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

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR RUMUS

DAFTAR TABEL

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 jatung 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 disebebkan oleh infeksi atau kelainan dari lahir. Penyakit jantung merupakan sebuah kondisi dimana jantung tidak dapat melaksanakan tugasnya dengan baik. Data Organisasi Kesehatan Dunia *World Health Organization* (WHO) menyebutkan, lebih dari 17 juta orang di dunia meninggal akibat penyakit jantung dan pembuluh darah.

Penyakit jantung sering dikenal sebagai *suddent death*. Tingginya factor kematian akibat penyakit jantung dapat di cegah dan ditekan factor resiko 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 di fungsikan 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 atau 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 dkk., (2020), membuktikan melalui hasil penelitian telah dijelaskan dan dilakukan pada pembuatan sistem klasifikasi penyakit jantung menggunakan *Naïve Bayes*. Pembuatan system 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 didapatlah 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 Naïve Bayes*. Implementasi metode ini juga akan menghitung tingkat akurasi, *presisi*, *recall* dan *f1-score* dari algoritma *Correlated Naïve Bayes* dalam melakukan klasifikasi penyakit jantung seseorang.

1.2. Rumusan Masalah

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

1. Atribut apa saja yang berpengaruh dalam klasifikasi penyakit jantung?
2. 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 atribut yang berpengaruh dalam klasifikasi dalam penyakit jantung.
2. 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

Data yang digunakan dalam penelitian ini adalah *heart.csv* sebuah data public yang diperoleh dari https://www.kaggle.com/search berjumlah 918 *record* data terdiri dari 11 atribut dan 1 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, manfat 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

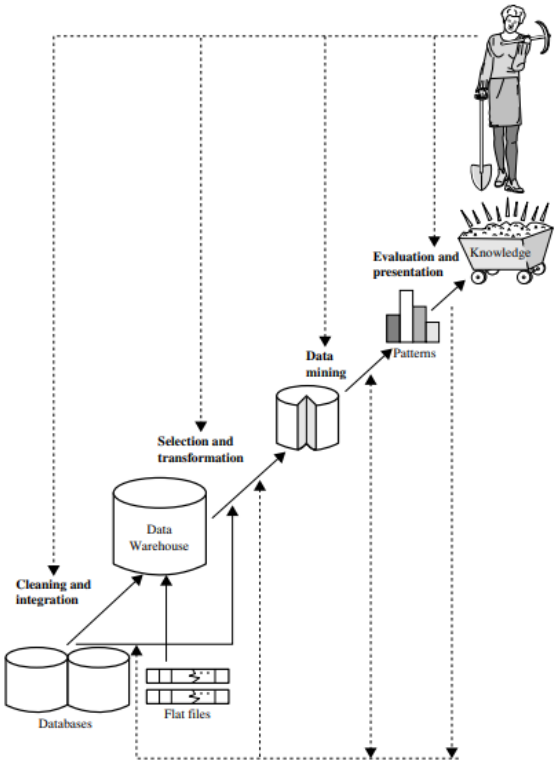
2.1. Data Mining

2.1.1. Pengertian Data Mining

Data mining merupakan pengolahan data dalam jumlah besar menjadi sebuah informasi atau pengetahuan suatu teknik. Proses pengolahan data dalam data mining membutuhkan algoritma-algoritma untuk melakukan ekstraksi menjadi sebuah informasi atau pola atau pengetahuan. Penggunaan dalam algoritma pada data mining diklasifikasikan berdasarkan masing-masing peranan data mining. (Suntoro, 2019).

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 seperti pada Gambar 2.1.:



Gambar 2.1. Proses *Knowledge Discovery Database* (Han et al., 2012)

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)

   Metode transformasi yang akan digunakan dalam penelitian ini adalah metode *min-max*. Normalisasi min-max merupakan perubahan ukuran pada data dari rentang asli menjadi sebuah nilai numerik dalam kisaran 0 dan 1. (Ambarwari dkk., 2017).

   Berikut ini akan digunakan untuk menghitung normalisasi dengan metode *min-max*: (Nurjanah dkk., 2017).

$$x_i^l = \frac{x_i - min_A}{max_A - min_A} (maxbaru_A - minbaru_A) + minbaru_A \qquad (2.1)$$

Keterangan:

x_i^l	: Nilai data baru hasil normalisasi *min-max*.
x_i	: Nilai data yang akan dinormalisasi.
min_A	: Nilai minimum data.
max_A	: Nilai maximum data.
$maxbaru_A$	: Nilai maximum dalam rentang.
$minbaru_A$	: Nilai minimum dalam rentang.

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. 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 ditulisakn dalam bentuk persamaan berikut :

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2.2)$$

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.

2.3. 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). Rumus algoritma correlated naïve bayes untuk klasifikasi:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{Q} P\ (X_I|Y)^{\ell}.R(X_i|Y)}{P(X)} \quad (2.3)$$

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 :

$$r = \frac{n.(\Sigma XY)-(\Sigma X).(\Sigma Y)}{\sqrt{(n.\Sigma X^2-(\Sigma X)^2}\sqrt{(n.\Sigma Y^2-(\Sigma Y)^2}} \quad (2.4)$$

$$R = r^2 \quad (2.5)$$

Keterangan :

R : R-Square fitur antar kelas.

r : Nilai Korelasi antar fitur kelas.

n : Total data pada dataset.

ΣXY: Total perkalian variabel X dengan variabel Y.

ΣX : Total variabel X.

ΣY : 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 <= r <= 1. Interprestasi koefisien korelasi nilai (r) ditunjukkan pada Tabel 2.1. dan Tabel 2.2. berikut:

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

Tabel 2.2. Tabel R-Square Korelasi

R-Square	Tingkat Hubungan
0 – 0.039601	Sangat Rendah

0.04 – 0,089401	Rendah
0.16 – 0.358801	Cukup
0.36 – 0.6384101	Kuat
0.64 – 1	Sangat Kuat

2.4. K-Fold Cross Validation

K-fold cross validation merupakan sebuah teknik pendekatan alternatif untuk melatih dan menguji data. Sebuah *dataset* terdiri dari N buah data akan dibagi menjadi k bagian yang sama. (Bramer, 2016). Gambaran proses berjalannya 3-fold cross validation, dapat dilihat pada Tabel 2.3. berikut:

Tabel 2.3. Tabel Model 3-*fold cross validation*

1	2	3
1	2	3
1	2	3

Keterangan:

: data testing

: data training

2.5. 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. (Bramer, 2016). Akurasi dapat dilihat pada Tabel 2.3. berikut:

Tabel 2.4. *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 dkk., 2020).

Rumus akurasi:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.6)$$

Rumus presisi:

$$\text{Presisi} = \frac{TP}{TP + FP} \quad (2.7)$$

Rumus *recall*:

$$\textit{Recall} = \frac{TP}{TP + FN} \quad (2.8)$$

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

Rumus *F1-score*:

$$F1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (2.9)$$

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 label yang digunakan sebagai input dalam perhitungan metode klasifikasi *Correlate Naïve Bayes*. Penjelasan masing-masing atribut yang digunakan dalam penelitian ini dapat dilihat pada Tabel 3.1. berikut:

Tabel 3.1. Penjelasan Atribut

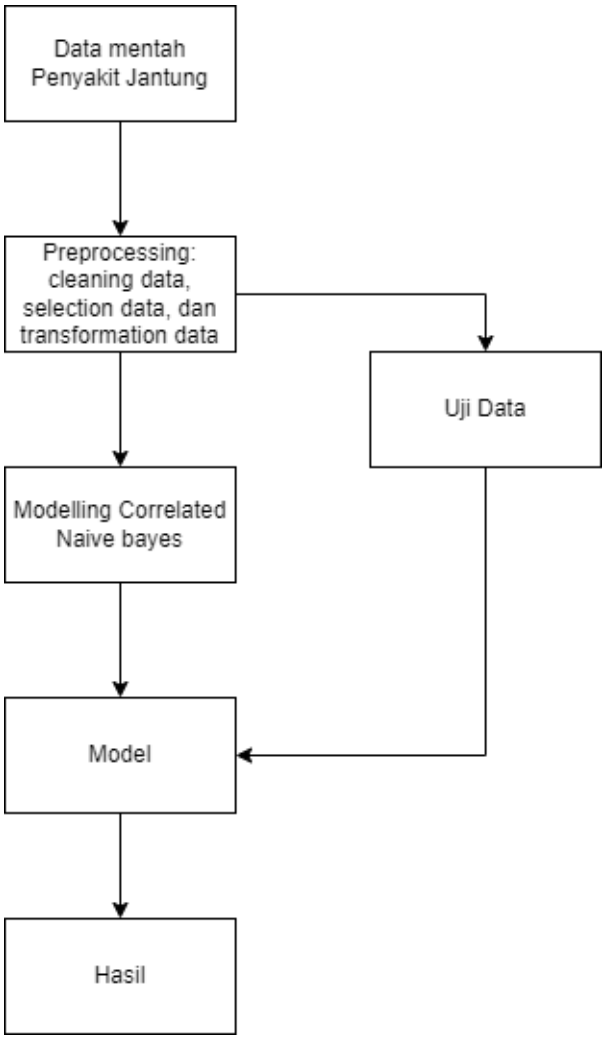
Atribut	Keterangan
Age	Usia
Sex	Gender/jenis kelamin
ChestPain Type	Jenis nyeri dada
RestingBP	Tekanan darah istirahat, diambil setelah duduk sekitar kurang dari 10 menit
Cholesterol	Produksi lemak oleh berbagai sel dalam tubuh
FastingBS	Mengukur gula darah setelah puasa selama 8 jam, sering disebut 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	Penurunan ST akibat olahraga
ST_Slope	Slope dari puncak ST setelah berolahraga
HeartDisease	Penyakit jantung

Tabel 3.2. Contoh Data Awal dari dataset Penyakit Jantung

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

3.2. Desain Alat Uji

Alur atau tahapan proses dalam penelitian ini dapat dilihat pada Gambar 3.1. berikut:



Gambar 3.1. Desai Alat Uji

3.3. Preprocessing

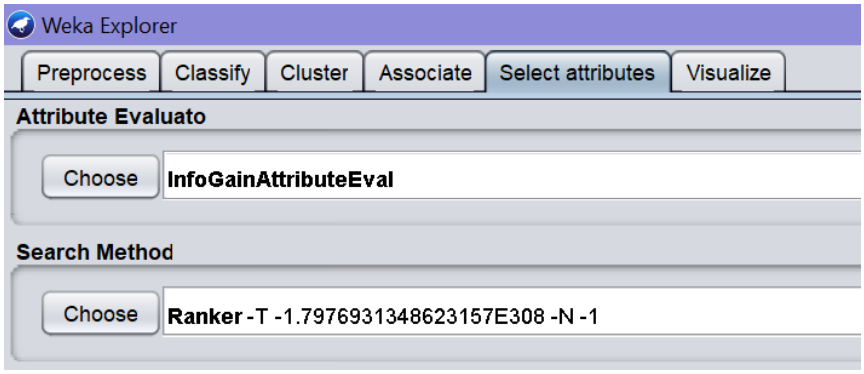
Data awal akan memulai tahap *processing* terlebih dahulu untuk menghilangkan data-data yang tidak lengkap, mengadung *error* dan tidak konsisten sehingga sistem akan menghasilkan *dataset* yang siap digunakan untuk proses selanjutnya. Tahap *processing* yang dilakukan melalui beberapa tahap, yaitu:

3.3.1. Pembersihan Data

Tahap pembersihan data dilakukan dengan mencari *missing value* dan mencari data *outlier*. Pencarian *missing value* dalam penelitian ini dilakukan dengan menggunakan *Microsoft Excel*, setiap data pada masing-masing atribut dilakukan pengurutan dari data terkecil hingga data terbesar. Jika pada urutan data ditemukan nilai kosong, maka data tersebut akan diganti. Hasil pencarian pada data ini tidak ditemukan adanya *missing value* pada *dataset.* Dalam mencari data *outlier*, digunakan pula bantuan *Microsoft Excel* untuk melakukan perhitungan data. Pencarian ini menemukan data yang memiliki *outlier*, sehingga data tersebut perlu dihapus untuk menghindari kesalahan pada saat penelitian. Jumlah data sebelum ditemukannya *outlier* adalah 918 record data, setelah ditemukan *outlier* dan data tersebut dihapus maka data yang dapat dinyatakan sudah bersih adalah 897 data.

3.3.2. Seleksi Data

Tahap seleksi data merupakan proses pemilihan atribut yang relevan dengan rangking 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. dapat dilihat pada Gambar 3.2. berikut:



Gambar 3.2. Perangkingan Atribut dengan Information Gain

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 dapat dilihat pada Tabel 3.3 berikut ini:

Tabel 3.3. Perangkingan Atribut menggunakan aplikasi Weka

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

3.3.3. Transformasi Data

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

1. Transformasi Atribut Sex

   Proses ini melakukan transformasi data *sex* kedalam bentuk numerik.

   M (Male) : 1

   F (Female) : 2

2. Transformasi Atribut ChestPain Type

Proses ini melakukan transformasi data *Chest Pain Type* kedalam bentuk numerik.

ATA : 1

NAP : 2

ASY : 3

3. Transformasi Atribut ExerciseAngina

Proses ini melakukan transformasi data *exercise angina* kedalam bentuk numerik.

Y : 1

N : 2

4. Transformasi Atribut ST_Slope

Proses ini melakukan transformasi dari *st_slope* kedalam bentuk numerik.

UP : 1

FLAT : 2

5. Normalisasi Min-max

Transformasi semua atribut menggunakan normalisasi *min-max*. Data tertinggi sebagai nilai max dan data terendah sebagai nilai min. Normalisasi yang dilakukan pada setiap data kolom pertama. Rumus yang digunakan adalah rumus *min-max* (2.1). Proses perhitungan dapat dilihat sebagai berikut:

Age = 40

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

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

$$x_i^l = 0{,}176470588$$

Resting BP = 140

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

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

$$x_i^l = 0{,}6$$

Cholesterol = 289

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

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

$$x_i^l = 0{,}685534591$$

Max HR = 172

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

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

$$x_i^l = 1$$

Transformasi data yang dilakukan meliputi perubahan data kategorial menjadi numerik dan menggunakan fungsi min-max untuk memperkecil rentang data antar atribut. Hasil transformasi yang dilakukan pada data Tabel 3.4. berikut:

Tabel 3.4. Data Hasil Transformas

No	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	0,176471	0	1	0,2	0,685535	1	1	0	0,666667	0	2
2	0,705882	1	0,5	1	0	0	0,783784	1	0	1	1
3	0	1	0	0,4	0,647799	0	0	1	0	1	2
4	0,647059	0	0	0,56	0,213836	0	0,135135	1	1	0	1
5	1	1	0,5	0,8	0,09434	0	0,324324	1	0	0	2
6	0,117647	1	0,5	0,6	1	1	0,972973	1	0	0	2
7	0,470588	0	0	0,4	0,358491	0	0,972973	1	0	0	1
8	1	1	0	0	0,176101	0	0,594595	1	0	0	2
9	0	1	1	0,2	0,169811	1	0,432432	0	1	0	1
10	0,647059	0	0	0,6	0,654088	0	0,297297	1	0	1	2

3.4. 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). Dapat dilihat pada Tabel 3.5. berikut:

Tabel 3.5. Data *Atribut Product MaxHR*

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

$$r = \frac{n.(\Sigma XY) - (\Sigma X).(\Sigma Y)}{\sqrt{(n.\Sigma X^2 - (\Sigma X)^2}\sqrt{(n.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$r = \frac{10.(8{,}702702702) - (5{,}513513513).(16)}{\sqrt{(10.4{,}260043828 - (16)^2}\sqrt{(10.28 - (16)^2}}$$

$$r = \frac{87{,}02702702 - 88{,}21621621}{\sqrt{42{,}6004 - 30{,}3988}\sqrt{280 - 256}}$$

$$r = \frac{-1{,}189189188}{\sqrt{12{,}2016}\sqrt{24}}$$

$$r = \frac{-1{,}189189188}{3{,}49308\ .\ 4{,}89898}$$

$$r = \frac{-1{,}189189188}{17{,}11252665}$$

$$r\ (MaxHR) = -0{,}069492321$$

$$R = r^2$$

$$R = -0{,}069492321^2$$

$$R\ (MaxHR) = 0{,}004829183$$

Nilai korelasi *R-Square* untuk atribut *MaxHR* telah diperoleh, untuk atribut lain dapat dilihat pada tabel 3.6 berikut:

Tabel 3.6. Hasil Perhitungan *R-Square*

Atribut	R
Age	0,002121763
Sex	0,340277778
ChestPain Type	0,330601093
RestingBP	0,221958144
Cholesterol	0,319272606
FastingBS	0,285714286
RestingECG	0,074074074
MaxHR	0,004829183
ExerciseAngina	0,010416667
Oldpeak	0,615384615
ST_Slope	0,007936508

Tahap selanjutnya akan dilakukan perhitungan terhadap nilai *prior probability*, probabilitas independen kelas Y dari semua fitur dalam vektor X, menghitung perkalian dan diakhiri dengan menghitung nilai *maximum* dari probabilitas akhir yang diperoleh. Perhitungan nilai *prior probability* dapat dilihat pada Tabel 3.7. berikut:

Tabel 3.7. Perhitungan *Prior Probability*

Prior Probability (Kelas)	Jumlah kelas	N	P (X|Kelas) (Kelas / N)
YES	4	10	0,4
NO	6	10	0,6

Tahap ini akan dilakukan perhitungan terhadap nilai probabilitas independen Y dari semua fitur kelas X dengan acuan data yang terdapat pada Tabel 3.4. dengan baris pertama dijadikan sebagai data testing.

P (Yes | Age = 0,176470588) = $\frac{0+1}{4}$ = 0,25

P (No | Age = 0,176471) = $\frac{1+1}{6}$ = 0,333333333

P (Yes | Sex = 0) = $\frac{2+1}{4}$ = 0,75

P (No | Sex = 0) = $\frac{2+1}{6}$ = 0,5

P (Yes | ChestPain Type = 1) = $\frac{1+1}{4}$ = 0,5

P (No | ChestPain Type = 1) = $\frac{1+1}{6}$ = 0,333333333

P (Yes | RestingBP = 0,2) = $\frac{1+1}{4}$ = 0,5

P (No| RestingBP = 0,2) = $\frac{1+1}{6}$ = 0,333333333

P (Yes | Cholesterol = 0,68553459) = $\frac{0+1}{4}$ = 0,25

P (No | Cholesterol = 0,68553459) = $\frac{1+1}{6}$ = 0,333333333

P (Yes | FastingBS = 1) = $\frac{2+1}{4}$ = 0,75

P (No | FastingBS = 1) = $\frac{1+1}{6}$ = 0,333333333

P (Yes | MaxHR = 1) = $\frac{0+1}{4}$ = 0,25

P (No | MaxHR = 1) = $\frac{1+1}{6}$ = 0,333333333

P (Yes | ExerciseAngina = 0) = $\frac{1+1}{4}$ = 0,5

P (No | ExerciseAngina = 0) = $\frac{1+1}{6}$ = 0,333333333

P (Yes | Oldpeak = 0,666666667) = $\frac{0+1}{4}$ = 0,25

P (No | Oldpeak = 0,666666667) = $\frac{1+1}{6}$ = 0,333333333

P (Yes | ST_Slope = 0) = $\frac{3+1}{4}$ = 1

P (No | ST_Slope = 0) = $\frac{0+1}{6}$ = 0,833333333

Tahap ini akan dilakukan perkalian antara hasil perhitungan probabilitas independen kelas Y dari semua fitur kelas X akan dengan hasil perhitungan R-Square, perhitungan akan dilakukan untuk semua tribute. Hasil dari perhitungan tersebut akan dikalikan dengan perhitungan prior probability untuk masing-masing kelas.

P (Yes | X) = ((P (Yes | Age = 0,176470588) * R (Yes | Age = 0,176470588) + P (Yes | Sex = 0) * R (Yes | Sex = 0) + P (Yes | ChestPain Type = 1) * R (Yes | ChestPain Type = 1) + P (Yes | RestingBP = 0,2) * R (Yes | RestingBP = 0,2) + P (Yes | Cholesterol = 0,68553459) * R (Yes | Cholesterol = 0,68553459)) + P (Yes | FastingBS = 1) * R (Yes | FastingBS = 1) + P (Yes | MaxHR = 1) * R (Yes | MaxHR = 1) + P (Yes | Exercise Angina = 0) * R (Yes | Exercise Angina = 0) + P (Yes | Oldpeak = 0,666666667) * R (Yes | Oldpeak = 0,666666667) + P (Yes | ST_Slope = 0) * R (Yes | ST_Slope = 0)) * P (X | Yes) = ((0,25 * 0,002121763) + (0,75 * 0,340277778) + (0,5 * 0,330601093) + (0,5 * 0,221958144) + (0,25 * 0,319272606) + (0,75 * 0,285714286) + (0,25 * 0,004829183) + (0,5 * 0,010416667) + (0,25 * 0,615384615) + (1 * 0,007936508)) * 0,4 = **0,39772822**

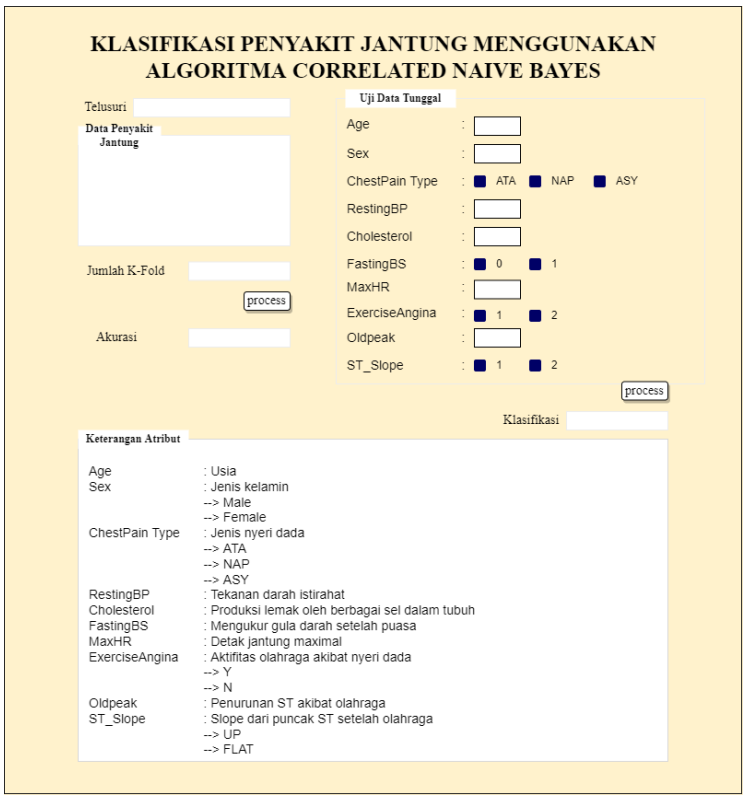
P (No | X) = ((P (No | Age = 0,176470588) * R (No | Age = 0,176470588) + P (No | Sex = 0) * R (No | Sex = 0) + P (No | ChestPain Type = 1) * R (No |

ChestPain Type = 1) + P (No | RestingBP = 0,2) * R (No | RestingBP = 0,2) + P (No | Cholesterol = 0,68553459) * R (No | Cholesterol = 0,68553459)) + P (No | FastingBS = 1) * R (No | FastingBS = 1) + P (No | MaxHR = 1) * R (No | MaxHR = 1) + P (No | Exercise Angina = 0) * R (No | Exercise Angina = 0) + P (No | Oldpeak = 0,666666667) * R (No | Oldpeak = 0,666666667) + P (No | ST_Slope = 0) * R (No | ST_Slope = 0)) * P (X | No) = ((0,33333333 * 0,002121763) + (0,5 * 0,340277778) + (0,33333333 * 0,330601093) + (0,33333333 * 0,221958144) + (0,33333333 * 0,319272606) + (0,33333333 * 0,285714286) + (0,33333333 + 0,004829183) + (0,33333333 * 0,010416667) + (0,33333333 * 0,615384615) + (0,83333333 * 0,007936508)) * 0,6 = **0,464111255**

P (Yes | X) = **0,39772822**
P (No | X) = **0,464111255**

Dari hasil yang didapatkan, maka perhitungan nilai *maximum* dari probabilitas akhir kelas *No* lebih besar daripada kelas *Yes*, sehingga data diklasifikasikan kedalam kelas *No*.

3.5. Desain *User Interface*



Keterangan *User Interface*:

1. *Button* telusuri berfungsi untuk memilih file dari computer.
2. *Field* Data Penyakit Jantung digunakan untuk menampilkan data siap pakai yang dipilih oleh *user*.
3. *Field* jumlah *k-fold* untuk menginputkan hasil dari jumlah *k-fold*
4. *Button process* berfungsi untuk memproses data dari *k-fold cross validation*, lalu
5. Button akurasi akan menampilkan hasil akurasi.
6. *Field* keterangan atribut digunakan untuk menampilkan keterangan dari atribut yang digunakan.
7. Tahap uji data tunggal dilakukan dengan memasukkan data yang akan diklasifikasi, kemudian *button* proses akan memproses data uji tunggal sehingga pada *field* akan menampilkan hasil klasifikasi penyakit jantung seseorang.

3.6. 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

DAFTAR PUSTAKA

Ambarwari, A., Adrian, Q. J., & Herdiyeni, Y. (2017). Terakreditasi SINTA Peringkat 2 Analisis Pengaruh Data Scaling Terhadap Performa Algoritme Machine Learning untuk Identifikasi Tanaman. *Masa Berlaku Mulai*, *1*(3), 117–122.

Bianto, M. A., Kusrini, K., & Sudarmawan, S. (2020). Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes. *Creative Information Technology Journal*, *6*(1), 75. https://doi.org/10.24076/citec.2019v6i1.231

Bramer, M. (2016). *Principles of Data Mining*.

Chawla, B. dan H. (2002). (2002). SMOTE: Synthetic Minority Over-sampling Technique Nitesh. *Ecological Applications*, *30*(2), 321–357.

Han, J., Kamber, M., & Pei, J. (2012). A study of data mining concepts and techniques. In *International Journal of Applied Engineering Research* (Vol. 9, Issue 27 Special Issue).

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., & Pujianto, 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

Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *1*(12), 1750–1757.

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.

Suntoro, J. (2019). 22-DATA MINING Algoritma dan Implementasi Menggunakan Bahasa Pemrograman PHP. *DATA MINING Algoritma Dan Implementasi Menggunakan Bahasa Pemrograman PHP*, *9*(9), 259–278.