan metric performa *accuracy* dan *f1 score*. berikut adalah rumus yang digunakan pada tahap evaluasi model,

Accuracy:

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{(True Positive + False Positive + True Negative + False Negative)}}$$

f1 score:

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1}$$

III. HASIL DAN PEMBAHASAN

A. Hasil Uji Coba Algoritma Logistic Regression
Tanpa PCA:
Akurasi = 84.17%
F1 score = 69.84%

Dengan PCA:
Akurasi = 84.36%

Perolehan hasil diatas dengan menggunakan default parameter yaitu penalty dengan nilai l2.

B. Hasil Uji Coba Algoritma Artificial Neural Network
Tanpa PCA:
Akurasi = 80.83%
F1 score = 65.67%

Dengan PCA:
Akurasi = 88.27%

Pada penelitian kami pada tahap uji coba algoritma ANN menghasilkan akurasi dengan PCA dan akurasi tanpa PCA. Hasil dari penelitian ini menghasilkan bahwa akurasi yang menggunakan PCA memperoleh nilai lebih besar dari pada nilai akurasi yang tidak menggunakan PCA.

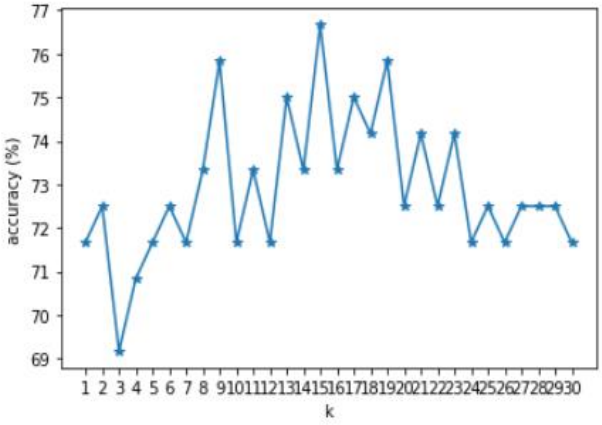
C. Hasil Uji Coba Algoritma Gaussian Naïve Bayes
Tanpa PCA:
Akurasi = 78.33%
F1 score = 53.57%

Dengan PCA:
Akurasi = 84.92%

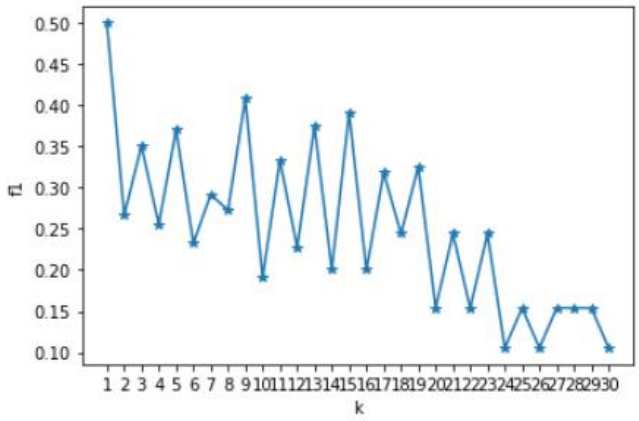
Pada tahap uji coba algoritma gaussian naive bayes memperoleh nilai akurasi 84.92% dengan PCA, 78.33% tidak menggunakan PCA dan menghasilkan nilai F1 score sebesar 53.57%. Dapat disimpulkan bahwa pada tahap uji coba algoritma gaussian naive bayes yang memiliki tingkat akurasi lebih tinggi adalah akurasi yang menggunakan PCA.

D. Hasil Uji Coba Algoritma K-Nearest Neighbors
Pada model ini kami mencoba untuk men-looping model ini dengan menggunakan nilai k(jumlah tetangga terdekat) yang bervariasi dari k=1 hingga k=31.

Pada k=15 didapatkan akurasi tertinggi yaitu sebesar 76.67% dengan nilai F1 sebesar 39.13
Pada k=1 didapatkan akurasi yang lebih rendah yaitu sebesar 71.67% namun dengan nilai F1 terbesar yaitu 49.99%



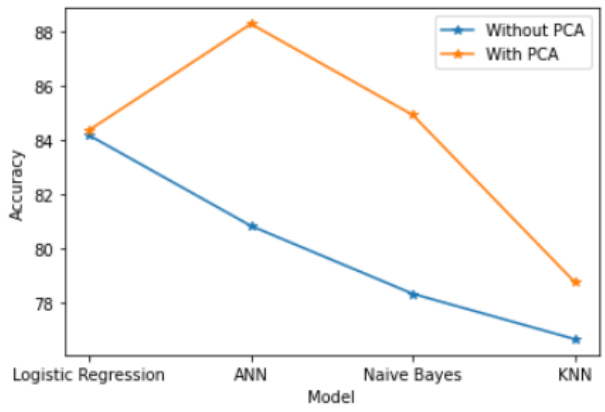
Gambar 5. Grafik dari akurasi metode KNN banding jumlah k



Gambar 6. Grafik Dari nilai F1 metode KNN banding jumlah k

IV. KESIMPULAN

Hasil dari penelitian yang kami coba dengan *heart failure clinical records dataset* menunjukkan bahwa tingkat akurasi tertinggi diperoleh dari Neural Network dengan menggunakan PCA yaitu 88.27%. Selain algoritma Neural Network, algoritma Logistic Regression, Gaussian Naive Bayes dan KNN mendapatkan nilai akurasi yang lebih tinggi dengan menggunakan PCA daripada tanpa menggunakan PCA. Berikut adalah hasil dari penelitian kami,



Gambar 7. Grafik Perbandingan antara model menggunakan PCA dan Tanpa PCA

REFERENSI

[1] Henkel, D. M., Redfield, M. M., Weston, S. A., Gerber, Y., & Roger, V. L. (2008). Death in Heart Failure. *Circulation: Heart Failure*, *1*(2), 91–97. https://doi.org/10.1161/CIRCHEARTFAILURE.107.743146

[2] Primadi,Oscar. 2017. "Penyakit Jantung Penyebab Kematian tertinggi, Kemenkes Ingatkan CERDIK". https://sehatnegeriku.kemkes.go.id/baca/umum/20170801/2521890/penyakit-jantung-penyebab-kematian-tertinggi-kemenkes-ingatkan-cerdik-2/.

[3] Willy, Tjin. 2019. "Pengobatan Gagal Jantung". https://www.alodokter.com/gagal-jantung/pengobatan.

[4] PT. Jawa Pos Grup Multimedia Redaksi, "Sepertiga Kematian di Dunia Dipicu Penyakit Jantung, Angkanya Segini," 2018. [Online]. Available: https://www.jawapos.com/kesehatan/29/09/2017/sepertiga-kematian-di-dunia-dipicu-penyakit-jantung-angkanya-segini. [Accessed: 23-Nov-2021]

[5] Hidayat,Anwar.2016."Pengertian dan Penjelasan Metodologi Penelitian".https://www.statistikian.com/2016/11/metodologi-penelitian.html

[6] https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

[7] Oliver, Andre.2021."Bikin data lebih mudah dibaca, yuk, kenalan dengan data preprocessing"."https://glints.com/id/lowongan/data-preprocessing-adalah/#.YZ8pjdBBxPY