

PROPOSAL TUGAS AKHIR
KLASIFIKASI PENYAKIT JANTUNG
MENGGUNAKAN ALGORITMA
CORRELATED NAÏVE BAYES



Disusun Oleh:

Stevani Maria Meilissa Sapca Sagurung

185314132

PROGRAM STUDI INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS SANATA DHARMA
YOGYAKARTA

2022

BAB I

PENDAHULUAN

Bagian pendahuluan akan memberikan informasi tentang latar belakang penelitian rumusan masalah, tujuan, batasan masalah, metodologi penelitian dan sistematika penulisan.

1.1. Latar Belakang

Penyakit jantung adalah kondisi dimana jantung mengalami gangguan. Gangguan itu sendiri terdiri dari bermacam-macam, bisa berupa gangguan pada pembuluh darah jantung, katup jantung, atau otot jantung. Penyakit jantung juga dapat disebabkan oleh infeksi atau kelainan lahir. *Cardiovascular Diseases* (CVD) merupakan salah satu penyakit mematikan nomor satu didunia. Berdasarkan laporan *World Health Organization* (WHO) angka kematian yang disebabkan oleh penyakit *Cardiovascular Disease* (CVD) mencapai 17,7 juta orang setiap tahunnya dan 31% merupakan penyebab dari seluruh kematian global. Angka kematian akibat penyakit *Cardiovascular Diseases* diprediksi akan terus meningkat dari tahun ke tahun dan di perkirakan pada tahun 2030 akan mencapai 23,3 juta kematian. Salah satu penyakit *Cardiovascular Diseases* ialah penyakit jantung. Berdasarkan riset kesehatan dasar (Riskesdas) tahun 2018, angka kejadian penyakit jantung dan pembuluh darah semakin meningkat dari tahun ke tahun. Setidaknya, 15 dari 1000 orang atau sekitar 2,7 juta individu di Indonesia menderita penyakit jantung. Penyakit jantung merupakan organ manusia yang berperan dalam system peredaran darah. Penyakit jantung adalah sebuah kondisi dimana jantung tidak dapat melaksanakan tugasnya dengan baik. Data WHO menyatakan bahwa sebanyak 7,3 juta penduduk dunia meninggal dikarenakan penyakit jantung. Meskipun penyakit jantung merupakan penyakit yang tidak menular, penyakit ini merupakan jenis penyakit yang mematikan nomor satu di dunia. Setiap tahun, lebih dari 2 juta orang amerika meninggal karena penyakit jantung/stroke.

Penyakit jantung merupakan gangguan yang terjadi pada system pembuluh darah besar. Sehingga menyebabkan jantung dan peredaran darah tidak bekerja sebagaimana mestinya. Penyakit jantung sering dikenal sebagai “*sudden death*”. Tingginya factor kematian akibat penyakit jantung dapat di cegah dan ditekan factor

resikonya, kurangnya pengetahuan masyarakat tentang gejala penyakit jantung. Kurang akuratnya peralatan yang digunakan jika hanya mengontrol gula darah dan tekanan darah, dan gaya hidup yang tidak sehat. Data laboratorium yang belum difungsikan secara efektif bisa digunakan untuk deteksi penyakit jantung. Dari data dapat diketahui bahwa banyak orang yang belum menanggapi penyebab penyakit ini dengan serius dan setelah melakukan pemeriksaan kesehatan dokter mendeteksi adanya penyakit dengan stadium yang sudah tinggi. Banyak alternative cara untuk mencegah bahkan menyembuhkan penyakit-penyakit tersebut seperti dengan melakukan operasi, penyinaran dan khemoterapi. Namun, kurangnya akses informasi/media menjadi alasan penderita terlambat untuk memeriksa diri ke dokter.

Terdapat hubungan antara kurangnya akses informasi/media dengan keterlambatan pemeriksaan awal penyakit jantung. Kurangnya akses untuk mencari informasi tentang penyakit serangan jantung ini menyebabkan peningkatan kematian setiap tahunnya. Karena itu, dibutuhkan sebuah system klasifikasi yang dapat memberikan informasi tentang penyakit serangan jantung serta dapat melakukan pengecekan klasifikasi secara dini tentang penyakit serangan jantung yang dialami oleh seseorang. Untuk melakukan sebuah klasifikasi system membutuhkan metode yang tepat dalam mengelola pengetahuan yang diadopsi dari pakar sehingga diperoleh hasil yang akurat.

Penelitian yang dilakukan oleh Bianto et al., (2020), membuktikan bahwa melalui hasil penelitian telah dijelaskan dan dilakukan pada pembuatan sistem klasifikasi penyakit jantung menggunakan *Naïve Bayes*. Pembuatan sistem ini menyimpulkan hasil akurasi dengan rata-rata akurasi senilai 90,61% rata-rata hasil nilai presisi senilai 87,44% dan rata-rata nilai recall senilai 87,95% dengan konfigurasi data yang terdapat pada *UCI Machine Learning* yang berisi 2 kelas klasifikasi dan 15 atribut dengan jumlah 303 data.

Menurut penelitian oleh Putra & Rini, (2019), menjelaskan bahwa penyakit jantung merupakan salah satu dari jenis PTM yang rentan menyerang terutama pria dengan usia dibawah 60 tahun. Oleh sebab itu penelitian ini berfokus untuk menyelidiki suatu algoritma, apakah memiliki tingkat akurasi yang tinggi guna pendeteksi penyakit jantung melalui objek menggunakan dataset (*heart disease*).

Berdasarkan cross validation dengan masing-masing algoritma yang ditetapkan, sehingga menghasilkan akurasi algoritma *Naive Bayes* 84,07%, *Support Vector Machine* 81,85%, *C.45* 74,81%, *Logistic Regression* 82,59%, *Back Propagation* 81,85%. Setelah mengeksekusi dataset dengan algoritma yang dipilih didapatkan algoritma *Naive Bayes* dengan tingkatan akurasi tertinggi dalam penelitian ini.

Berdasarkan uraian diatas, penelitian ini mencoba membangun system untuk melihat tingkat prediksi untuk melakukan klasifikasi penyakit jantung dengan menggunakan metode *Correlated Naive Bayes*. Implementasi metode ini juga akan menghitung tingkat akurasi, *presisi*, *recall* dan *f1-score* dari algoritma *Correlated Naive Bayes* dalam melakukan klasifikasi penyakit jantung seseorang.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka rumusan masalahnya adalah:

1. Bagaimana menyusun data penyakit jantung menjadi beberapa kelas (terkena penyakit jantung atau tidak) berdasarkan atribut yang ada?
2. Atribut apa saja yang berpengaruh dalam klasifikasi penyakit jantung ?
3. Bagaimana hasil akurasi, presisi, dan f-1 score dari pengelompokkan penyakit jantung?

1.3. Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Mengetahui kelas seseorang yang terkena penyakit jantung dan yang tidak terkena penyakit jantung.
2. Mengetahui atribut yang berpengaruh dalam klasifikasi dalam penyakit jantung
3. Menghasilkan sebuah system yang dapat mengklasifikasi tingkat keakuratan penyakit jantung.

1.4. Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Membantu dan mempermudah pihak rumah sakit atau pelaku kesehatan dalam

mengidentifikasi penyakit jantung.

2. Hasil penelitian ini dapat dijadikan penambahan pengetahuan dan referensi yang dapat dikembangkan, khususnya dalam bidang yang berkaitan dengan algoritma *Correlated Naïve Bayes*.

1.5. Batasan Masalah

Batasan-batasan masalah dalam penelitian ini adalah:

1. Metode yang digunakan adalah *Correlated Naïve Bayes*.
2. Data yang digunakan adalah [heart.csv](https://www.kaggle.com/search) yang didapat dari <https://www.kaggle.com/search> , berjumlah 918 record data.
3. Atribut yang digunakan yaitu *Age*, *Sex*, *ChestPain Type*, *RestingBP*, *Cholesterol*, *FastingBS*, *RestingECG*, *MaxHR*, *ExerciseAngina*, *Oldpeak*, *ST_Slope*, dan *HeartDisease* sebagai label.

1.6. Sistematika Penulisan

Langkah-langkah sistematika penulisan, sebagai berikut :

- a. Bab I Pendahuluan
Bab ini berisi penjelasan secara garis besar bagian dari penelitian yaitu latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan.
- b. Bab II Landasan Teori
Berisi penjelasan tentang teori-teori yang menjadi acuan dalam mendukung penulisan tentang klasifikasi penyakit jantung menggunakan algoritma *Correlated Naïve Bayes*.
- c. Bab III Metode Penelitian
Berisi penjelasan tentang data, tahap-tahap penelitian data, perhitungan algoritma *Correlated Naïve Bayes*, peralatan penelitian dan desain *user interface*.

BAB II

LANDASAN TEORI

2.1. Data Mining

2.1.1 Pengertian Data Mining

Data mining adalah suatu proses kegiatan menemukan pola yang menarik dari data dalam jumlah besar yang disimpan dalam penyimpanan database atau penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistic atau matematika.

2.1.2 Proses *Knowledge Discovery Database*

Pada data mining terdapat langka-langkah yang dikenal sebagai proses *Knowledge Discovery Database* (KDD). Adapun langkah-langkah proses KDD adalah sebagai berikut:

1. Pembersihan data (data cleaning)

Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan.

2. Integrasi data (data integration)

Integrasi data merupakan proses menggabungkan data dari berbagai database atau sumber data kedalam satu database baru.

3. Seleksi data (data selection)

Seleksi data merupakan proses pemilihan data pada database yang sering tidak dipakai oleh karena itu hanya data yang sesuai untuk di analisis yang akan diambil dari database.

4. Transformasi data (data transformation)

Transformasi data merupakan data diubah atau digabung kedalam format yang sesuai untuk di proses dalam data mining.

5. Proses mining (data mining)

Proses data mining merupakan proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi pola (pattern evaluation)

Evaluasi pola merupakan tahap untuk mengidentifikasi pola-pola menarik kedalam based yang ditemukan sehingga menghasilkan pengetahuan yang jelas dipahami.

7. Presentasi pengetahuan (knowledge presentation)

Presentasi pengetahuan merupakan proses mempresentasikan pengetahuan hasil dari penambangan data dengan menggunakan teknik visualisasi untuk membuat data dipahami oleh pengguna.

2.2. *Synthetic Minority Oversampling Technique (SMOTE)*

Synthetic Minority Oversampling Technique (SMOTE) merupakan teknik untuk mengatasi ketidakseimbangan pada kelas yang memiliki suatu *dataset* diperkenalkan oleh Chawla, Bowyer dan Hall (2002).

Metode SMOTE yang digunakan memungkinkan adanya *overfitting*. *Overfitting* dapat terjadi pada kelas minoritas yang diduplikasi sehingga adanya kemungkinan data latih yang sama. (Kasanah et al., 2019)

2.3. Naïve Bayes

Klasifikasi bayes sederhana yang lebih dikenal sebagai naïve Bayesian classifier dapat di asumsikan bahwa efek dari suatu nilai atribut sebuah kelas yang diberikan adalah bebas dari atribut-atribut lain. Asumsi ini disebut class conditional independence yang dibuat untuk memudahkan perhitungan-perhitungan, pengertian ini dianggap “naïve”, dalam bahasa lebih sederhana naïve itu mengasumsikan bahwa kemunculan suatu term kata dalam suatu kalimat tidak dipengaruhi kemungkinan kata-kata yang lain dalam kalimat padahal kenyataannya bahwa kemungkinan kata dalam kalimat sangat

dipengaruhi kemungkinan keberadaan kata-kata dalam kalimat. (Sulaksono & Darsono, 2015)

Secara umum, teorema bayes dapat ditulis dalam bentuk persamaan berikut :

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Keterangan :

X : data dengan class yang belum diketahui.

H : Hipotesis data merupakan suatu class spesifik.

P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X.

P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H.

P(H) : Probabilitas hipotesis H.

P(X) : Probabilitas X

Salah satu hal berpotensi untuk menambah nilai akurasi dari naïve bayes classifier adalah dengan nilai korelasi atribut terhadap kelas. Perhitungan korelasi nilai atribut terhadap kelas akan menjadi dasar ketepatan dari klasifikasi yang tidak hanya probabilitas namun juga seberapa besar korelasi atribut dengan kelas.

2.4. Correlated Naïve Bayes

Metode correlated naïve bayes classifier merupakan sebuah pengembangan dari metode naïve bayes. Pada metode correlated naïve bayes classifier memperhitungkan nilai korelasi (R-Square) antara variabel bebas (X) terhadap variabel terikat (Y). penambahan parameter korelasi digunakan untuk mengukur tinggi rendahnya derajat hubungan antara variabel bebas (X) terhadap variabel terikat (Y). (Marzuki et al., 2018)

Rumus algoritma correlated naïve bayes untuk klasifikasi:

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^Q P(X_i|Y)^{\ell \cdot R(X_i|Y)}}{P(X)}$$

Keterangan :

X : Data dengan kelas yang belum diketahui.

Y : Hipotesis data X merupakan suatu kelas spesifik.

$P(X|Y)$: Probabilitas hipotesis Y berdasarkan kondisi Y.

$P(Y)$: Probabilitas hipotesis Y (prior probability)

$P(X)$: Probabilitas dari X.

$R(X_i|Y)$: R-Square setiap atribut dari data X berdasarkan kondisi hipotesis Y.

$\prod_{i=1}^q (X_i|Y)$: Probabilitas setiap atribut dari data X berdasarkan kondisi hipotesis Y

$R(X_i|Y)$: R-Square setiap atribut dari data X berdasarkan kondisi hipotesis Y.

τ : Bilangan laplacian

Berikut merupakan persamaan untuk menentukan perhitungan korelasi atribut :

$$r = \frac{n \cdot (\sum XY) - (\sum X) \cdot (\sum Y)}{\sqrt{(n \cdot \sum X^2 - (\sum X)^2)} \sqrt{(n \cdot \sum Y^2 - (\sum Y)^2)}}$$
$$R = r^2$$

Keterangan :

R : R-Square fitur antar kelas.

r : Nilai Korelasi antar fitur kelas.

n : Total data pada dataset.

$\sum XY$: Total perkalian variabel X dengan variabel Y.

$\sum X$: Total variabel X.

$\sum Y$: Total variabel Y.

ΣX^2 : Total variabel X yang dikuadratkan.

ΣY^2 : Total variabel Y yang dikuadratkan.

$(\Sigma X)^2$: Kuadrat dari total variabel X.

$(\Sigma Y)^2$: Kuadrat dari total variabel Y.

Nilai (r) memiliki ketentuan dari nilai koefisien korelasi yakni $-1 \leq r \leq 1$.

Interprestasi koefisien korelasi nilai (r) ditunjukkan pada Tabel 2.1. berikut ini

:

Tabel 2.1. Tabel Koefisien Korelasi

Interval Koefisien	Tingkat Hubungan
0 – 0.199	Sangat Rendah
0.20 – 0.299	Rendah
0.40 – 0.599	Cukup
0.60 – 0.799	Kuat
0.80 – 1	Sangat Kuat

2.5. K-Fold Cross Validation

K-fold cross validation adalah teknik yang diadopsi untuk menggunakan semua dataset yang tersedia dan dibagi menjadi data testing dan data training.

Gambaran proses berjalannya 3-fold cross validation, dapat dilihat pada Tabel

2.2. berikut ini :

Tabel 2.2. Tabel Model 3-fold cross validation

1	2	3
1	2	3
1	2	3

Keterangan:



: data testing



: data training

2.6. Confusion Matrix

Confusion matrix adalah suatu metode yang biasa digunakan ketika melakukan perhitungan akurasi pada *data mining*. *Confusion matrix* diilustrasikan dengan tabel yang berisi jumlah data uji yang benar dan data uji yang salah diklasifikasikan. Akurasi dapat dilihat pada Tabel 2.3. berikut ini :

Tabel 2.3. *Confusion Matrix*

Actual	Diklasifikasikan	
	+	-
+	True positives	False negatives
-	False positives	True negatives

Keterangan:

True positives (TP) : jumlah data positif yang diklasifikasikan menjadi nilai positif

False positives (FP) : jumlah data positif yang diklasifikasikan menjadi nilai negatif

False negative (FN) : jumlah data negative yang diklasifikasikan menjadi nilai positif

True negative (TN) : jumlah data negative yang diklasifikasikan menjadi nilai negative

Matriks yang digunakan untuk mengevaluasi model adalah akurasi yang didefinisikan sebagai perbandingan jumlah data yang diprediksikan secara benar terhadap total jumlah data. *Presisi* didefinisikan untuk menggambarkan perbandingan *true positif* terhadap total data yang diprediksi positif. *Recall* didefinisikan sebagai perbandingan *true positif* terhadap total data *positif*. (HARIYANI et al., 2020)

Rumus akurasi:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN}$$

Rumus presisi:

$$\text{Presisi} = \frac{TP}{TP + FP}$$

Rumus *recall*:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Setelah hasil klasifikasi diukur kebenarannya, maka akan dilakukan perhitungan kombinasi nilai untuk dijadikan sebagai pengukuran *F1-score*. *F1-score* adalah rata-rata harmonik antara presisi, dan *recall*.

Rumus *F1-score*:

$$F1\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

BAB III

METODE PENELITIAN

3.1. Data

Data yang akan digunakan adalah data Penyakit Jantung yang diperoleh dari situs kaggle.com. Dari data penyakit jantung diperoleh 918 data record dan terdapat 11 atribut dan 1 kelas yang digunakan sebagai input dalam perhitungan metode klasifikasi Correlate Naïve Bayes. Atribut-atribut tersebut antara lain yang berhubungan dengan ciri-ciri terkena penyakit jantung seperti kolesterol, gula darah, dan nyeri dada. Atribut yang berhubungan dengan kondisi fisik seorang yang memiliki penyakit jantung diperhatikan berapa angka cek yang dilakukan setiap waktunya sesuai dengan atribut yang digunakan.

Penjelasan mengenai rincian atribut :

Atribut	Keterangan
Age	Usia
Sex	Gender/jenis kelamin
ChestPain Type	Jenis nyeri dada
RestingBP	Tekanan darah istirahat, diambil setelah duduk sekitar kurang dari 10 menit
Cholesterol	Produksi lemak oleh berbagai sel dalam tubuh
FastingBS	Mengukur gula darah setelah puasa selama 8 jam, sering disebut sebagai pemeriksaan gula darah puasa
RestingECG	Pemeriksaaan EKG yang dilakukan pada saat pasien dalam kondisi istirahat (dalam kondisi berbaring)
MaxHR	Detak jantung maximal
ExerciseAngina	Aktifitas olahraga akibat nyeri dada
Oldpeak	
ST_Slope	
HeartDisease	Penyakit jantung

Contoh data awal

No	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
3	37	M	ATA	130	283	1	ST	98	N	0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
6	39	M	NAP	120	339	1	Normal	170	N	0	Up	0
7	45	F	ATA	130	237	0	Normal	170	N	0	Up	0
8	54	M	ATA	110	208	0	Normal	142	N	0	Up	0
9	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
10	48	F	ATA	120	284	1	Normal	120	N	0	Up	0

3.2. *Synthethic Minority Oversampling Technique (SMOTE)*

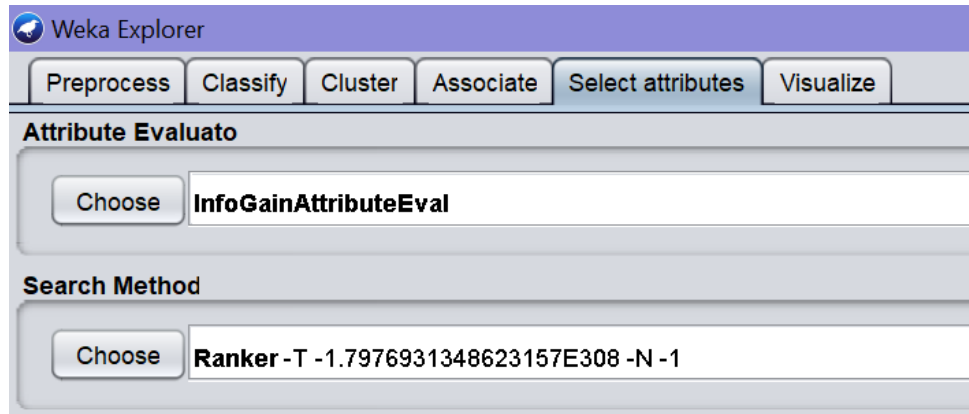
Penelitian ini tidak menggunakan *Synthethic Minority Oversampling Technique* (SMOTE) karena data yang digunakan dalam penelitian sudah *relative* seimbang dimana data yang digunakan bernilai 40% untuk ya dan 60% untuk tidak dimana data ya berjumlah 480 data dan tidak berjumlah 510 data. Hal tersebut membuat penelitian ini tidak menggunakan SMOTE karena data yang digunakan sudah *relative* seimbang.

3.3. Cleaning Data

Tahap data *cleaning* merupakan proses untuk membersihkan data yang tidak lengkap, mengandung *error*, dan tidak konsisten serta *missing value* atau kondisi adanya data yang hilang atau tidak lengkap. Data yang digunakan pada penelitian ini tidak memiliki *missing value*.

3.4. Data Selection

Tahap data selection merupakan proses pemilihan atribut yang relevan dengan ranking atau urutan bobot pada data menggunakan metode *information gain*. Untuk memilih data atau atribut yang relevan, maka proses seleksi atribut akan menggunakan *Waikoto Environment for Knowledge Analysis* (WEKA) tools versi 3.9.4. Data yang melewati tahap data selection berjumlah 918 dengan 11 atribut dan 1 label, atribut tersebut merupakan atribut dari ciri-ciri terkena penyakit jantung seperti umur, jenis kelamin, tipe nyeri dada, tekanan darah, dan kolesterol. Nama-nama atribut meliputi age, sex, chest pain type, resting bp, cholesterol, fasting bs, resting ecg, max hr, exercise angina, old peak, st slope, *heart disease* sebagai label.



Pemilihan atribut dengan menggunakan weka menghasilkan ranked yang dimana hasil dari ranked pada penelitian ini yang diambil adalah yang nilai ranked nya lebih dari 0.01 sedangkan nilai yang kurang dari 0.01 tidak digunakan. Hasil dari pemeringkatan atribut, berikut ini :

No	Ranked	Attributes
1	0.29932	ST_Slope
2	0.22504	ChestPainType
3	0.18997	ExerciseAngina
4	0.16143	Oldpeak
5	0.12748	MaxHR
6	0.08337	Cholesterol
7	0.0696	Age
8	0.06849	Sex
9	0.05488	FastingBS
10	0.01552	RestingBP
11	0.00872	RestingECG

3.5. Data Transformation

Tahap ini akan dilakuakn perubahan tipe data pada data yang melekat pada atribut untuk mempermudah proses penambangan data. Berikut merupakan proses dari *transformasi* data yang berjalan:

3.5.1 Transformasi atribut Sex

Proses ini melakukan transformasi data *sex*, data yang berkaitan dengan jenis kelamin akan dikelompokkan berdasarkan ketentuan sebagai berikut:

M (Male) : 1

F (Female) : 2

3.5.2 Transformasi atribut Chest Pain Type

Proses ini melakukan transformasi data *Chest Pain Type*, data yang berkaitan dengan tipe nyeri dada akan dikelompokkan berdasarkan ketentuan sebagai berikut:

ATA : 1

NAP : 2

ASY : 3

3.5.3 Transformasi atribut Exercise Angina

Proses ini melakukan transformasi data *exercise angina*, data akan dikelompokkan berdasarkan ketentuan sebagai berikut:

N : 2

Y : 1

3.5.4 Transformasi atribut ST_Slope

Proses ini melakukan transformasi dari *st_slope*, data akan dikelompokkan berdasarkan ketentuan sebagai berikut:

UP : 1

FLAT : 2

3.5.5 Normalisasi min-max attribute Age

Proses ini melakukan normalisasi min-max data age, data tersebut terdiri dari usia termuda 36 dan usia tertua 60. Nilai tersebut akan dihitung menggunakan normalisasi min-max berdasarkan rumus sebagai berikut

Age = 40

$$x_i^l = \frac{x_i - \min_A}{\max_A - \min_A} (\max_{baru_A} - \min_{baru_A}) + \min_{baru_A}$$

$$x_i^l = \frac{40 - 37}{54 - 37} (1) + 0$$

$$x_i^l = \frac{3}{17} (1) + 0$$

$$x_i^l = 0,176470588$$

3.5.6 Normalisasi min-max attribute Resting BP

Proses ini melakukan normalisasi min-max data resting bp, data tersebut terdiri dari data terkecil 100 dan data tertinggi 160. Nilai tersebut akan dihitung menggunakan normalisasi min-max berdasarkan rumus sebagai berikut:

$$\text{Resting BP} = 140$$

$$x_i^l = \frac{x_i - \min_A}{\max_A - \min_A} (\max_{baru_A} - \min_{baru_A}) + \min_{baru_A}$$

$$x_i^l = \frac{140 - 110}{160 - 110} (1) + 0$$

$$x_i^l = \frac{30}{50} (1) + 0$$

$$x_i^l = 0,6$$

3.5.7 Normalisasi min-max attribute Cholesterol

Proses ini melakukan normalisasi min-max data kolesterol, data tersebut terdiri dari data terkecil 164 dan data tertinggi 339. Nilai tersebut akan dihitung menggunakan normalisasi min-max berdasarkan rumus sebagai berikut:

$$\text{Cholesterol} = 289$$

$$x_i^l = \frac{x_i - \min_A}{\max_A - \min_A} (\max_{baru_A} - \min_{baru_A}) + \min_{baru_A}$$

$$x_i^l = \frac{289 - 180}{339 - 180} (1) + 0$$

$$x_i^l = \frac{109}{159} (1) + 0$$

$$x_i^l = 0,685534591$$

3.5.8 Normalisasi min-max attribute Max HR

Proses ini melakukan normalisasi min-max data kolesterol, data tersebut terdiri dari data terkecil 98 dan data tertinggi 172. Nilai tersebut akan dihitung menggunakan normalisasi min-max berdasarkan rumus sebagai berikut:

$$\text{Max HR} = 172$$

$$x_i^l = \frac{x_i - \min_A}{\max_A - \min_A} (\max_{baru_A} - \min_{baru_A}) + \min_{baru_A}$$

$$x_i^l = \frac{172 - 98}{172 - 98} (1) + 0$$

$$x_i^l = \frac{74}{74} (1) + 0$$

$$x_i^l = 1$$

Transformasi data yang dilakukan meliputi perubahan data kategorial menjadi numeric dan menggunakan fungsi min-max untuk memperkecil rentang data antar atribut. Hasil transformasi yang dilakukan pada data tabel:

Data hasil transformasi

No	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	0,176470588	1	1	0,6	0,685534591	0	1	2	0	1	2
2	0,705882353	2	2	1	0	0	0,783783784	2	1	2	1
3	0	1	1	0,4	0,647798742	1	0	2	0	1	2
4	0,647058824	2	3	0,56	0,213836478	0	0,135135135	1	1.5	2	1
5	1	1	2	0,8	0,094339623	0	0,324324324	2	0	1	2
6	0,117647059	1	2	0,2	1	1	0,972972973	2	0	1	2
7	0,470588235	2	1	0,4	0,358490566	0	0,972972973	2	0	1	1
8	1	1	1	0	0,176100629	0	0,594594595	2	0	1	2
9	0	1	3	0,6	0,169811321	0	0,432432432	1	1.5	2	1
10	0,647058824	2	1	0,2	0,65408805	1	0,297297297	2	0	1	2

3.6. Modelling Correlated Naïve Bayes

Tahap pembentukan model, data telah melalui tahap *preprocessing* akan dibentuk modelnya terlebih dengan menggunakan algoritma *Correlated Naïve Bayes*. Sebelum masuk dalam tahap perhitungan, data akan dibagi menjadi beberapa bagian menggunakan *K-fold cross validation*. Data yang digunakan dalam perhitungan menggunakan data hasil transformasi yang akan dibagi menjadi data *training* dan data *testing*. Untuk permodelan metode Correlated Naïve Bayes akan digunakan data pada atribut *maxHR* Dengan data testing yang digunakan adalah data pada baris pertama. Perhitungan dimulai dengan mencari nilai korelasi (R-Square).

MaxHR (X)	HeartDisease (Y)	X^2	Y^2	ΣXY
1	2	1	4	2
0,783783784	1	0,61431702	1	0,783783784
0	2	0	4	0
0,135135135	1	0,018261505	1	0,135135135
0,324324324	2	0,105186267	4	0,648648648
0,972972973	2	0,946676406	4	1,945945946
0,972972973	1	0,946676406	1	0,972972973
0,594594595	2	0,353542732	4	1,18918919
0,432432432	1	0,186997808	1	0,432432432
0,297297297	2	0,088385683	4	0,594594594
(ΣX) 5,513513513	(ΣY) 16	(ΣX^2) 4,260043828	(ΣY^2) 28	(ΣXY) 8,702702702

3.7. Spesifikasi Alat Penelitian

Peralatan yang digunakan untuk menyelesaikan penelitian ini adalah sebagai berikut:

a. Perangkat Keras

Perangkat keras yang digunakan dalam penelitian ini adalah:

1. Merk : Asus

2. Processor : Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz
3. RAM : 8,00 GB
4. HDD : 1TB

b. Perangkat Lunak

Perangkat lunak yang digunakan dalam penelitian ini adalah:

1. Windows 10
2. Weka Tools versi 3.9.4
3. Draw io

3.8. Desain Interface

KLASIFIKASI PENYAKIT JANTUNG MENGGUNAKAN ALGORITMA CORRELATED NAIVE BAYES

Telusuri

Data Penyakit Jantung

Jumlah K-Fold

Akurasi

Uji Data Tunggal

Age :

Sex :

ChestPain Type : ☐ ATA ☐ NAP ☐ ASY

RestingBP :

Cholesterol :

FastingBS : ☐ 0 ☐ 1

MaxHR :

ExerciseAngina : ☐ 1 ☐ 2

Oldpeak :

ST_Slope : ☐ 1 ☐ 2

Keterangan Atribut

Age : Usia

Sex : Jenis kelamin
--> Male
--> Female

ChestPain Type : Jenis nyeri dada
--> ATA
--> NAP
--> ASY

RestingBP : Tekanan darah istirahat

Cholesterol : Produksi lemak oleh berbagai sel dalam tubuh

FastingBS : Mengukur gula darah setelah puasa

MaxHR : Detak jantung maksimal

ExerciseAngina : Aktivitas olahraga akibat nyeri dada
--> Y
--> N

Oldpeak :
--> UP
--> FLAT

Daftar Pustaka

- Bianto, M. A., Kusriani, K., & Sudarmawan, S. (2020). Perancangan Sistem Klasifikasi Penyakit Jantung Menggunakan Naïve Bayes. *Creative Information Technology Journal*, 6(1), 75.
<https://doi.org/10.24076/citec.2019v6i1.231>
- Chawla, B. dan H. (2002). (2002). SMOTE: Synthetic Minority Over-sampling Technique Nitesh. *Ecological Applications*, 30(2), 321–357.
- HARIYANI, Y. S., HADIYOSO, S., & SIADARI, T. S. (2020). Deteksi Penyakit Covid-19 Berdasarkan Citra X-Ray Menggunakan Deep Residual Network. *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, 8(2), 443. <https://doi.org/10.26760/elkomika.v8i2.443>
- Kasanah, A. N., Muladi, M., & Pujiyanto, U. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 3(2), 196–201. <https://doi.org/10.29207/resti.v3i2.945>
- Marzuki, J. I., Mataram, K., & Bar, N. T. (2018). KOMPARASI AKURASI METODE CORRELATED NAIVE BAYES CLASSIFIER DAN NAIVE BAYES CLASSIFIER UNTUK DIAGNOSIS PENYAKIT DIABETES Hairani , Gibran Satya Nugraha , Mokhammad Nurkholis Abdillah , Muhammad Innuddin InfoTekJar (Jurnal Nasional Informatika dan Teknolog. *InfoTekJar (Jurnal Nasional Informatika Dan Teknologi Jaringan)*, 3(1), 6–11.
- Putra, P. D., & Rini, D. P. (2019). Prediksi Penyakit Jantung dengan Algoritma Klasifikasi. *Prosiding Annual Research Seminar 2019*, 5(1), 978–979.
- Sulaksono, J., & Darsono. (2015). Sistem pakar penentuan penyakit gagal jantung menggunakan metode naive bayes classifier. *Seminar Nasional Teknologi Informasi Dan Multimedia 2015*, 6–8.