Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung

Deo Haganta Depari[1], Yuni Widiastiwi[2], Mayanda Mega Santoni[3]
[1,2,3]Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta
[1,2,3]Jl. RS. Fatmawati Raya, Pd. Labu, Depok, Jawa Barat 12450
deohd@upnvj.ac.id[1], widiastiwi@yahoo.com[2], megasantoni@upnvj.ac.id[3]

Abstrak. Jantung sebuah rongga organ berotot yang memompa darah melalui pembuluh darah dengan kontraksi berirama yang terus berulang merupakan salah satu organ manusia yang berperan dalam sistem peredaran darah. Jantung sebagai salah organ terpenting dalam tubuh memiliki resiko kematian jika ada kelainan yang terjadi pada jantung. Beberapa masalah pada jantung dibagi menjadi dua yaitu penyakit jantung dan serangan jantung. WHO berdasarkan data menyatakan bahwa ada sebanyak 7,3 juta penduduk di dunia yang meninggal dikarenakan penyakit jantung. Penelitian ini menggunakan kumpulan data pasien penyakit jantung "Personal Key Indicators of Heart Disease" dan menerapkan algoritma klasifikasi Decision Tree, Naive Bayes dan Random Forest. Tujuan dari penelitian ini adalah untuk bagaimana mengolah dan melakukan analisa data, bagaimana penerapan metode Decision Tree, Naive Bayes dan Random Forest pada klasifikasi penyakit jantung, kemudian bagaimana hasil akurasi metode-metode yang digunakan tersebut, bagaimana hasil perbandingan antara Decision Tree, Naive Bayes dan Random Forests yang digunakan dan metode apa yang merupakan terbaik dari klasifikasi penyakit jantung. Hasil dari penelitian ini adalah evaluasi performa metode klasifikasi Decision Tree, Naive Bayes dan Random Forest. Dimana nilai akurasi metode Decision Tree sebesar 0.71%, Naive Bayes sebesar 0.72% dan Random Forest sebesar 0.75%.

Kata Kunci: Penyakit Jantung, Decision Tree, Naive Bayes, Perbandingan

1 Pendahuluan

Jantung sebuah rongga organ berotot yang memompa darah melalui pembuluh darah dengan kontraksi berirama yang terus berulang merupakan salah satu organ manusia yang berperan dalam sistem peredaran darah. Darah kemudian menyuplai oksigen dan nutrisi pada tubuh, dimana darah juga membantu menghilangkan sisa-sisa dari metabolisme. Letak Jantung berada di rongga dada sekitar sebelah kiri. Jantung sebagai salah organ terpenting dalam tubuh memiliki resiko kematian jika ada kelainan yang terjadi pada jantung. Beberapa masalah pada jantung dibagi menjadi dua yaitu penyakit jantung dan serangan jantung. WHO berdasarkan data menyatakan bahwa ada sebanyak 7,3 juta penduduk di dunia yang meninggal dikarenakan penyakit jantung. Tipe penyakit jantung terjadi dikarenakan jantung tidak dapat melaksanakan tugasnya dengan baik, seperti Otot jantung yang lemah atau adanya celah antara serambi kanan dan serambi kiri [9].

Mengambil data dari tulisan Center for Disease Control and Prevention, penyakit jantung merupakan penyebab utama meninggalnya wanita, pria dan beberapa kelompok ras dan juga etnis di Amerika Serikat, paling tidak ada satu orang yang meninggal setiap 36 detik dari penyakit jantung. Indonesia berdasarkan data Riset Kesehatan Dasar (Riskesdas) pada tahun 2018 memiliki angka kejadian penyakit jantung dan juga pembuluh darah yang terus meningkat dari tahun ke tahun, 15 dari 1000 orang atau sekitar 4,2 juta individu di Indonesia menderita penyakit jantung. Penyakit jantung yang terjadi perlu didiagnosa oleh dokter dengan menjalankan serangkaian tes dan evaluasi, yaitu pemeriksaan fisik, tes darah, tes dalam kategori non invasif seperti stress test dan Mengambil data dari tulisan Center for Disease Control and Prevention, penyakit jantung merupakan penyebab utama meninggalnya wanita, pria dan beberapa kelompok ras dan juga etnis di Amerika Serikat, paling tidak ada satu orang yang meninggal setiap 36 detik dari penyakit jantung. Indonesia berdasarkan data Riset Kesehatan Dasar (Riskesdas) pada tahun 2018 memiliki angka kejadian penyakit jantung dan juga pembuluh darah yang terus meningkat dari tahun ke tahun, 15 dari 1000 orang atau sekitar 4,2 juta individu di Indonesia menderita penyakit jantung. Penyakit jantung yang terjadi perlu didiagnosa oleh dokter dengan menjalankan serangkaian tes dan evaluasi, yaitu pemeriksaan fisik, tes darah, tes dalam kategori non invasif seperti stress test dan elektrokardiogram dan tes invasif seperti katerisasi pada jantung. Semua tes ini akan menghasilkan data pasien seperti kadar kolesterol, tekanan darah dan beberapa data lainnya yang kemudian dapat membantu dalam proses mendiagnosis penyakit pasien [10].

Industri kesehatan telah berkembang sangat pesat salah satunya dikarenakan teknologi yang berkembang pesat, sehingga perkembangan ini membuka pintu untuk melakukan lebih banyak penelitian. Kesehatan merupakan hal yang penting untuk dijaga, dimana kesalahan pengobatan atau pencegahan dapat menyebabkan kehilangan nyawa, terutama pada penyakit jantung. Peningkatan teknologi pada industri kesehatan salah satunya adalah melakukan digitalisasi informasi medis terutama data pasien, dengan harapan informasi yang telah digitalisasi dapat digunakan untuk penelitian dan hasil penelitian tersebut dapat meningkatkan layanan industri Kesehatan [11].

Tentunya perkembangan teknologi yang pesat ini mempunyai beberapa masalah yang perlu diperhatikan, seperti bagaimana penanganan informasi medis yang didapatkan akan mempunyai ukuran data yang cukup besar dikarenakan perkembangan industri, hal yang perlu kita perhatikan adalah bagaimana bisa kita uraikan, kemudian analisa sehingga kita mendapatkan informasi yang penting dan berguna dimana kemudian dapat diaplikasikan dalam layanan pada industri kesehatan. Machine Learning kemudian memegang peran penting pada proses penanganan dan analisa data medis, dengan Machine Learning kita dapat mencoba berbagai metode-metode yang tersedia untuk menemukan pola atau informasi yang penting dan juga berguna dimana diharapkan dapat meningkatkan kualitas layanan kesehatan saat ini.

Dalam beberapa penelitian yang memiliki kemiripan dalam melakukan klasifikasi penyakit jantung, metode-metode yang digunakan umumnya menghasilkan nilai akurasi yang berbeda. Penelitian Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes yang diterbitkan pada tahun 2019. Menggunakan metode Naive Bayes menghasilkan nilai rata rata akurasi sebesar 90,61%. Penelitian Analisis Penyakit Jantung Koroner Menggunakan Decision Tree yang diterbitkan pada tahun 2021 menjelaskan bahwa dengan menggunakan metode CART Decision Tree, mendapatkan akurasi sebesar 80%. Penelitian Prediction of Heart Diseases using Random Forest yang diterbitkan Maret 2021 menjelaskan bahwa penerapan klasifikasi pada 303 data penyakit jantung menggunakan algoritma Random Forest, mendapatkan hasil akurasi sebesar 86.69%.

Mengacu pada Teorema No Free Lunch dimana dijelaskan bahwa semua algoritma umum nya akan bekerja secara optimal ketika kinerjanya rata rata untuk setiap fungsi dan tujuan yang mungkin. Hal ini kemudian memberikan pertanyaan bagaimana jika ada tiga metode klasifikasi yang ingin digunakan, bagaimana caranya membandingkan kedua algoritma tersebut dan memilih algoritma yang terbaik. Berdasarkan penelitian terdahulu diatas, maka pada penelitian akan dilaksanakan dengan mendapatkan data, kemudian mengolahnya berdasarkan atribut yang relevan, lalu menggunakan metode Decision Tree, Naive Bayes dan Random Forest untuk mendapatkan hasil akurasi pada setiap metode model prediktif yang optimal. Penelitian ini diharapkan dapat membantu peneliti di bidang penyakit jantung dalam memilih model-model atau metode metode yang tepat didalam penelitian nya dan dapat membantu peningkatan layanan ahli medis.

2 Landasan Teori

Penelitian di memiliki beberapa tinjauan pustaka seperti *Data Mining, Machine Learning* dan penelitian sebelumnya yang relevan dengan penelitian ini.

2.1 *Machine Learning*

Suatu kumpulan teknik yang memungkinkan penggunanya untuk melakukan implementasi algoritma dimana memiliki sifat dapat beradaptasi untuk mengorganisir data masukkan berdasarkan fitur yang sama secara otomatis dan juga melakukan prediksi dikenal dengan nama Machine Learning [12]. Dalam beberapa dekade yang lalu, Machine Learning menjadi salah satu alat yang sangat umum untuk digunakan pada setiap pekerjaan yang membutuhkan ekstraksi informasi dari suatu kumpulan data dengan ukuran yang besar. Istilah Machine Learning sendiri menunjuk kepada suatu otomasi yang mendeteksi pola penting dalam data [13]. Scikit-Learn merupakan salah satu scientific libraries yang menjadi pilihan untuk penembangan algoritma. Scikit-Learn menyediakan berbagai implementasi dari algoritma machine learning yang cukup dikenal, dan di desain untuk mudah digunakan [14].

2.1.1 *Decision Tree*

Decision Tree merupakan suatu struktur yang mempunyai dasar dari proses dimana sifatnya sekuensial. Awal proses akan dimulai dari akar, kemudian dilanjut dengan melakukan evaluasi pada suatu fitur dan salah satu dari dua cabang yang dimiliki, kemudian dilakukan terus menerus hingga cabang terakhir atau yang disebut daun, dimana umumnya menjadi suatu representasi target yang dicari sudah tercapai. Dikarenakan struktur dari decision tree sendiri tidak terkena pengaruh oleh nilai yang ada pada setiap fitur data, maka mudah untuk decision tree bekerja secara efisien pada kumpulan data yang sebelumnya tidak melalui pra proses normalisasi Scikit-learn pada Decision Tree dapat melatih binary decision tree dengan ukuran Gini dan cross-entropy impurity.
Gini Impurity didefinisikan dengan rumus berikut:

$$I\ \boldsymbol{gini}(\boldsymbol{x}) = \sum_x \quad p(x|y)(1-p(x|y)) \qquad (1)$$

Dimana total akan selalu diperpanjang ke semua kelas yang ada. Hal inilah yang umum untuk dijadikan suatu ukuran dan juga digunakan sebagai nilai bawaan di dalam Scikit-Learn. Jika diberikan suatu sampel maka Gini Impurity mengukur kemungkinan kesalahan klasifikasi suatu label jika suatu label secara acak di pilih dengan probabilitas distribusi cabang. Cross-entropy Impurity didefinisikan dengan rumus berikut:

$$I\ \boldsymbol{cross-entropy}\ (\boldsymbol{x}) = -\sum_x \quad p(x|y)\ log\ p(x|y) \qquad (2)$$

Ukuran ini didasarkan dari teori informasi dan berasumsi tidak adanya nilai hanya ketika suatu sampel yang dimiliki suatu kelas individu hadir pada pemisahan tetapi akan bernilai maksimum ketika ada distribusi yang seragam antar kelas. Serupa dengan Gini Impurity hanya saja memiliki perbedaan yaitu cross-entropy mengizinkan penggunanya untuk memilih pemisahan yang meminimalisir ketidakpastian di dalam klasifikasi, sedangkan Gini impurity akan langsung meminimalisasi kemungkinan kesalahan dalam melakukan klasifikasi [4].

2.1.2 Naive Bayes

Naive Bayes merupakan suatu kelompok yang tidak saja mudah untuk dilatih penggolonganya tetapi juga memiliki sifat yang kuat didalam menentukan suatu probabilitas dari suatu hasil yang telah diberikan dari beberapa kondisi menggunakan teorema bayes. Teorema bayes didapatkan dari contoh jika ada suatu kejadian yang mempunyai kemungkinan terjadi, contohnya kegiatan x dan kegiatan y, jika adanya korelasi antara dua kegiatan tersebut, dengan menggunakan peraturan produk maka kita mendapatkan rumus:

$$\langle P\ (x\ \cap\ y) = P\ (x \mid y)\ P\ (y) \qquad (3)$$
$$\langle P\ (y\ \cap\ x) = P\ (y \mid x)\ P\ (x) \qquad (4)$$

Dengan pertimbangan bahwa kejadian x dan kejadian y bersimpangan dan memiliki sifat komutatif, anggota pertama nya sama maka disinilah kita mendapatkan teorema Bayes:

$$P\ (x \mid y) = P\ (y \mid a)\ P\ (x)\ /\ P\ (y) \qquad (5)$$

Kondisi naif adalah alasan kenapa adanya penggolong Naive Bayes, hal ini juga mengimplikasikan bahwa adanya independensi yang mempunyai syarat dari sebab-sebab yang ada.

Scikit-learn pada Naive Bayes mengimplementasikan tiga variasi Naive Bayes yang didasarkan dengan banyaknya kemungkinan distribusi. Salah satu dari tiga variasi itu adalah Gaussian, distribusi yang sifatnya kontinu dengan ciri ciri dari rata rata dan variansinya. Gaussian Naive Bayes sangat berguna ketika bekerja dengan nilai yang terus menerus dimana probabilitasnya dapat dimodelkan dengan distribusi Gaussian:

$$P(a) = \frac{1}{\sqrt{2\pi\sigma^{\ 2}}} e^{-\frac{(a-\mu)^{\ 2}}{2\sigma^{\ 2}}} \qquad (6)$$

Kondisi dari kemungkinan P($a_{\ i}$|b) juga disebar secara Gaussian, maka dari itu, sangat penting untuk melakukan estimasi rata rata dan juga variansi dari setiap pendekatan yang maksimal. Rumus yang didapat seperti berikut:

$$L(\mu; \sigma^2; a_i|b) = log \prod_k P(a_i^{(k)}|b) = \sum_k log\, P(a_i^{(k)}|b) \quad (7)$$

Nilai *k* menunjuk ke sampel yang ada di kumpulan data dan $P(a_i^{(k)}|b)$adalah *Gaussian* [4].

2.1.3 Random Forest

Random Forest merupakan suatu set dari decision trees yang dibangun dengan sampel yang dipilih secara acak tetapi memiliki peraturan membelah simpul yang berbeda. Model ini bekerja dengan menggunakan subset dari suatu fitur pada setiap pohon, kemudian mencoba mencari ambang batas terbaik didalam memisahkan data. Sehingga hasilnya akan memiliki banyak pohon yang dilatih dengan cara yang lebih lemah dan masing-masing akan menghasilkan prediksi yang berbeda.

Hasil tersebut kemudian dapat diinterpretasi dengan dua cara, yang paling umum adalah berdasarkan suara terbanyak sehingga akan di pertimbangkan sebagai kelas yang benar. Tetapi scikit-learn melakukan implementasi algoritmanya berdasarkan rata-rata dari hasil tersebut sehingga menghasilkan prediksi yang sangat akurat. Sehingga walaupun a secara teoritis berbeda, rata-rata probabilitas dari Random Forest yang terlatih tidak dapat sangat berbeda dari sebagian besar prediksi (jika tidak, harus ada titik stabil yang berbeda); oleh karena itu kedua metode tersebut sering kali mengarah ke hasil yang sebanding. Untuk model Random Forest dengan scikit-learn ada beberapa parameter yang dapat diatur, seperti untuk menentukan jumlah pohon yang ingin model ini bangun(n_estimators).

2.1.4. Imbalanced Data

Imbalanced Data merupakan kumpulan data yang memiliki proporsi kelas yang tidak seimbang. Kelas yang porsi nya lebih banyak disebut kelas mayoritas, sedangkan kelas yang porsi nya lebih sedikit disebut kelas minoritas. Hal ini kemudian dapat menjadi suatu masalah dikarenakan dengan sedikitnya data dengan kelas positif, maka model akan lebih banyak mengambil waktu untuk belajar data negatif dibanding data positif. Hal ini kemudian menyebabkan penilaian model *Machine Learning* menjadi bias akan data baru atau data uji, sehingga tidak hanya model cenderung overfitting, tetapi juga performa dalam melakukan klasifikasi pada data positif sangatlah berkurang. Berikut Tabel derajat ketidakseimbangan dengan proporsi kelas minoritas [15].

Tabel. 1 Tabel derajat ketidakseimbangan dengan proporsi kelas minoritas

Derajat ketidakseimbangan	Proporsi Kelas Minoritas
Ringan	20-40% dari kumpulan data
Sedang	*1-20% dari kumpulan data*
Ekstrim	*<1 % dari kumpulan data*

2.2. *Data Mining*

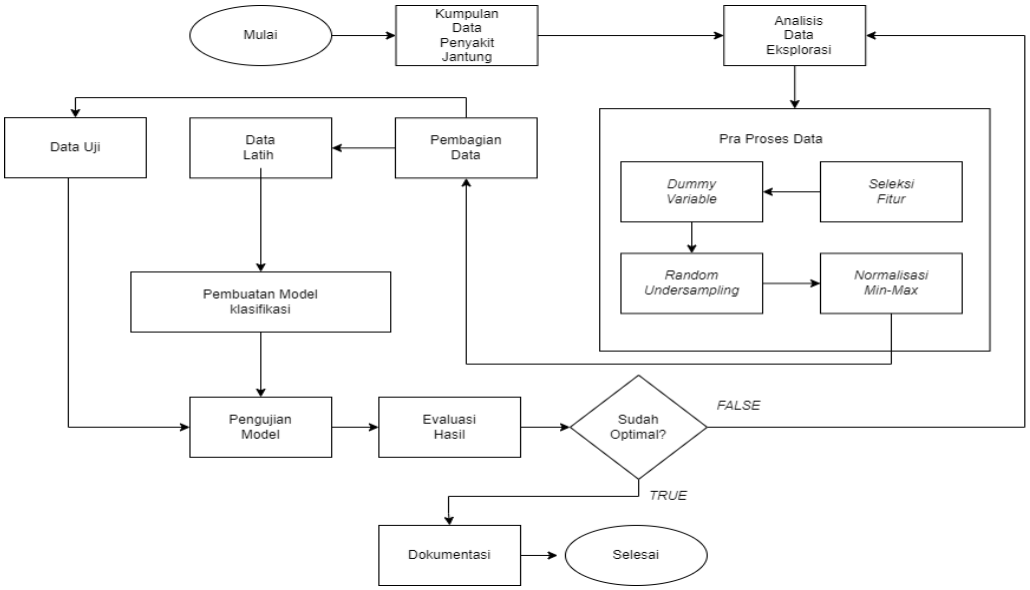
Data mining adalah bidang yang tergabung dari beberapa bagian keilmuan, dimana bidang ini mempergabungkan beberapa teknik yaitu mempelajari suatu mesin, mengenal pola, melakukan analisis yang bersifat otomatis dari suatu kumpulan data yang jumlah nya besar atau kumpulan data yang sifatnya kompleks dengan tujuan untuk mendapatkan suatu informasi seperti pola dan hubungan antar data, yang kemudian dapat digunakan dalam membuat visualisasi, statistik atau basis data [16].

a. Pemilihan Data, dilakukan sebelum menjalankan proses untuk mendapatkan suatu informasi. Data yang terpilih dari sekumpulan data sebelumnya kemudian akan disimpan pada tempat yang terpisah.
b. Praproses atau pembersihan merupakan tahap yang perlu dilakukan, proses pembersihan terdiri dari memeriksa apakah ada data yang bersifat inkonsisten, membuang data duplikat yang ada dan memperbaiki setiap kesalahan yang ada pada data tersebut contohnya seperti kesalahan pengetikan, penulisan atau pencetakan (tipografi). Selain itu data ini akan diperkaya dengan data atau informasi yang sifatnya relevan baik dari internal atau eksternal.

c. Transformasi adalah suatu proses yang tidak akan sama untuk setiap data dikarenakan proses ini tergantung dengan jenis atau informasi yang akan dicari. Proses akan membantu data menjadi sesuai untuk digunakan dalam proses data mining
d. Data Mining, proses Data Mining seperti terjemahan kasar nya adalah menggali data tersebut, dengan tujuan untuk mendapatkan suatu pola atau informasi yang sebelumnya tidak ada dengan teknik atau metode yang bermacam dan bervariasi. Teknik, metode dan juga algoritma memegang peran penting dan sangat bergantung dalam proses keseluruhannya
e. Penafsiran atau Evaluasi dari hasil proses data mining baik hasil yang merupakan suatu pola data atau informasi sangat penting untuk diperiksa apakah hasil tersebut konflik dengan hipotesis, kesimpulan atau informasi yang sudah ada sebelum proses ini, kemudian hasilnya perlu diubah menjadi bentuk yang mudah untuk dimengerti oleh pihak pihak yang mempunyai kepentingan dalam keseluruhan proses ini.

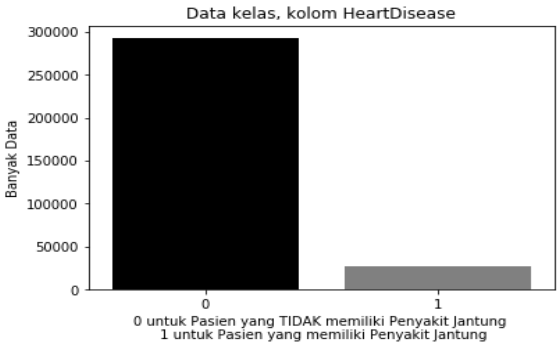
3 Hasil Pembahasan

Metodologi Penelitian terdiri dari beberapa tahapan yaitu dari Pengumpulan Data Penyakit Jantung, pra proses Data, kemudian pembagian Data Latih dan Data Uji, Pembuatan model klasifikasi, proses Klasifikasi setiap model dan Evaluasi hasil setiap model, kemudian jika belum optimal atau masih dapat ditingkatkan maka proses tahapan penelitian diulang kembali hingga hasil nya sudah optimal dan kemudian penelitian dapat didokumentasi lalu selesai, seperti yang terlihat pada Gambar 1.



Gambar. 1. *Flowchart* Metode Penelitian

Pengumpulan Data Penyakit Jantung mendapatkan kumpulan data yang berjudul "Personal Key Indicators of Heart Disease". Diajukan oleh Kamil Pytlak, format kumpulan data adalah csv. Awalnya Kumpulan data penyakit jantung ini berasal dari CDC dan bagian dari Behavioral Risk Factor Surveillance System (BRFSS), yang melakukan survey melalui telepon untuk mendapatkan data di 50 negara bagian, distrik Kolombia dan tiga wilayah Amerika Serikat. Kumpulan data yang paling baru adalah 15 Februari 2022, dimana terdiri dari 401958 baris dan 279 kolom. Kemudian oleh Kamil Pyltak Kumpulan diolah sehingga data yang digunakan pada penelitian ini adalah pilihan kolom/variabel yang relevan, kemudian dibersihkan sehingga bisa digunakan untuk keperluan Machine Learning. (Kamil Pyltak, 2022). Sehingga data yang digunakan pada penelitian ini adalah hasil proses pengolahan oleh Kamil Pyltak, dimana kumpulan data memiliki 18 kolom dan 319795 baris. Kemudian proses analisis data eksplorasi dilanjut dengan melihat banyak jumlah pasien yang memiliki penyakit jantung dan yang tidak memiliki penyakit jantung, hasilnya sebagai berikut:



Gambar. 2. Grafik Batang Perbandingan data HeartDisease pada pasien yang memiliki dan tidak memiliki penyakit jantung.

```
Jumlah data kelas HeartDisease 319795
Jumlah data dibagi dengan kelas
0    292422
1     27373
Name: HeartDisease, dtype: int64
0 untuk Pasien yang TIDAK memiliki Penyakit Jantung
1 untuk Pasien yang memiliki Penyakit Jantung
```

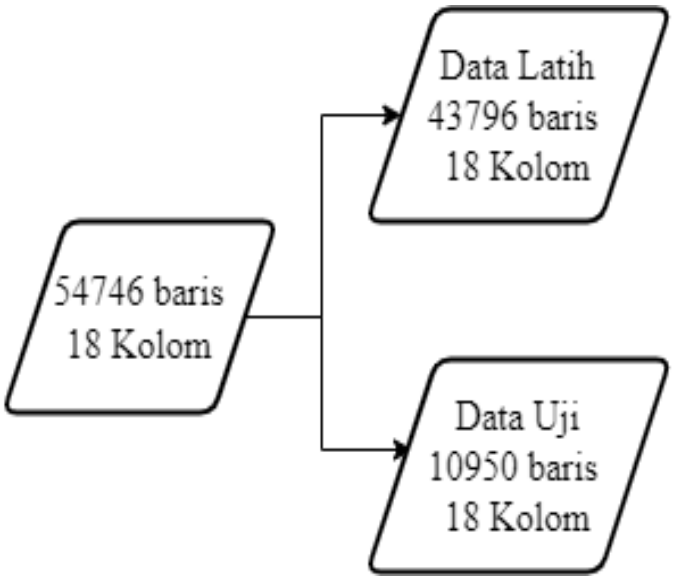
Gambar. 3. Perbandingan data HeartDisease pada pasien yang memiliki dan tidak memiliki penyakit jantung.

Pada Gambar 2 dan 3, yang menunjukkan Grafik Batang dan angka Perbandingan data HeartDisease pada pasien yang memiliki dan tidak memiliki penyakit jantung. Terlihat bahwa di kumpulan data, kasus pasien yang tidak memiliki penyakit jantung sangat signifikan, sebesar 91 %. Sehingga hal ini menciptakan bias dan ada kemungkinan besar hasil klasifikasi dengan data uji akan selalu bernilai "0" (Pasien yang tidak memiliki penyakit jantung). Untuk kasus kumpulan data seperti ini, penelitian ini akan melakukan pengolahan kumpulan data dengan teknik Random Undersampling, sehingga proporsi data dari pasien yang memiliki dan tidak memiliki penyakit jantung menjadi sama. Kumpulan data yang digunakan akan di pra proses terlebih dahulu dengan tujuan untuk memaksimalkan hasil dari klasifikasi, beberapa tahapan pada pra proses yaitu:

a. label Encoding, proses mengubah suatu fitur yang memiliki nilai kategori, dimana nilai dari setiap kolom akan diganti menjadi angka.
b. Random Undersampling, dilakukan karena kumpulan data pasien yang tidak memiliki penyakit jantung lebih banyak proporsi nya dibandingkan dengan pasien yang memiliki penyakit jantung. Hasil dari proses ini adalah kumpulan data dimana kasus pasien penyakit jantung dan tidak penyakit jantung dengan proporsi nya sama
c. Normalisasi data, dilakukan untuk memastikan nilai kumpulan data menjadi ke dalam skala umum

Penelitian akan dilanjutkan dengan membagi data yang telah di pra proses pada tahap sebelumnya menjadi dua bagian, data latih dan data uji. Menggunakan fungsi train_test_split yang di ambil dari modul sklearn.model_selection, kumpulan data sebanyak 54746 baris dan 18 kolom dibagi menjadi 80% data latih sebesar 43796 baris 18 kolom dan 20% data uji sebesar 10950 baris 18 kolom.

ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)



Gambar. 4. *Pembagian data menjadi data uji dan data latih*

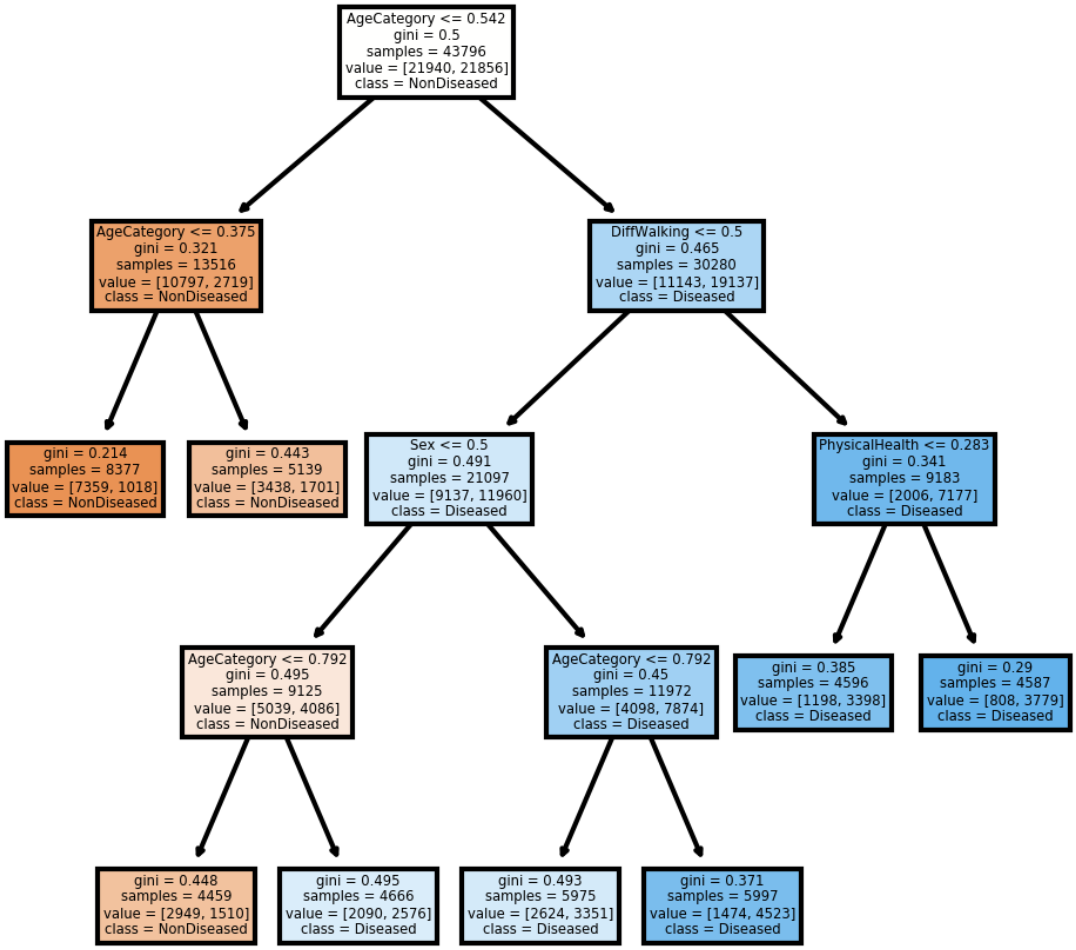
Pada tahap Pembuatan model dan tahapan klasifikasi model, dari data yang telah di praproses dan dibagi menjadi data fitur latih, data kelas latih, data fitur uji dan data kelas uji. Penelitian akan dilanjutkan dengan membuat model klasifikasi Naive Bayes, Decision Tree, dan Random Forest yang diambil dari modul sklearn, kemudian setiap model klasifikasi akan melalui pelatihan dengan masukkan data fitur latih dan data kelas latih. Untuk model Decision Tree parameter untuk jumlah minimum sampel yang diperlukan untuk memisahkan simpul internal(min_samples_split) akan diberikan nilai 2. Untuk model Random Forest parameter untuk menentukan jumlah pohon yang ingin model ini bangun(n_estimators) akan diberikan nilai 1000, dimana semakin besar nilai tersebut program akan berjalan lebih lambat tetapi akan memberikan kinerja yang terbaik. Setelah melalui pelatihan setiap model akan diuji melalui tahapan klasifikasi menggunakan data fitur uji, dimana hasil klasifikasi oleh setiap model tersebut kemudian dibandingkan dengan data kelas uji. Akurasi dari setiap model akan dihitung berdasarkan berapa banyak hasil klasifikasi yang dilakukan oleh model sesuai dengan data latih kelas.

Pada tahap Hasil dan Evaluasi Model, dari proses Pembuatan model dan klasifikasi penelitian akan dilanjutkan dengan mengevaluasi performa setiap model. Tabel Data Uji yang digunakan sebagai berikut:

Tabel. 2 Tabel Data Uji

Index	HeartDisease Test	BMI	Smoking	…	SkinCancer
1	0	0.1663648437	1	...	*0*
2	*0*	0.3128093686	1		0
3	*1*	0.2258843414	0		1
4	*1*	0.6188579017	1		0
…	...	…	…		…
5	*1*	0.3370759387	0		0

Visualisasi dari Model Decision Tree sebagai berikut:



Gambar. 5. Visualisasi *Decision Tree*

Pada Gambar 5 terlihat bahwa yang menjadi root adalah fitur AgeCategory dengan nilai kurang atau sama dengan 0.542. Berikut 5 Peraturan yang berlaku pada decision tree dalam menentukan kelas data uji.

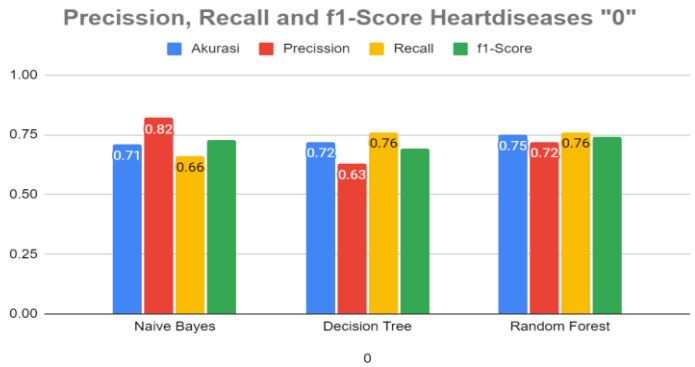
a. Peraturan 1: Jika nilai AgeCategory lebih kecil atau sama dengan 0.375, maka data akan di klasifikasikan sebagai pasien tidak memiliki penyakit jantung
b. Peraturan 2: Jika nilai AgeCategory lebih kecil atau sama dengan 0.542, maka data akan di klasifikasikan sebagai pasien tidak memiliki penyakit jantung
c. Peraturan 3: Jika nilai AgeCategory lebih kecil atau sama dengan 0.542, nilai DiffWalking lebih kecil atau sama dengan 0.5, nilai Sex <= 0.5, Maka data akan di klasifikasikan sebagai pasien tidak memiliki penyakit jantung.
d. Peraturan 4: Jika nilai AgeCategory lebih kecil atau sama dengan 0.542, nilai DiffWalking lebih kecil atau sama dengan 0.5, nilai Sex lebih besar dari 0.5 Maka data akan di klasifikasikan sebagai pasien yang memiliki penyakit jantung.
e. Peraturan 5: Jika nilai AgeCategory lebih kecil atau sama dengan 0.542, nilai DiffWalking lebih kecil atau sama dengan 0.5 Nilai PhysicalHealth lebih kecil atau sama dengan 0.283 Maka data akan di klasifikasikan sebagai pasien yang memiliki penyakit jantung

Berikut Tabel 3, Gambar 6 dan 7 yang menunjukan bagaimana evaluasi dan perbandingan hasil waktu eksekusi metode dan akurasi antara model Naive Bayes, Decision Tree dan Random Forest:

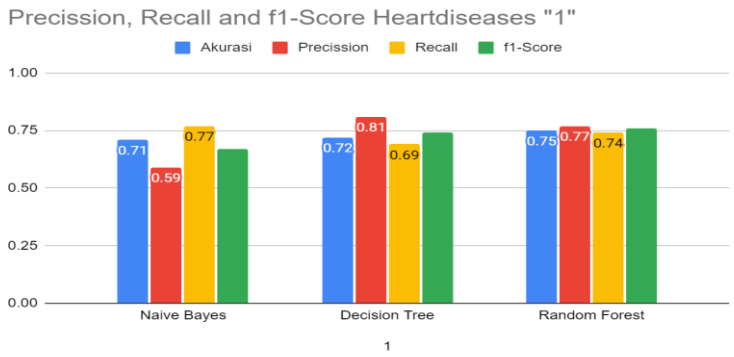
ISSN: 2655-139X (ONLINE)
ISSN: 0216-4221 (PRINT)

Tabel. 3 Tabel Hasil Evaluasi Model

			Precission	Recall	F1-Score
Naive Bayes	Kelas 0		0.82	*0.66*	*0.73*
	Kelas 1		0.59	0.77	0.67
	Waktu Komputasi	0.189 detik			
	Akurasi	0.71			
Decision Tree	*Kelas 0*		0.63	0.76	0.69
	Kelas 1		0.81	0.69	0.74
	Waktu Komputasi	0.188 detk			
	Akurasi	0.72			
Random Forest	*Kelas 0*		0.72	0.76	0.74
	Kelas 1		0.77	0.74	0.76
	Waktu Komputasi	69 detik			
	Akurasi	0.75			



Gambar. 6. Grafik Batang perbandingan Akurasi, Precission, Recall dan F1-Score untuk data dengan kelas pasien tidak memiliki penyakit jantung



Gambar. 7. Grafik Batang perbandingan Akurasi, Precission, Recall dan F1-Score untuk data dengan kelas pasien yang memiliki penyakit jantung

4 Kesimpulan dan Saran

Berdasarkan penelitian yang telah dilakukan dan hasil yang telah didapatkan maka dapat disimpulkan beberapa hal:

a. Nilai precision tertinggi pada pasien yang pasien tidak memiliki penyakit jantung (0) dan yang memiliki penyakit jantung (1) didapatkan oleh model Naive Bayes dan Decision Tree
b. Nilai Recall tertinggi pada pasien yang tidak memiliki penyakit jantung (0) dan yang memiliki penyakit jantung (1) didapatkan oleh model Decision Tree juga Random Forest dan Naive Bayes
c. Nilai f1-score tertinggi pada yang tidak memiliki penyakit jantung (0) dan yang memiliki penyakit jantung (1) didapatkan oleh model NaiveBayes juga Random Forest dan Random Forest
d. Nilai accuracy model Naive Bayes sebesar 71%, model Decision Tree sebesar 72% dan nilai Random Forest sebesar 75%.
e. Model Random Forest merupakan metode terbaik untuk digunakan dikarenakan hasil klasifikasi nya yang paling tinggi yaitu sebesar 75%.

Untuk kasus metode klasifikasi dimana dibutuhkan juga kecepatan mendapatkan hasil, maka Model Random Forest tidak lagi menjadi metode yang terbaik, dikarenakan untuk mendapatkan hasil akurasi yang tinggi maka semakin banyak juga jumlah pohon yang model ini perlu bangun sehingga mengakibatkan waktu yang lebih lama, dari hasil penelitian juga terlihat bahwa Random Forests membutuhkan waktu yang sangat signifikan jauh yaitu 69 detik dengan banyak pohon 1000 untuk mendapatkan akurasi 0.75%. Sehingga metode Decision Tree pada kasus ini merupakan model terbaik dikarenakan memiliki akurasi sebesar 72% dan dengan waktu eksekusi sebesar 0.118 detik
adsasd
Saran untuk penelitian selanjutnya adalah untuk melakukan banyak percobaan didalam proses menangani juga mengolah data dan pembuatan model, sehingga meningkatkan akurasi, kecepatan proses klasifikasi dan tetap menghindari overfitting data.

Referensi

[1] T. Ariwibowo, "Perbandingan Metode Imputasi Mean, Median, Modus, Dan 1-Nn Pada Hasil Klasifikasi K-Nearest Neigbour (K-Nn)," *Universitas Pembangunan Nasional "Veteran" Jakarta*, 2019.
[2] R. Donovan, "Heart Disease: Risk Factors, Prevention, and More." https://www.healthline.com/health/heart-disease (diakses Jul 20, 2022).
[3] NEJM Catalyst, "Healthcare Big Data and the Promise of Value-Based Care," 2018, [Daring]. Tersedia pada: https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0290.
[4] G. Bonaccorso, "Machine Learning Algorithms: A reference guide to popular algorithms for data science and machine learning," 2017.
[5] S. Shalev-Shwartz dan S. Ben-David, "Understanding Machine Learning," *Cambridge University Press*, 2014.
[6] F. É. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, dan Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res. 12*, 2011.
[7] Google Developers, "Imbalanced Data." https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data (diakses Jul 20, 2022).
[8] Kusrini dan E. T. Luthfi, "Algoritma Data Mining," *Yogyakarta: CV. ANDI OFFSET*, 2009.
[9] World Health Organization, "Cardiovascular diseases."
[10] M. Pal dan S. Parija, "Prediction of Heart Diseases using Random Forest," *J. Phys. Conf. Ser.*, vol. 1817, no. 1, hal. 012009, Maret 2021, doi: 10.1088/1742-6596/1817/1/012009.