KLASIFIKASI PENYAKIT HATI MENGGUNAKAN RANDOM FOREST DAN KNN

Ketut Widya Kayohana
Ilmu Komputer, Universitas Bumigora
Jl. Ismail Marzuki No.22, Cilinaya, Kec. Cakranegara, Kota Mataram, Nusa Tenggara Bar. 83127
ketut.widya@universitasbumigora.ac.id

ABSTRAK

Penderita liver (hati) meningkat dari tahun ke tahun. Hal ini dikarenakan oleh gaya hidup yang tidak sehat misalnya pengkonsumsian alkohol, serta karena keterlambatan diagnosa penyakit. Penyakit hati merupakan salah satu penyakit yang menjadi masalah nasional di semua negara, baik di negara-negara berkembang seperti Indonesia maupun di negara - negara maju. Data yang digunakan pada penelitian ini adalah data sekunder yang diambil melalui internet. 15 variabel diantaranya gender, age, education, currentSmoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartrate, glucose dan 1 'class' dengan label Heart_stroke. metode logistic regression, K-Nearest Neighbors (KNN), Random Forest, dan Stochastic Gradient Descent (SGD) memiliki nilai akurasi sebesar 68%, 82%, 90% dan 48%. Dari perbandingan hasil tingkat akurasi metode yang digunakan, metode dengan tingkat akurasi terbesar didapatkan oleh Random Forest dan tingkat akurasi terkecil adalah Stochastic Gradient Desce

Kata Kunci : *Random Forest, Klasifikasi Penyakit Hati, KNN*

1. PENDAHULUAN

Penderita liver (hati) meningkat dari tahun ke tahun. Hal ini dikarenakan oleh gaya hidup yang tidak sehat misalnya pengkonsumsian alkohol, serta karena keterlambatan diagnosa penyakit. Penyakit hati merupakan salah satu penyakit yang menjadi masalah nasional di semua negara, baik di negara-negara berkembang seperti Indonesia maupun di negara - negara maju. Penyakit ini dapat terjadi pada semua golongan umur, mulai dari anak-anak, remaja, dewasa sampai orang tua [1].

Beberapa penyakit yang menyerang hati salah satunya adalah Hepatitis. Menurut data WHO virus Hepatitis B menyerang 350 juta orang di dunia terutama Asia Tenggara dan Afrika yang menyebabkan kematian sebanyak 1,2 juta pertahun [2]. Ada beberapa faktor yang dapat mengakibatkan penyakit hati ini yang terus meningkat. Diantaranya infeksi atau peradangan pada jaringan hati, obesitas, dan kelainan sistem kekebalan tubuh. Untuk mendiagnosa penyakit hati perlu dilakukan tes darah. Berdasarkan hasil tes darah tersebut dapat diketahui seorang pasien menderita penyakit hati atau tidak[3]. Adapun jenis-jenis penyakit hati yang umum antara lain yaitu sirosis, kanker hati atau hepatoma, abses hati, kolesistitis dan perlemakan hati non alkoholik[4].

Salah satu cara untuk mendeteksi penyakit hati adalah dengan melakukan klasifikasi. Oleh karena itu, penelitian ini dilakukan untuk membantu dalam menyelesaikan masalah tertentu dengan machine learning yang dapat berfungsi untuk klasifikasi penyakit hati. Dimana pada penelitian ini diperlukan suatu metode yang dapat mengolah data. Salah satu metodenya yaitu menggunakan teknik dari machine learning.

Rendon forest didasarkan pada teknologi pohon keputusan, yang memungkinkan hutan tersebut mengatasi masalah nonlinier. Metode ini menggabungkan pohon. Untuk mengidentifikasi variabel penjelas yang berkaitan dengan suatu variabel respon, random forest membuat ukuran pentingnya variabel penjelas (variable important). Dalam bidang biostatistika, hal ini berlaku pada masalah seleksi gen pada data microarray [5].

Metode K-Nearest Neighbor (KNN) merupakan metode untuk mengklasifikasikan objek berdasarkan nilai (K) objek terdekat. Selain itu juga menggunakan ekstraksi ciri HSV (hue, saturation, dan lightness) untuk mengekstrak warna guna membantu proses pengklasifikasian tingkat kematangan buah belimbing. fungsionalitas warna digunakan untuk menyederhanakan pemrosesan gambar[6].

Penggunaan metode K-NN adalah metode melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Dimana terdapat 3 variabel yang akan diolah yaitu variable age (umur), education (pendidikan), glucose (gula), (Tenyearchd).

2. KAJIAN PUSTAKA

Kesehatan merupakan salah satu hal terpenting dalam hidup manusia, hal ini menjadi dasar bahwasanya banyak ditemukannya temuan- temuan ilmiah baik berupa temua obat-obatan, alat kesehatan atau penemuan-penemuan baru di bidang kesehatan. Sejak dahulu banyak sekali penyakitpenyakit bermunculan entah itu datangnya dari virus, bakteri, parasite, sel kanker dan lain-lain. Salah satunya adalah organ tubuh hati[7].

Sistem Pakar adalah suatu program komputer yang mensimulasikan penilaian dan perilaku manusia atau organisasi yang memiliki pengetahuan dan pengalaman ahli dalam bidang tertentu. Biasanya sistem itu mengandung basis pengetahuan, akumulasi pengalaman dan perangkat aturan untuk menerapkan kodisi setiap suatu situasi tertentu yang dijelaskan

dalam suatu program[8]. Untuk memudahkan pendeteksian penyakit liver maka diperlukan alat atau mesin yang dapat mendeteksi penyakit liver secara lebih efisien dengan mengklasifikasikan pasien yang menderita penyakit liver dan yang tidak menderita penyakit liver. Salah satu metode yang mungkin adalah penggunaan pembelajaran mesin[9]. Pembelajaran mesin (machine learning) merupakan salah satu bidang ilmu yang merupakan bagian dari kecerdasan buatan (artificial intelligence), dan merupakan suatu sistem yang dapat belajar dan mengambil keputusan sendiri tanpa adanya pemrograman berulang-ulang oleh manusia, dan merupakan suatu sistem yang belajar dari data yang dimiliki oleh komputer. Hal ini memungkinkan komputer menjadi cerdas[10].

KNN adalah pendekatan yang mudah diterapkan dan merupakan metode lama yang biasa digunakan dalam klasifikasi. knn memiliki tingkat efisiensi yang tinggi dan dapat memberikan tingkat akurasi klasifikasi yang tinggi dalam beberapa kasus. Metode K-NN memiliki keunggulan unik dalam memprediksi data yang diperoleh dari prediksi penyakit liver di Indonesia[11].

3. METODE PENELITIAN

Data yang digunakan pada penelitian ini adalah data sekunder yang diambil melalui internet. Untuk menyelesaikan penelitian ini, diagram aliran penelitian yang digunakan ditunjukkan pada Gambar 1. Pada tahapan pertama melakukan pengumpulan data penyakit hati, tahapan kedua melakukan preprocessing, tahapan ketiga yaitu data imbalance, tahap keempat melakukan modeling, dan tahap terakhir adalah evaluasi.

3.1. Preprocessing

Pada proses preprocessing merupakan tahap awal untuk melakukan suatu pengujian model algoritma. Dataset yang digunakan akan diolah menjadi data bersih yang siap untuk diuji. Preprocessing ini bertujuan untuk menghilangkan noise dan menyamakan bentuk data agar sesuai dan dapat diuji dengan model algoritma yang akan digunakan. Proses preprocessing ini mencakup dua tahapan yaitu cleaning data dan transform data. Proses cleaning data dilakukan dengan menghilangkan noise, membuang duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan pada data. transform data merupakan langkah terakhir preprocessing pada penelitian ini yang bertujuan untuk perubahan struktur fitur data yang digunakan agar menyesuaikan dengan kebutuhan model algoritma[12].

3.2. Data Imbalance

Data imbalance atau data tidak seimbang merupakan suatu keadaan data dimana datanya mengalami ketidak seimbangan dalam klasifikasi pembelajaran mesin yang terdapat rasio tidak proporsional disetiap kelas. Biasanya kelompok kelas data yang memiliki jumlah data lebih sedikit dikenal dengan kelompok minoritas (minority), sedangkan kelompok kelas data yang jumlah data yang banyak dikenal dengan istilah mayoritas (mayority).

Pada penelitian ini jumlah data yang mengidap penyakit liver atau hati memiliki jumlah data yang berbeda dengan jumlah data yang tidak mengalami penyakit hati. Oleh karena itu, jumlah data tersebut akan dilakukan class imbalance dengan menggunakan SMOTE.

3.3. Modeling

Modeling merupakan proses untuk membangun model prediktif atau deskriptif berdasarkan data yang dimiliki. Pada modeling ini dapat menerapkan teknik dan algoritma yang tepat untuk mencari pola tersembunyi pada dataset[13]. Tujuan dari proses modeling ini yaitu menghasilkan model yang dapat digunakan untuk memprediksi hasil masa depan, mengidentifikasi tren, atau memberikan wawasan baru tentang data yang ada. Model-model ini dapat digunakan untuk mengambil keputusan yang lebih baik, memahami perilaku konsumen, mendeteksi anomali, dan mengoptimalkan proses bisnis. Berikut model yang digunakan pada penelitian ini:

3.4. Regresi Logistik

Regresi logistik biner merupakan salah satu teknik analisis statistika dengan satu atau lebih variabel bebas dan satu variabel respon. Variabel bebas dapat berupa data kategorik maupun kontinu, sedangkan untuk variabel respon harus berskala kategorik [1]. Kemudian menggunakan hubungan ini untuk memprediksi nilai dari salah satu faktor tersebut berdasarkan faktor yang lain. Prediksi biasanya memiliki jumlah hasil yang terbatas, seperti ya atau tidak serta memiliki nilai 1 dan 0.

3.5. KNN (K-Nearest Neighbor)

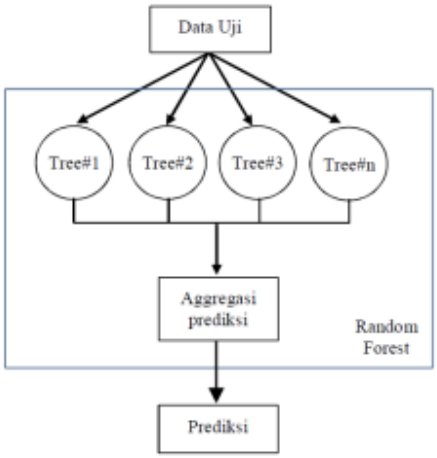
Algoritma K-Nearest Neighbor merupakan metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumnya[13]. Prinsip kerja algoritma dari KNN yaitu menentukan dan mencari jarak terdekat dengan nilai k neighbor terdekat dalam data training dengan data yang akan diuji. Nilai k yang terbaik untuk algoritma ini tergantung berdasarkan nilai sebuah data, dimana biasanya nilai k yang tinggi mengurangi efek kesalahan atau noise pada proses klasifikasi, akan tetapi membuat sebuah batasan antar klasifikasi menjadi tidak maksimal [4]. Hitung jarak dari data untuk dibandingkan dengan dataset training, dapat dihitung dengan persamaan jarak Euclidean

$$dist\ (p,q) = \sqrt{\sum_{i-1} n(pi - qi)^2}$$

Ket:
Dist (p,q) = jarak antara p dan 1
Pi = nilai ke-i pada data p
Qi = nilai ke-i pada data q

3.6. Random Forest

Random forest adalah algoritma klasifikasi dan regresi yang menjadi bagian dari kelompok ensemble learning. Metode random forest merupakan pengembangan dari decision tree dimana setiap decision tree telah dilakukan proses pelatihan dengan menggunakan sampel individu[14]. Metode ini menerapkan metode bootstrap aggregating (bagging) dan random feature selection[15].



Gambar 2. Proses Prediksi Random Forest

3.7. Stochastic Gradient Descent (SGD)

Merupakan metode optimasi sederhana berbasis statistik yang efisien digunakan mencari nilai koefisien untuk meminimalkan loss (error) function pada skala besar. Proses algoritma SGD adalah dengan menemukan nilai θ yang dapat meminimalkan fungsi J(θ). Untuk menentukan nilai awal θ digunakan algoritma pencarian, kemudian pada setiap iterasi nilai θ agar terus diperbaharui sampai menemukan titik minimum atau nilai J yang paling minimum.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Pada penelitian ini dataset yang digunakan untuk melakukan eksperimen dalam penelitian yaitu menggunakan dataset penderita penyakit jantung. Pada *dataset* ini berjumlah 4238 data yang akan dibagi menjadi 2 bagian yaitu 20% untuk data uji (*testing*) dan 80% digunakan untuk data latih (*training*). Berikut tampilan data di tunjukkan pada gambar 3

Gender	age	education	currentSmo	cigsPerDay	BPMeds	prevalentS	prevalentH	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Heart_	stroke
Male	39	postgradu	0	0	0	no	0	0	195	106	70	26.97	80	77	No	
Female	46	primarysc	0	0	0	no	0	0	250	121	81	28.73	95	76	No	
Male	48	uneducate	1	20	0	no	0	0	245	127.5	80	25.34	75	70	No	
Female	61	graduate	1	30	0	no	1	0	225	150	95	28.58	65	103	yes	
Female	46	graduate	1	23	0	no	0	0	285	130	84	23.1	85	85	No	
Female	43	primarysc	0	0	0	no	1	0	228	180	110	30.3	77	99	No	
Female	63	uneducate	0	0	0	no	0	0	205	138	71	33.11	60	85	yes	
Female	45	primarysc	1	20	0	no	0	0	313	100	71	21.68	79	78	No	
Male	52	uneducate	0	0	0	no	1	0	260	141.5	89	26.36	76	79	No	
Male	43	uneducate	1	30	0	no	1	0	225	162	107	23.61	93	88	No	
Female	50	uneducate	0	0	0	no	0	0	254	133	76	22.91	75	76	No	
Female	43	primarysc	0	0	0	no	0	0	247	131	88	27.64	72	61	No	
Male	46	uneducate	1	15	0	no	1	0	294	142	94	26.31	98	64	No	
Female	41	graduate	0	0	1	no	1	0	332	124	88	31.31	65	84	No	
Female	39	primarysc	1	9	0	no	0	0	226	114	64	22.35	85	NA	No	
Female	38	primarysc	1	20	0	no	1	0	221	140	90	21.35	95	70	yes	
Male	48	graduate	1	10	0	no	0	0	232	138	90	22.37	64	72	No	
Female	46	primarysc	1	20	0	no	0	0	291	112	78	23.38	80	89	yes	
Female	38	primarysc	1	5	0	no	0	0	195	122	84.5	23.24	75	78	No	

Gambar 3. *Dataset Penyakit Jantung*

Data diatas memiliki sebanyak 15 variabel diantaranya gender, age, education, currentSmoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartrate, glucose dan 1 ‘class’ dengan label Heart_stroke.

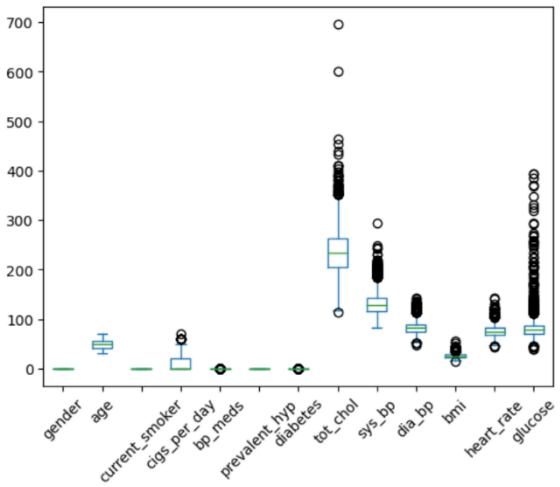
4.2. Pre-Processing

4.2.1. Data Duplikat

Data duplikat merujuk pada salinan atau replika yang identik dari entri data tertentu dalam suatu kumpulan data. Dalam konteks ini, entri data mengacu pada baris atau rekaman individu dalam basis data atau tabel. Ketika ada beberapa entri yang memiliki nilai yang sama untuk setiap kolom yang relevan, entri tersebut dianggap sebagai data duplikat. Pada dataset yang kami gunakan, terdapat 489 data duplikat. Data tersebut tentu akan dihapus untuk menghasilkan dataset yang lebih baik. Setelah data duplikat dibersihkan, didapatkan data yang kami miliki yaitu sebanyak 3749 data.

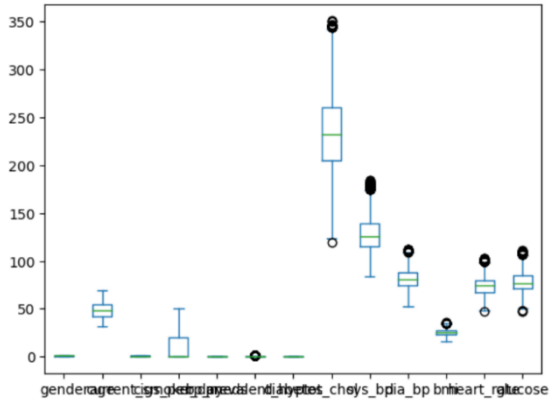
4.2.2. Data Outlier

Data outlier merujuk pada observasi yang signifikan atau tidak biasa yang berbeda secara signifikan dari pola umum atau tren dalam kumpulan data. Outlier dapat muncul sebagai nilai yang jauh di luar rentang normal atau sebagai nilai yang sangat berbeda dari pola data yang ada. Outlier sering kali merupakan hasil dari kesalahan pengukuran, gangguan, atau peristiwa langka yang tidak mewakili keadaan umum. Berikut tampilan data oulier pada dataset kami ditunjukan pada gambar 4.



Gambar 4. *Data Outlier*

Setelah melakukan perbaikan pada Data outlier, didapatkan hasil plot dataset kami sebagai berikut.



Gambar 5. *Data No Outlier*

4.2.3. Class Imbalance

Data imbalance mengacu pada ketidakseimbangan distribusi kelas dalam kumpulan data. Ini terjadi ketika jumlah sampel dalam setiap kelas tidak seimbang atau tidak proporsional. Ketidakseimbangan data dapat menyebabkan masalah dalam pemodelan atau analisis data, karena model cenderung menjadi bias terhadap kelas mayoritas. Berikut tampilan jumlah dataset kami sebelum melalukan class imbalance pada gambar 6

```
df.heart_stroke.value_counts()
```

```
0    2771
1     417
Name: heart_stroke, dtype: int64
```

Gambar 6. *Class imbalance*

Terdapat ketidak seimbangan pada dataset kami, dimana sebanyak 2771 data berlabel normal dan sebanyak 417 data berlabel penyakit hati. Hal ini tentu akan membuat model menjadi bias karena kelas mayoritas. Solusi dalam menyelesaikan masalah imbalance data sangat beragam. Salah satunya dengan cara SMOTE (Syntetic Minority Over-sampling Technique).

SMOTE merupakan teknik untuk mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara membuat instance baru dari minority class dengan pembentukan convex kombanasi dari instances yang saling berdekatan. Dari teknik-teknik Data Balancing yang digunakan serta pada literatur, teknik SMOTE merupakan yang paling banyak dipakai dalam penanganan Imbalance. Pada literatur juga ditemukan bahwa jumlah dataset mempengaruhi performa yang didapatkan dalam klasifikasi. Semakin banyak dataset yang dipakai semakin rendah performa klasifikasi.

Oleh karena itu, teknik untuk penanganan imbalance data yang akan digunakan pada penelitian klasifikasi penyakit hati adalah oversampling SMOTE. Teknik ini digunakan karena dapat membuat dataset menjadi seimbang. Terlihat pada gambar 7. Jumlah data sebelum melakukan SMOTE berbeda dengan jumlah data setelah melakukan SMOTE. Selain itu juga pada gambar tersebut dapat dilihat hasil dari distribusi kelas sesudah melakukan SMOTE sehingga datanya lebih seimbang.

```
smt = SMOTETomek()
re_X, re_y = smt.fit_resample(X, y)
```

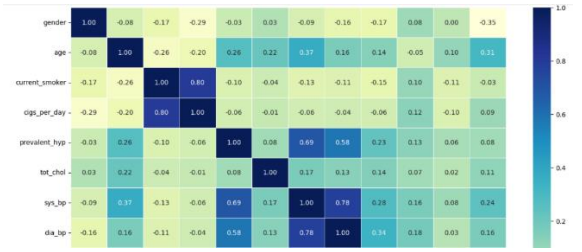
```
re_y.value_counts()
```

```
0    2756
1    2756
Name: heart_stroke, dtype: int64
```

Gambar 7. *Class imbalance*

4.3. Matrix Korelasi

Matriks korelasi adalah alat statistik yang digunakan untuk mengukur hubungan antara dua atau lebih variabel dalam bentuk matriks. Ini menunjukkan tingkat korelasi antara setiap pasangan variabel dalam kumpulan data. Matriks korelasi berguna dalam analisis multivariat untuk memahami hubungan antara variabel dan mengidentifikasi pola atau tren yang mungkin ada di antara mereka. Nilai korelasi berkisar antara -1 hingga 1, di mana nilai 1 menunjukkan korelasi positif sempurna, -1 menunjukkan korelasi negatif sempurna, dan 0 menunjukkan tidak adanya korelasi linier antara variabel. Berikut pada gambar 8 dan 9 ditampilkan hasil matrix korelasi.



Gambar 8. *Matrix korelasi*

Gambar 9. *Matrix korelasi*

4.4. Modelling

Sebelum melakukan pengimplementasia metode, jumlah dataset terbaru yaitu sebanyak 3749 data. Pada tahap ini data akan dibagi menjadi 2 bagian yaitu 20% untuk data uji (*testing*) dan 80% digunakan untuk data latih (*training*). Metode yang digunakan adalah metode Logistik Regres, K-Nearest Neighbors (KNN), Random Forest, dan Stochastic Gradient Descent (SGD). Keempat metode tersebut adalah metode

pembelajaran mesin yang termasuk dalam kategori supervised learning. Artinya, mereka menggunakan data latihan yang berlabel yang memiliki pasangan input dan output yang diketahui untuk melatih model.

Tujuan utama dari keempat metode ini adalah melakukan prediksi atau estimasi berdasarkan data input yang diberikan. Metode ini digunakan untuk membangun model yang dapat memprediksi kelas atau nilai target untuk data baru yang belum pernah dilihat sebelumnya. Selain itu Metode ini dapat digunakan baik untuk masalah klasifikasi (ketika output adalah variabel diskrit) maupun regresi (ketika output adalah variabel kontinu). Meskipun KNN cenderung lebih sering digunakan untuk klasifikasi, semua metode ini dapat diadaptasi untuk kedua tujuan tersebut. Berikut untuk tampilan hasil tingkat akurasi dari ke-empat metode tersebut.

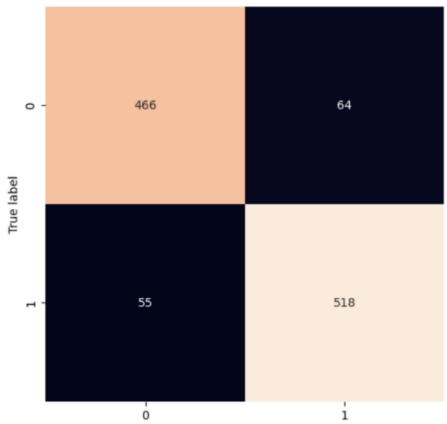
```
{'logistic regression': 0.6872166817769719,
 'knn': 0.8204895738893926,
 'random forest': 0.9038984587488668,
 'sgd': 0.485040797824116}
```

Gambar 10. *Modelling*

Dari hasil pengujian, dengan dilakukan evaluasi baik secara *confusion matrix* maupun nilai akurasi terbukti bahwa pengujian yang dilakukan dengan metode logistic regression, K-Nearest Neighbors (KNN), Random Forest, dan Stochastic Gradient Descent (SGD) memiliki nilai akurasi sebesar 68%, 82%, 90% dan 48%. Dari perbandingan hasil tingkat akurasi metode yang digunakan , metode dengan tingkat akurasi terbesar didapatkan oleh Random Forest dan tingkat akurasi terkecil adalah Stochastic Gradient Descent (SGD).

4.5. Counfusion Matrix

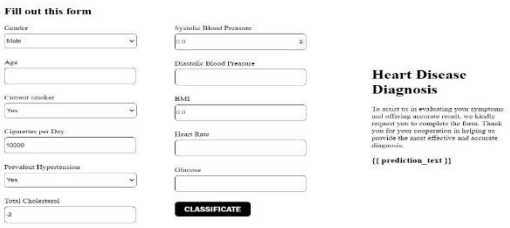
Untuk nilai Confusion Matrix yang dihasilkan, maka diperoleh nilai true positive sebesar 466, true negative sebesar 518, false positive sebesar 64, dan untuk nilai false negative nya sebesar 55.



Gambar 11. *Confusion Matrix*

4.6. Tampilan Interface

Pada website ini akan terlihat tampilan untuk melakukan klasifikasi dari penyakit hati. Pada tampilan ini user akan menginputkan data yang diminta pada website tersebut sesuai dengan gambar 12. Setelah mengisis inputan selanjutnya user menekan button classificate yang tersedia pada website sehingga akan muncul hasil dari klasifikasi berdasarkan inputan yang telah dimasukkan apakah user atau pasien terkena penyakit hati atau tidak.



Gambar 12. *Tampilan Website*

4.7. Hasil dan Analisis

Dari hasil penelitian yang telah dilakukan pada penelitian ini sehingga mendapatkan hasil nilai akurasi maupun secara confusion matrix terbukti bahwa pengujian yang dilakukan dengan metode logistic regression, K-Nearest Neighbors (KNN), Random Forest, dan Stochastic Gradient Descent (SGD) memiliki nilai akurasi sebesar 68%, 82%, 90% dan 48%. Dari perbandingan hasil tingkat akurasi metode yang digunakan, metode dengan tingkat akurasi terbesar didapatkan oleh Random Forest dan tingkat akurasi terkecil adalah Stochastic Gradient Descent (SGD). Sehingga dapat disimpulkan bahwa metode Random Forest dapat melakukan klasifikasi penyakit hati dengan baik.

5. KESIMPULAN

penelitian ini sehingga mendapatkan hasil nilai akurasi maupun secara confusion matrix terbukti bahwa pengujian yang dilakukan dengan metode logistic regression, K-Nearest Neighbors (KNN), Random Forest, dan Stochastic Gradient Descent (SGD) memiliki nilai akurasi sebesar 68%, 82%, 90% dan 48%. Dari perbandingan hasil tingkat akurasi metode yang digunakan, metode dengan tingkat akurasi terbesar didapatkan oleh Random Forest dan tingkat akurasi terkecil adalah Stochastic Gradient Descent (SGD)

Didalam penggunaanmetode harus memperhatikan setiap variable yang digunakan guna mengoptimalkan setiap metode yang digunakan.

DAFTAR PUSTAKA

[1] M. Nurkholifah, Jasmarizal, Y. Umar, and Rahmaddeni, “Analisa Performa Algoritma Machine Learning Dalam Prediksi Penyakit Liver,” *J. Indones. Manaj. Inform. dan Komun.*, vol. 4, no. 1, pp. 164–172, 2023, doi: 10.35870/jimik.v4i1.149.

[2] I. Setiawati, A. P. Wibowo, and A. Hermawan, “Pendahuluan Tinjauan Pustaka Penelitian

Sebelumnya Klasifikasi," *J. Inf. Syst. Manag.*, vol. 1, no. 1, pp. 13–17, 2019.

[3] F. L. D. Cahyanti, F. Sarasati, W. Astuti, and E. Firasari, "Klasifikasi Data Mining Dengan Algoritma Machine Larning Untuk Prediksi Penyakit Liver," *Technol. J. Ilm.*, vol. 14, no. 2, p. 134, 2023, doi: 10.31602/tji.v14i2.10093.

[4] D. Akuntansi, F. Ekonomika, and U. Gadjah, "Ekplorasi Pengaruh Levers Of Control terhadap Kinerja Kreatif pada Perusahaan Startup Indonesia: Studi Kasus pada Perusahaan Startup PT Bukalapak," vol. 11, no. 4, pp. 480–503, 2023.

[5] N. K. Dewi, S. Y. Mulyadi, and U. D. Syafitri, "Penerapan Metode Random Forest Dalam Driver Analysis," *Forum Stat. Dan Komputasi*, vol. 16, no. 1, pp. 35–43, 2012, [Online]. Available: http://journal.ipb.ac.id/index.php/statistika/article/view/5443

[6] D. I. Muhammad, E. Ermatita, and N. Falih, "Penggunaan K-Nearest Neighbor (KNN) untuk Mengklasifikasi Citra Belimbing Berdasarkan Fitur Warna," *Inform. J. Ilmu Komput.*, vol. 17, no. 1, p. 9, 2021, doi: 10.52958/iftk.v17i1.2132.

[7] A. Desiani, "Perbandingan Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Penyakit Hati," *Simkom*, vol. 7, no. 2, pp. 104–110, 2022, doi: 10.51717/simkom.v7i2.96.

[8] A. Ramadhan, A. Huda, Syerlie Annisa, and I. Efendi, "Perancangan Rule-Based Classification bagi Guru Baru Teknik Informatika," *SATIN - Sains dan Teknol. Inf.*, vol. 8, no. 2, pp. 142–151, 2022, doi: 10.33372/stn.v8i2.919.

[9] N. M. Farhan and B. Setiaji, "Indonesian Journal of Computer Science," *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

[10] L. A. Dewi, "Klasifikasi Machine Learning Untuk Mendeteksi Penyakit Jantung Dengan Algoritma K-Nn, Decision Tree dan Random Forest," *Repository.Uinjkt.Ac.Id*, 2023, [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/71124%0Ahttps://repository.uinjkt.ac.id/dspace/bitstream/123456789/71124/1/LIZKY ASKA DEWI-FST.pdf

[11] S. Royal, "METODE K-NEAREST NEIGHBOR DALAM Lestari," vol. 1, no. 1, pp. 1–9, 2024.

[12] S. Alacsel, "Forward Chaining Methods in Expert Systems in the Prevention and Treatment of Carp Diseases," *J. Inf. dan Teknol.*, vol. 5, pp. 13–18, 2022, doi: 10.37034/jidt.v5i1.227.

[13] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 6–10, 2020, [Online]. Available: http://bit.ly/datasetcardio.

[14] M. Erkamim, S. Suswadi, M. Z. Subarkah, and E. Widarti, "Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM," *J. Sist. Inf. Bisnis*, vol. 13, no. 2, pp. 127–134, 2023, doi: 10.21456/vol13iss2pp127-134.

[15] A. Ramadhan, B. Susetyo, and Indahwati, "Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan," *J. Pendidik. dan Kebud.*, vol. 4, no. 2, pp. 169–182, 2019, doi: 10.24832/jpnk.v4i2.1327.