JSI : Jurnal Sistem Informasi *(E-Journal), VOL.14, NO.1, April 2022*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

PENERAPAN DATA MINING DAN TEKNOLOGI MACHINE LEARNING PADA KLASIFIKASI PENYAKIT JANTUNG

Iman Saladin B. Azhar[1], Winda Kurnia Sari[2]
[1,2] Fakultas Ilmu Komputer, Universitas Sriwijaya
e-mail: imansaladin@unsri.ac.id, windakurniasari@unsri.ac.id

Abstrak

Saat ini, dalam dunia kesehatan, data analisis dapat diproses untuk mendeteksi dan mendiagnosa penyakit. Dengan perkembangan teknologi, peranan data mining, dan kebutuhan studi digunakan untuk memecahkan masalah tersebut. Maka dari itu, kami memutuskan untuk mengklasifikasikan penyakit jantung menggunakan 3 teknik machine learning: Logistic Regression, K-Nearest Neighbors, Random Forest, dan Tuned K-Nearest Neighbors dengan bahasa pemrograman python. Dataset yang digunakan dalam penelitian ini mempunyai 13 fitur, 1 variabel label, dan 303 contoh di mana 138 menderita karena penyakit cardiovascular dan 165 sehat. Pengukuran yang digunakan untuk membandingkan kinerja teknik data mining yaitu akurasi, presisi, recall, dan f-measure. Hasilnya menunjukkan bahwa Logistic Regression merupakan teknik dengan kinerja terbaik dan mendapatkan akurasi tertinggi 88,52%.

Kata kunci: *penyakit jantung, data mining, machine learning, klasifikasi*

Abstract

In today's healthcare, data analysis can be processed to identify and diagnose various kind of diseases. Through technological development, the role of data mining and study purposes are used to solve the problem. In this case, we decided to classify heart disease using 3 kind of machine learning techniques: Logistic Regression, K-Nearest Neighbors, Random Forest, and Tuned K-Nearest Neighbors with the python programming language. The dataset in this research contained 13 features, 1 label, and 303 instances where 139 of them had cardiovascular and the other 164 were healthy. Measurement by Accuracy, Precision, Recall, and F-measure applied in order to compare all performance of each techniques. The results shown that Logistic Regression performed as the best technique with highest accuracy by 88,52%.

Keywords*: heart disease, data mining, machine learning, classification*

1. PENDAHULUAN

Salah satu penyakit mematikan di dunia yang menyerang banyak orang dikalangan paruh baya adalah penyakit jantung, dan dalam banyak kasus dapat menyebabkan komplikasi yang fatal [6]. Kebanyakan orang bingung dengan istilah penyakit jantung dan penyakit cardiovascular, di mana keadaan yang memacu situasi yang dapat menyebabkan serangan jantung, nyeri dada, atau bahkan stroke. Berdasarkan data WHO, diperkirakan 17 juta orang meninggal setiap tahun karena cardiovascular, khususya serangan jantung dan stroke [9]. Oleh karena itu, penting untuk menjaga kebiasaan sehat dalam kontribusi pencegahan cardiovascular. Banyak tes yang dapat dilakukan untuk mendiagnosa awal dari cardiovascular, yaitu ECG, tekanan darah, kolesterol, dan gula darah.

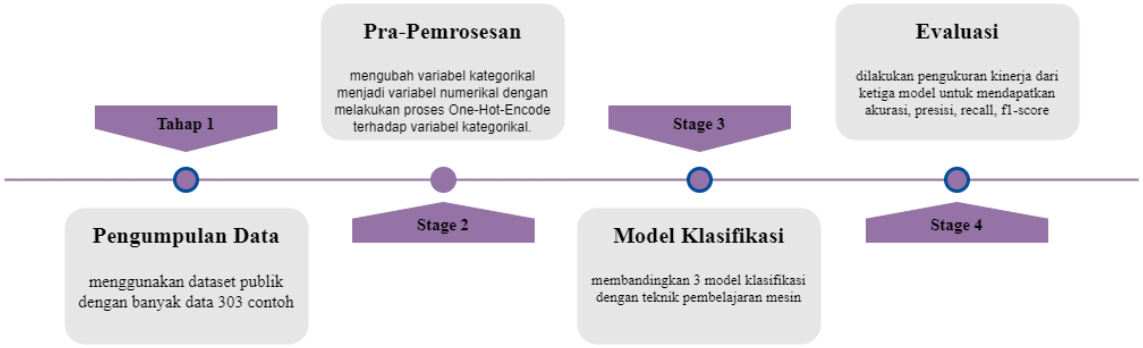
Selain faktor-faktor diatas, kebiasaan gaya hidup seperti kebiasaan makan, kurangnya olahraga dan aktivitas fisik, serta obesitas juga dianggap sebagai faktor risiko utama dari penyakit jantung [1],[5]. Beberapa jenis penyakit jantung yang paling umum

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.14, NO.1, April 2022*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

yang sering terjadi seperti jantung koroner, kardiomiopati, jantung bawaan, penyakit katup jantung, endokarditis, aritmia, dan tumor jantung. Sulit untuk menentukan secara manual kemungkinan terkena penyakit jantung berdasarkan faktor risiko nya [8]. *Machine learning* atau disebut pembelajaran mesin merupakan salah satu teknik data mining yang dapat memprediksi keluaran dari data yang ada. Penelitian ini menggunakan 3 dasar teknik pembelajaran mesin untuk mengklasifikasikan apakah mempunyai penyakit atau tidak, berdasarkan faktor-faktor risiko. Penelitian ini juga melakukan *tuning* pada 3 teknik pembelajaran mesin untuk dapat membandingkan hasil prediksi yang lebih akurat.

2. METODOLOGI PENELITIAN

Pada penelitian ini dilakukan klasifikasi penyakit jantung dengan menggunakan teknik pembelajaran mesin untuk menentukan mana yang mempunyai penyakit dan tidak. Adapun tahapan-tahapan metodologi penelitian yang dilakukan sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Metodologi Penelitian

2.1. Dataset

Kumpulan data atau disebut *dataset* yang digunakan pada penelitian ini didapat dari Kaggle *dataset*, sebuah dataset publik yang terdiri dari 13 atribut, 2 label, dan 303 sampel data dengan 165 yang mengidap penyakit dan 138 yang tidak sakit. Atribut-atribut dari dataset ini sebagai berikut.

1. ***age*** - usia dalam tahun
2. ***sex*** - (1 = laki-laki; 0 = perempuan)
3. ***cp*** - jenis nyeri dada
   - 0: Angina tipikal: nyeri dada berhubungan dengan penurunan suplai darah ke jantung
   - 1: Angina atipikal: nyeri dada yang tidak berhubungan dengan jantung
   - 2: Nyeri non-angina: biasanya kejang esofagus (tidak berhubungan dengan jantung)
   - 3: Tanpa gejala: nyeri dada tidak menunjukkan tanda-tanda penyakit
4. ***trestbps*** - tekanan darah istirahat (dalam mm Hg saat masuk ke rumah sakit) apa pun di atas 130-140 biasanya menjadi perhatian

5. ***chol*** - kolesterol serum dalam mg/dl
   - serum = LDL + HDL + .2 * trigliserida
   - di atas 200 adalah penyebab kekhawatiran
6. ***fbs*** - (gula darah puasa > 120 mg/dl) (1 = benar; 0 = salah)
   - '>126' mg/dL menandakan diabetes
7. ***restecg*** - hasil elektrokardiografi istirahat
   - 0: Tidak ada yang perlu diperhatikan
   - 1: Kelainan Gelombang ST-T
     - dapat berkisar dari gejala ringan hingga masalah berat
     - menandakan detak jantung tidak normal
   - 2: Kemungkinan atau pasti hipertrofi ventrikel kiri
     - Ruang pemompaan utama jantung membesar
8. ***thalach*** - detak jantung maksimum tercapai
9. ***exang*** - angina akibat olahraga (1 = ya; 0 = tidak)
10. ***oldpeak*** - ST depresi yang disebabkan oleh olahraga relatif terhadap istirahat terlihat stres jantung selama latihan jantung yang tidak sehat akan lebih stres
11. ***slope*** - kemiringan segmen ST latihan puncak
    - 0: Upsloping: detak jantung lebih baik dengan olahraga (jarang)
    - 1: Flatsloping: perubahan minimal (khas jantung sehat)
    - 2: Downslopins: tanda-tanda jantung yang tidak sehat
12. ***ca*** - jumlah pembuluh darah besar (0-3) yang diwarnai dengan fluoroskopi
    - pembuluh berwarna berarti dokter dapat melihat darah yang lewat
    - semakin banyak pergerakan darah semakin baik (tidak ada gumpalan)
13. ***thal*** - thalium hasil stres
    - 1,3: biasa
    - 6: cacat tetap: dulu cacat tapi ok sekarang
    - 7: cacat reversibel: tidak ada pergerakan darah yang tepat saat berolahraga

2.2. Pra-Premosesan

Pada prapemrosesan data dilakukan *split* data dengan pembagian data pelatihan 80% dan data pengujian 20%. Dalam penelitian ini, menggunakan fungsi *dummies* yang digunakan untuk mengubah data yang berupa kategorikal menjadi numerikal dengan melakukan proses *One-Hot-Encode* terhadap data kategorikal. *One-Hot-Encode* adalah proses untuk membuat kolom baru dari data kategorikal di mana setiap kategori menjadi kolom baru dengan nilai 0 atau 1 yang artinya 0 mewakili tidak ada dan 1 mewakili ada.

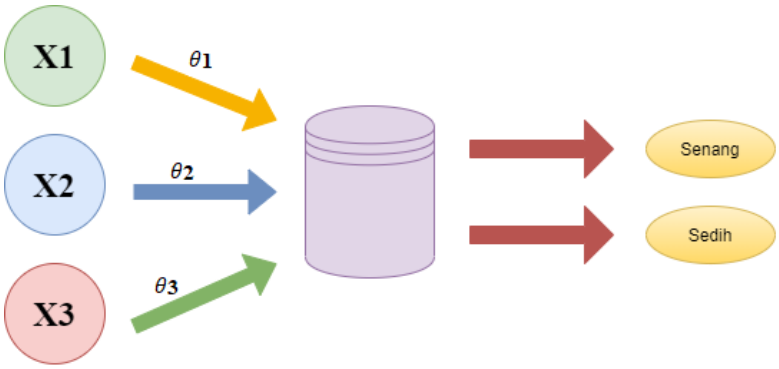
2.3. Model Klasifikasi

Terdapat 3 model yang dibangun dari pengklasifikasian dengan teknik pembelajaran mesin, di mana dari ketiga model di *tuning* dengan menambah atau mengubah parameter. Proses *tuning* dilakukan agar model menjadi lebih akurat untuk membantu diagnosa penyakit.

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.14, NO.1, April 2022*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

1) Logistic Regression

Logistic Regression merupakan algoritma analisis prediksi regresi yang efisien. Penerapannya efisien ketika variabel tidak bebas dari dataset adalah biner. *Logistic Regression* digunakan dalam mendeskripsikan dan menganalisis data untuk menjelaskan hubungan antara satu variabel biner tidak bebas dan satu atau lebih variabel bebas. [10]. Model *Logistic regression* terdapat pada Gambar 2 dibawah ini.



Gambar 2. Model Logistic Regression

Di mana X1, X2, dan X3 adalah masukan. $\theta 1$, $\theta 2$, dan $\theta 3$ adalah bobot. Sedangkan "Senang" dan "Sedih" adalah keluaran.

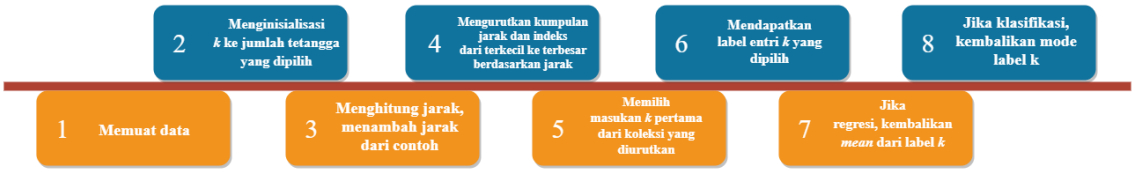
2) K-Nearest Neighbors

Pada pengklasifikasian k-NN, keanggotaan kelas yang menjadi keluaran. Sebuah objek diklasifikasikan oleh vote terbanyak dari tetangganya, dengan objek yang ditugaskan ke kelas yang paling umum di antara *k* tetangga terdekatnya. *k* adalah bilangan bulat positif. Jika k = 1, maka objek hanya ditugaskan ke kelas tetangga tunggal terdekat. Pilihan terbaik dari *k* tergantung pada data, umumnya, nilai *k* yang lebih besar akan mengurangi efek *noise* pada klasifikasi [2] tetapi membuat batas antar kelas menjadi kurang jelas. *k* yang baik dapat dipilih dengan berbagai teknik heuristik, seperti optimasi *hyperparameter*. Untuk lebih jelas, berikut Gambar 3 menunjukkan bagaimana tahapan dari algoritma KNN.

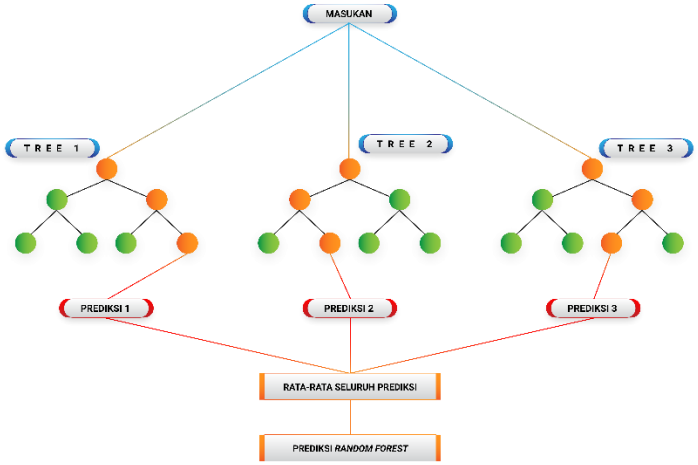
3) Random Forest

Random forest merupakan algoritma klasifikasi yang menggunakan ensemble pohon klasifikasi, yang masing-masing ditetapkan dengan menerapkan sampel bootstrap dari data [5]. Untuk pembuatan pohon, variabel dipilih secara acak sebagai himpunan kandidat variabel pada setiap pemisahan. Random forest memiliki kinerja yang mumpuni dalam tugas-tugas klasifikasi seperti *robustness* dalam hal set fitur yang besar, penggabungan interaksi antara variabel prediktor, dan kualitas tinggi dan implementasi bebas [3]. Diagram struktur Random Forest ditunjukkan pada Gambar 4.

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.14, NO.1, April 2022*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id



Gambar 3. Tahapan Algoritma KNN



Gambar 4. Model Random Forest

2.4. Evaluasi

Dalam data mining, evaluasi untuk klasifikasi berupa *confusion matrix* untuk tahu seberapa baik model yang dibuat dengan diberikannya informasi dari perbandingan hasil pengklasifikasian yang dikerjakan oleh model dengan hasil kelas yang sebenarnya. Hasil dari confusion matrix yaitu mendapatkan nilai akurasi, presisi, recall, dan f-measure. Tabel 1 menampilkan rumus perhitungan *confusion matrix*.

Tabel 1. Confusion Matrix

	Prediksi:	
Aktual:	A	B
A	YES	NO
B	NO	YES

Dimana, jika nilai YES(A) akan menghasilkan prediksi yang positif dan itu memang hasil yang benar, sebaliknya jika YES(B) maka akan menghasilkan prediksi yang negatif dan juga benar. Jika terdapat multilabel atau lebih dari dua label maka dapat menggunakan *confusion matrix* seperti penelitian ini [10].

3. HASIL DAN ANALISIS

Penelitian ini menggunakan teknik data mining untuk mengklasifikasikan data yang mempunyai penyakit jantung dan tidak, dengan bantuan teknik pembelajaran mesin.

Teknik yang digunakan berupa Logistic Regression, K-Nearest Neighbors, dan Random Forest. Di mana ketiga metode pembelajaran mesin dilatih dan diuji sehingga mendapatkan model klasifikasi terbaik. Kemudian dilakukan *tuning* parameter terhadap ketiga metode tersebut sehingga dapat dibandingkan hasilnya pada Tabel 2. Proses yang dilakukan antara lain, memasukkan dataset, kemudian data di proses dengan mengubah data kategorikal menjadi data numerik, setelah itu dibangun model, terakhir melatih dan menguji dengan metode klasifikasi. Gambar 5 menunjukkan hasil confusion matrix dan klasifikasi dari pengujian Logistic Regression. Gambar 6 menunjukkan hasil confusion matrix dan klasifikasi dari pengujian K-Nearest Neighbors. Gambar 7 menunjukkan hasil *confusion matrix* dan klasifikasi dari pengujian Random Forest.

```
Train Result:
================================================
Accuracy Score: 85.95%
________________________________________________
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.87   0.85      0.86       0.86          0.86
recall       0.81   0.90      0.86       0.86          0.86
f1-score     0.84   0.87      0.86       0.86          0.86
support    111.00 131.00      0.86     242.00        242.00
________________________________________________
Confusion Matrix:
 [[ 90  21]
 [ 13 118]]

Test Result:
================================================
Accuracy Score: 88.52%
________________________________________________
CLASSIFICATION REPORT:
                0     1  accuracy  macro avg  weighted avg
precision    0.88  0.89      0.89       0.89          0.89
recall       0.85  0.91      0.89       0.88          0.89
f1-score     0.87  0.90      0.89       0.88          0.88
support     27.00 34.00      0.89      61.00         61.00
________________________________________________
Confusion Matrix:
 [[23  4]
 [ 3 31]]
```

Gambar 5. Confusion Matrix dan Klasifikasi Logistic Regression

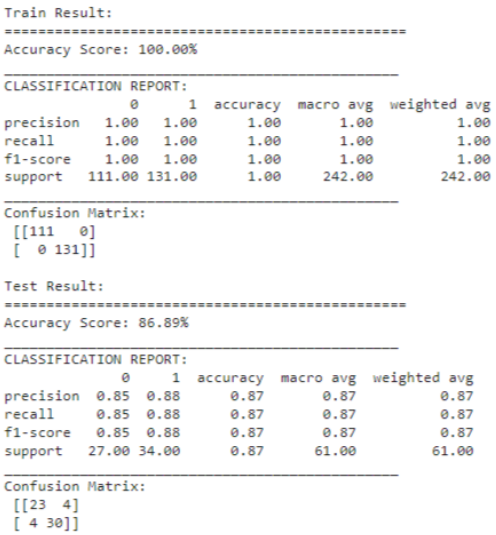
```
Train Result:
================================================
Accuracy Score: 89.26%
________________________________________________
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.91   0.88      0.89       0.90          0.89
recall       0.85   0.93      0.89       0.89          0.89
f1-score     0.88   0.90      0.89       0.89          0.89
support    111.00 131.00      0.89     242.00        242.00
________________________________________________
Confusion Matrix:
 [[ 94  17]
 [  9 122]]

Test Result:
================================================
Accuracy Score: 78.69%
________________________________________________
CLASSIFICATION REPORT:
                0     1  accuracy  macro avg  weighted avg
precision    0.77  0.80      0.79       0.78          0.79
recall       0.74  0.82      0.79       0.78          0.79
f1-score     0.75  0.81      0.79       0.78          0.79
support     27.00 34.00      0.79      61.00         61.00
________________________________________________
Confusion Matrix:
 [[20  7]
 [ 6 28]]
```

Gambar 6. Confusion Matrix dan Klasifikasi k-NN

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.14, NO.1, April 2022*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

Hasil pengujian klasifikasi dengan Logistic Regression mendapatkan nilai akurasi 88,52% lebih besar dibandingkan hasil pelatihan data nya, namun tetap nilai presisi, recall, dan f-measure stabil. Sedangkan pada pelatihan klasifikasi menggunakan K-Nearest Neighbors mendapatkan akurasi 89,26% dan menurun pada saat pengujian dilakukan menjadi 78,69%. Hasil yang mengejutkan terjadi pada klasifikasi menggunakan Random Forest di mana pelatihan mendapatkan 100% akurat, namun berkurang ketika dilakukan pengujian sebesar 86,89%.

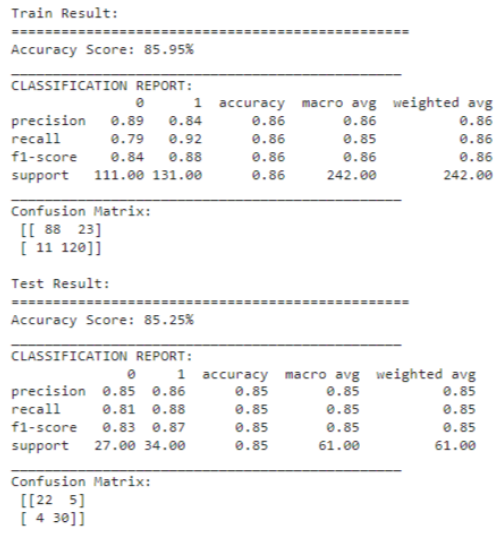


```
Train Result:
================================================
Accuracy Score: 100.00%
________________________________________________
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    1.00   1.00      1.00       1.00          1.00
recall       1.00   1.00      1.00       1.00          1.00
f1-score     1.00   1.00      1.00       1.00          1.00
support    111.00 131.00      1.00     242.00        242.00
________________________________________________
Confusion Matrix:
 [[111   0]
 [  0 131]]

Test Result:
================================================
Accuracy Score: 86.89%
________________________________________________
CLASSIFICATION REPORT:
                0     1  accuracy  macro avg  weighted avg
precision    0.85  0.88      0.87       0.87          0.87
recall       0.85  0.88      0.87       0.87          0.87
f1-score     0.85  0.88      0.87       0.87          0.87
support     27.00 34.00      0.87      61.00         61.00
________________________________________________
Confusion Matrix:
 [[23  4]
 [ 4 30]]
```

Gambar 7. Confusion Matrix dan Klasifikasi Random Forest

Dengan ketiga hasil dari teknik klasifikasi pembelajaran mesin diatas, penelitian ini mencoba menaikkan akurasi pada K-Nearest Neighbors dengan menambahkan parameter pengujian menggunakan *range* pada *k* nya. Sehingga didapatkan nilai *confusion matrix* dan klasifikasi yang meningkat dibandingkan sebelumnya. Hasil dapat ditunjukkan pada Gambar 8.



```
Train Result:
================================================
Accuracy Score: 85.95%
________________________________________________
CLASSIFICATION REPORT:
                0      1  accuracy  macro avg  weighted avg
precision    0.89   0.84      0.86       0.86          0.86
recall       0.79   0.92      0.86       0.85          0.86
f1-score     0.84   0.88      0.86       0.86          0.86
support    111.00 131.00      0.86     242.00        242.00
________________________________________________
Confusion Matrix:
 [[ 88  23]
 [ 11 120]]

Test Result:
================================================
Accuracy Score: 85.25%
________________________________________________
CLASSIFICATION REPORT:
                0     1  accuracy  macro avg  weighted avg
precision    0.85  0.86      0.85       0.85          0.85
recall       0.81  0.88      0.85       0.85          0.85
f1-score     0.83  0.87      0.85       0.85          0.85
support     27.00 34.00      0.85      61.00         61.00
________________________________________________
Confusion Matrix:
 [[22  5]
 [ 4 30]]
```

Gambar 8. Confusion Matrix dan Klasifikasi *Tuning* K-Nearest Neighbors

Tabel 2. Perbandingan Akurasi

Model	Akurasi %
Logistic Regression	88.52
K-Nearest Neighbors	78.69
Random Forest Classifier	86.89
Tuned K-Nearest Neighbors	85.25

4. KESIMPULAN

Dari pengujian yang dilakukan untuk mengklasifikasikan penyakit jantung menggunakan teknik data mining dan pembelajaran mesin, didapatkan kesimpulan bahwa prediksi yang lebih akurat dilakukan oleh Logistic Regression. Pengujian dilakukan dengan jumlah data yang sama, terdapat 303 sampel data dengan 80% data pelatihan dan 20% pengujian. Dari pengukuran presisi, recall, dan f-measure setiap metode mendapatkan nilai yang stabil sehingga tidak terlalu menurunkan nilai pengujian akurasi.

Untuk penelitian kedepan, diharapkan mendapatkan model yang lebih banyak dan hasil yang lebih maksimal sehingga prediksi kelasnya semakin akurat dengan metode pembelajaran mendalam (*deep learning*) dan dengan data yang besar.

REFERENCES

[1] Choi S, Kim K, Kim SM, Lee G, Jeong SM, Park SY, Kim YY, Son JS, Yun JM, Park SM. Association of obesity or weight change with coronary heart disease among young adults in South Korea. JAMA internal medicine. 2018 Aug 1;178(8):1060-8. [4]

[2] Everitt BS, Landau S, Leese M, Stahl D. Miscellaneous clustering methods. Cluster analysis. 2011:215-55. [7]

[3] Janitza S, Tutz G, Boulesteix AL. Random forest for ordinal responses: prediction and variable selection. Computational Statistics & Data Analysis. 2016 Apr 1;96:57-73. [9]

[4] Kurnia Sari W, Palupi Rini D, Firsandaya Malik R. Text Classification Using Long Short-Term Memory With GloVe Features. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika. 2020 Feb 4;5(2):85-100. [10]

[5] Lassale C, Tzoulaki I, Moons KG, Sweeting M, Boer J, Johnson L, Huerta JM, Agnoli C, Freisling H, Weiderpass E, Wennberg P. Separate and combined associations of obesity and metabolic health with coronary heart disease: a pan-European case-cohort analysis. European heart journal. 2018 Feb 1;39(5):397-406. [3]

[6] Latha CB, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. 2019 Jan 1;16:100203. [1]

[7] Sun Y, Li G, Zhang J, Qian D. Prediction of the strength of rubberized concrete by an evolved random forest model. Advances in Civil Engineering. 2019 Dec 28;2019. [8]

[8] Tan JH, Hagiwara Y, Pang W, Lim I, Oh SL, Adam M, San Tan R, Chen M, Acharya UR. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. Computers in biology and medicine. 2018 Mar 1;94:19-26. [5]

[9] WHO. Cardiovascular diseases (CVDs). 17 May 2017. https://www.who.int/news-room/fact-sheets/detail/cardiovasculardiseases-(cvds). Accessed 20 Dec 2019. [2]

[10] Zhu C, Idemudia CU, Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked. 2019 Jan 1;17:100179. [6]