**PROPOSAL TUGAS AKHIR**

**KLASIFIKASI POTENSI 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 ISI** ii](#_Toc106561288)

[**DAFTAR GAMBAR** iv](#_Toc106561289)

[**DAFTAR RUMUS** v](#_Toc106561290)

[**DAFTAR TABEL** vi](#_Toc106561291)

[**BAB I** 1](#_Toc106561292)

[**PENDAHULUAN** 1](#_Toc106561293)

[**1.1.** **Latar Belakang** 1](#_Toc106561294)

[**1.2.** **Rumusan Masalah** 3](#_Toc106561297)

[**1.3.** **Tujuan Penelitian** 3](#_Toc106561301)

[**1.4.** **Manfaat Penelitian** 3](#_Toc106561306)

[**1.5.** **Batasan Masalah** 3](#_Toc106561307)

[**1.6.** **Sistematika Penulisan** 3](#_Toc106561308)

[**BAB II** 5](#_Toc106561309)

[**LANDASAN TEORI** 5](#_Toc106561310)

[**2.1.** **Data Mining** 5](#_Toc106561313)

[**2.1.1.** **Pengertian Data Mining** 5](#_Toc106561317)

[**2.1.2.** **Proses Knowledge Discovery Database** 5](#_Toc106561322)

[**2.2.** **Naïve Bayes** 7](#_Toc106561323)

[**2.3.** **Correlated Naïve Bayes** 8](#_Toc106561328)

[**2.4.** **K-Fold Cross Validation** 10](#_Toc106561341)

[**2.5.** **Confusion Matrix** 10](#_Toc106561349)

[**BAB III** 13](#_Toc106561350)

[**METODE PENELITIAN** 13](#_Toc106561351)

[**3.1.** **Data** 13](#_Toc106561355)

[**3.2.** **Desain Alat Uji** 15](#_Toc106561360)

[**3.3.** **Preprocessing** 15](#_Toc106561366)

[**3.3.1.** **Pembersihan Data** 16](#_Toc106561373)

[**3.3.2.** **Seleksi Data** 16](#_Toc106561381)

[**3.3.3.** **Transformasi Data** 17](#_Toc106561390)

[**3.4.** **Modelling Correlated Naïve Bayes** 21](#_Toc106561397)

[***3.5.*** **Desain *User Interface*** 26](#_Toc106561405)

[***3.6.*** **Spesifikasi Alat Penelitian** 27](#_Toc106561414)

[**DAFTAR PUSTAKA** 28](#_Toc106561415)

# **DAFTAR GAMBAR**

Gambar 2.1. Proses *Knowledge Discovery Database*............................................ 5

Gambar 2.2. 5­-*Fold Cross Validation*................................................................... 10

Gambar 3.1. Desain Alat Uji................................................................................. 15

Gambar 3.2. Perangkingan atribut dengan *Information Gain*............................... 16

Gambar 3.3. Desain *Interface*............................................................................... 26

# **DAFTAR RUMUS**

Rumus 2.1. Normalisasi *Min-Max*.......................................................................... 6

Rumus 2.2. *Teorema Bayes*..................................................................................... 7

Rumus 2.3. *Correlated Naive Bayes*....................................................................... 8

Rumus 2.4. Korelasi atribut.................................................................................... 9

Rumus 2.5. R-*Square* antar kelas............................................................................ 9

Rumus 2.6. Akurasi............................................................................................... 11

Rumus 2.7. Presisi................................................................................................. 12

Rumus 2.8. *Recall*................................................................................................. 12

Rumus 2.9. F1-Score............................................................................................ 12

# **DAFTAR TABEL**

Tabel 2.1. Tabel Koefisien Korelasi......................................................................... 9

Tabel 2.2. Tabel R-*Square Korelasi*......................................................................... 9

Tabel 2.3. Tabel Model 3­-*Fold Cross Validation*....................................................10

Tabel 2.4. Tabel *Confusion Matrix*.........................................................................11

Tabel 3.1. Penjelasan Atribut..................................................................................13

Tabel 3.2. Contoh Data Awal data dataset Penyakit Jantung..................................14

Tabel 3.3. Perangkingan Atribut menggunakan aplikasi WEKA............................17

Tabel 3.4. Data Hasil Transformasi........................................................................20

Tabel 3.5. Data Atribut *Product Max*HR................................................................21

Tabel 3.6. Hasil Perhitungan R-*Square*...................................................................22

Tabel 3.7. Perhitungan *Prior Probability*...............................................................23

# **BAB I**

# **PENDAHULUAN**

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

## **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. (Kementrian Kesehatan-Peringatan Hari Jantung Sedunia, 2021).

Penyakit jantung sering dikenal sebagai *suddent 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 di fungsikan secara efektif bisa digunakan untuk deteksi penyakit jantung (Widiastuti dkk., 2014). 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 cara alternatif mencegah bahkan menyembuhkan penyakit jantung tersebut seperti dengan melakukan operasi, penyinaran dan khemoterapi. Namun, kurangnya akses informasi/media menjadi alasan penderita terlambat untuk memeriksa diri ke dokter dan menyebabkan peningkatan kematian setiap tahunnya. Karena itu, dibutuhkan sebuah sistem 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 (Bianto et al., 2020).

Menurut penelitian oleh Putra & Rini, (2019), menjelaskan bahwa penelitian penyakit jantung berfokus untuk menyelidiki suatu algoritma, apakah memiliki tingkat akurasi yang tinggi melalui 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 didapatlah algoritma *Naive Bayes* dengan tingkatan akurasi tertinggi dalam penelitian ini.

Hasil dari penelitian Komparasi akurasi metode *Correlated Naïve Bayes* dan *Naïve Bayes* yang digunakan untuk menyelesaikan kasus klasifikasi penyakit diabetes, menyebutkan bahwa metode *Correlated Naïve Bayes* menghasilkan 67,15% sedangkan metode *Naïve Bayes* menghasilkan sebesar 64,33%. Metode Correlated Naïve Bayes memiliki akurasi yang lebih tinggi dibandingkan metode *Naïve Bayes* (Hairani et al., 2018). Selain itu, ada penelitian yang melakukan pembobotan korelasi pada *Naïve Bayes Classifier*. Penelitian tersebut membuktikan bahwa nilai dari akurasi *Naïve Bayes Classifier* menghasilkan 77, 7773333% dan didapatkan peningkatan akurasi menggunakan *Correlated Naïve Bayes* menghasilkan 84,889% meningkatkan akurasi sebesar 7,111666667% secara signifikan setelah menggunakan metode *Correlated Naïve Bayes* (Muktamar et al., 2015).

Berdasarkan uraian diatas, penelitian ini mencoba membangun sistem untuk melihat potensi klasifikasi penyakit jantung 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.



## **Rumusan Masalah**

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

1. Bagaimana mengetahui atribut apa saja yang berpengaruh dalam klasifikasi penyakit jantung menggunakan dataset *heart.csv*?
2. Bagaimana hasil dari akurasi, presisi, dan f-1 score dari pengelompokkan penyakit jantung?

## **Tujuan Penelitian**

Tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Mengetahui atribut yang mempengaruhi hasil dari klasifikasi data penyakit jantung.
2. Menghasilkan sebuah sistem yang dapat mengetahui tingkat akurasi dalam melakukan klasifikasi terhadap penderita yang berpotensi memiliki penyakit jantung.

## **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*.

## **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.

## **Sistematika Penulisan**

Langkah-langkah sistematika penulisan, sebagai berikut :

1. **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.

1. **Bab II Landasan Teori**

Berisi penjelasan tentang teori-teori yang menjadi acuan dalam mendukung penulisan tentang klasifikasi penyakit jantung menggunakan algoritma Correlated Naïve Bayes.

1. **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**



## **Data Mining**



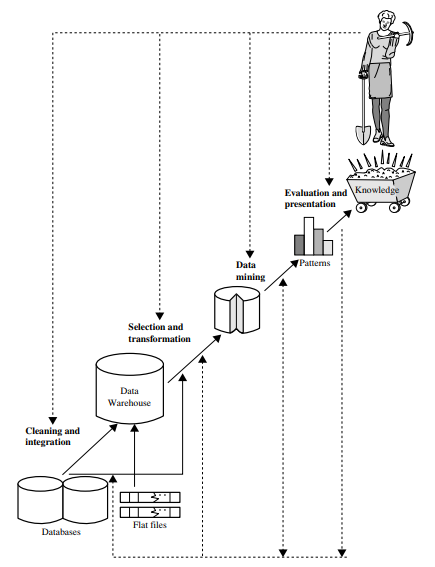
### **Pengertian Data Mining**

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).



### **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. berikut:



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.

1. Integrasi data (data integration)

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

1. 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.

1. 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).

(2.1)Keterangan:

: Nilai data baru hasil normalisasi *min-max*.

: Nilai data yang akan dinormalisasi.

: Nilai minimum data.

: Nilai maximum data.

: Nilai maximum dalam rentang.

: Nilai minimum dalam rentang.

1. Proses mining (data mining)

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

1. Evaluasi pola (pattern evaluation)

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

1. Presentasi pengetahuan (knowledge presentation)

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

## **Naïve Bayes**

Klasifikasi bayes sederhana yang lebih dikenal sebagai naïve Bayesian classifier yang diasumsikan 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 yang 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) = (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.



## **Correlated Naïve Bayes**

Metode *correlated naïve bayes classifier* merupakan pengembangan dari metode *naïve bayes* dimana pada metode *correlated naïve bayes classifier* akanmenghitung nilai yang terdapat dari korelasi atau *R-Square* diantara variabel bebas (X) pada variabel terikat (Y). Adanya penambahan pada parameter korelasi ini digunakan untuk mengukur tinggi rendahnya derajat hubungan variabel bebas (X) terhadap variabel terikat (Y) (Hairani et al., 2018). Rumus dari algoritma *correlated naïve bayes* untuk klasifikasi yang dapat dilihat pada Persamaan 2.3 berikut:

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

: Probabilitas setiap atribut dari data X berdasarkan kondisi hipotesis Y

: R-Square setiap atribut dari data X berdasarkan kondisi hipotesis Y.

: Bilangan laplacian

Berikut merupakan persamaan untuk menentukan perhitungan korelasi atribut :

(2.4)

(2.5)

Keterangan :

R : R-Square fitur antar kelas.

r : Nilai Korelasi antar fitur kelas.

n : Total data pada dataset.

: Total perkalian variabel X dengan variabel Y.

: Total variabel X.

: Total variabel Y.

: Total variabel X yang dikuadratkan.

: Total variabel Y yang dikuadratkan.

: Kuadrat dari total variabel X.

: 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 |



## **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). Pada umumnya k-fold cross validation digunakan untuk mengurangi waktu komputasi dengan menjaga keakuratan estimasi (Rachmat dkk., 2020).

Gambaran proses berjalannya 3-*fold cross validation*, dapat dilihat pada Tabel 2.3. berikut:

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

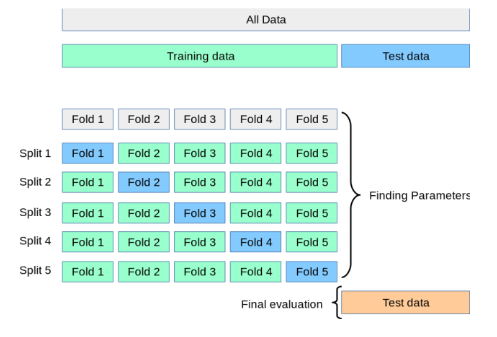
|  |  |  |
| --- | --- | --- |
| 1 | 2 | 3 |
| 1 | 2 | 3 |
| 1 | 2 | 3 |

Keterangan:

: data testing

: data training

Skenario untuk k-fold cross validation dengan k = 5 dapat dilihat pada Gambar 2.2 berikut:



Gambar 2.2. 5-*fold Cross Validation* (Admojo & Ahsanawati, 2020)



## **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 2.4. 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

Nilai akurasi didefinisikan sebagai perbandingan antara data yang diklasifikasi secara benar terhadap jumlah keseluruhan data (Satria dkk., 2020). Rumus akurasi terdapat pada Persamaan 2.6 berikut:

Akurasi = (2.6)

Presisi didefinisikan sebagai tingkat ketetapan antara permintaan informasi yang diminta oleh pengguna terhadap jawaban dari permintaan tersebut (Cahyanti dkk, 2020). Rumus presisi terdapat pada Persamaan 2.7 berikut:

Presisi = (2.7)

*Recall* merupakan rasio dari item relevan yang dipilih dengan total keseluruhan dari jumlah item yang tersedia (Baharuddin dkk., 2019). Rumus recall terdapat pada persamaan 2.8 berikut:

*Recall* = (2.8)

*F-measure* merupakan perhitungan nilai *balance* atau rata-rata akurasi dalam informasi yang diperoleh dari hasil presisi dan *recall*. *F-measure* juga kadang disebut dengan nama *F1-Score* (Kartika & Gondohanindijo, 2020). Rumus *F1-Score* dijabarkan pada persaman 2.9 berikut:

*F1-score* = (2.9)

# **BAB III**

# **METODE PENELITIAN**



## **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:

|  |  |
| --- | --- |
| 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 3.1. Penjelasan Atribut

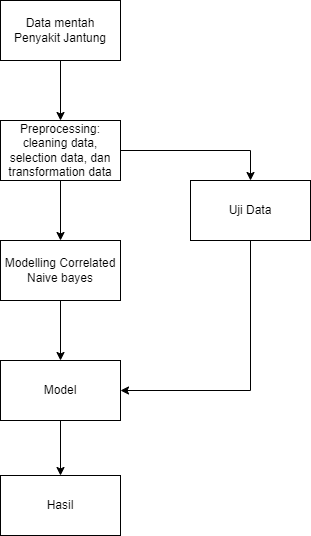
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST\_Slope | HeartDisease |
| 1 | 40 | M | ATA | 140 | 289 | 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 |

Tabel 3.2. Contoh Data Awal dari dataset Penyakit Jantung



## **Desain Alat Uji**

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



Gambar 3.1. Desai Alat Uji



## **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:



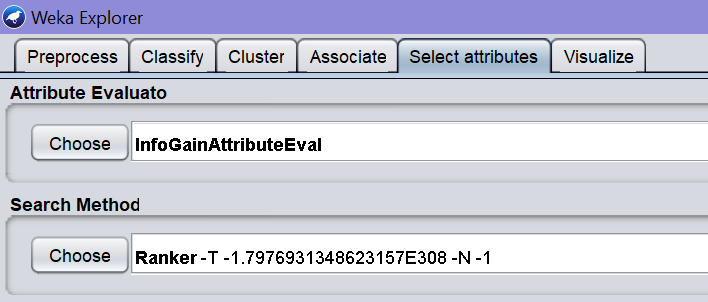
### **Pembersihan Data**

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.



### **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 |



### **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

1. Transformasi Atribut ChestPain Type

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

ATA : 1

NAP : 2

ASY : 3

1. Transformasi Atribut ExerciseAngina

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

Y : 1

N : 2

1. Transformasi Atribut ST\_Slope

Proses ini melakukan transformasi dari *st\_slope* kedalam bentuk numerik.

UP : 1

FLAT : 2

1. Normalisasi Min-max

Transformasi semua atribut menggunakan normalisasi *min-max.* Datatertinggi 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

Resting BP = 140

Cholesterol = 289

Max HR = 172

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:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

Tabel 3.4. Data Hasil Transformas



## **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) |  |  |  |
| 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 |
| () 5,513513513 | ()  16 | () 4,260043828 | ()  28 | ()  8,702702702 |

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 | Ranking | Keterangan |
| Age | 0,002121763 | 11 | Sangat rendah |
| Sex | 0,340277778 | 2 | Cukup |
| ChestPain Type | 0,330601093 | 3 | cukup |
| RestingBP | 0,221958144 | 6 | Cukup |
| Cholesterol | 0,319272606 | 4 | Cukup |
| FastingBS | 0,285714286 | 5 | cukup |
| RestingECG | 0,074074074 | 7 | rendah |
| MaxHR | 0,004829183 | 10 | Sangat rendah |
| ExerciseAngina | 0,010416667 | 8 | cukup |
| Oldpeak | 0,615384615 | 1 | Sangat kuat |
| ST\_Slope | 0,007936508 | 9 | Sangat rendah |

Notes : pilih terlebih dahulu atribut yang akan digunakan dalam penelitian berdasarakan perankingan dari r-square.

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) = = 0,25

P (No | Age = 0,176471) = = 0,333333333

P (Yes | Sex = 0) = = 0,75

P (No | Sex = 0) = = 0,5

P (Yes | ChestPain Type = 1) = = 0,5

P (No | ChestPain Type = 1) = = 0,333333333

P (Yes | RestingBP = 0,2) = = 0,5

P (No| RestingBP = 0,2) = = 0,333333333

P (Yes | Cholesterol = 0,68553459) = = 0,25

P (No | Cholesterol = 0,68553459) = = 0,333333333

P (Yes | FastingBS = 1) = = 0,75

P (No | FastingBS = 1) = = 0,333333333

P (Yes | MaxHR = 1) = = 0,25

P (No | MaxHR = 1) = = 0,333333333

P (Yes | ExerciseAngina = 0) = = 0,5

P (No | ExerciseAngina = 0) = = 0,333333333

P (Yes | Oldpeak = 0,666666667) = = 0,25

P (No | Oldpeak = 0,666666667) = = 0,333333333

P (Yes | ST\_Slope = 0) = = 1

P (No | ST\_Slope = 0) = = 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**

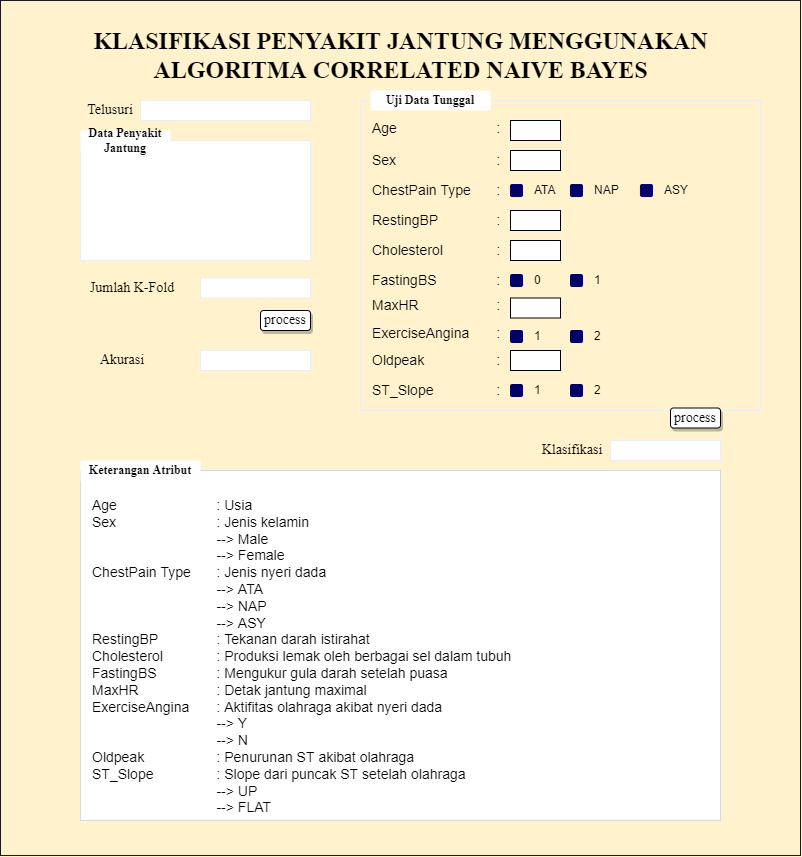
P (Yes | X) = **0,39772822**

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*.



## **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.

## **Spesifikasi Alat Penelitian**

Peralatan yang digunakan untuk menyelesaikan penelitian ini adalah sebagai berikut:

1. 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
5. 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**

Admojo, F. T., & Ahsanawati. (2020). *Klasifikasi Aroma Alkohol Menggunakan Metode KNN*. *1*(2), 34–38.

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.

Baharuddin, M. M., Hasanuddin, T., & Azis, H. (2019). *Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca*. *11*(28), 269–274.

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*.

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, *1*(2), 39–43. https://doi.org/10.33096/ijodas.v1i2.13

Hairani, Nugraha, G. S., Abdillah, M. N., & Innuddin, M. (2018). Komparasi Akurasi Metode Correlated Naive Bayes Classifier Dan Naive Bayes Classifier Untuk Diagnosis Penyakit Diabetes. *InfoTekJar (Jurnal Nasional Informatika Dan Teknologi Jaringan)*, *3*(1), 6–11.

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).

Kartika, E., & Gondohanindijo, J. (2020). Rancang Bangun Model Sentimen Analisis Review Produk Pada Toko Online Menggunakan Naive Bayes. *Seminar Nasional Hasil …*, 201–212. http://conference.upgris.ac.id/index.php/snhp/article/view/1222%0Ahttp://conference.upgris.ac.id/index.php/snhp/article/download/1222/869

Muktamar, B. A., Setiawan, N. A., & Adji, T. B. (2015). Pembobotan Korelasi Pada Naïve Bayes Classifier. *Seminar Nasional Teknologi Informasi Dan Multimedia 2015*, *2*, 43–47.

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.

Rachmat, R., Chrisnanto, Y. H., & Umbara, F. R. (2020). Sistem Prediksi Mutu Air Di Perusahaan Daerah Air Minum Tirta Raharja Menggunakan K–Nearest Neighbors (K–NN). *Prosiding SISFOTEK*, 189–193. http://www.seminar.iaii.or.id/index.php/SISFOTEK/article/view/211

Satria, F., Zamhariri, & Syaripudin, M. A. (2020). Prediksi Ketepatan Waktu Lulus Mahasiswa Menggunakan Algoritma C4.5 Pada Fakultas Dakwah Dan Ilmu Komunikasi UIN Raden Intan Lampung. *Jurnal Ilmiah Matrik*, *22*(1), 28–35. https://doi.org/10.33557/jurnalmatrik.v22i1.836

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). Data Mining Algoritma dan Implementasi Menggunakan Bahasa Pemrograman PHP. *Data Mining Algoritma Dan Implementasi Menggunakan Bahasa Pemrograman PHP*, *9*(9), 259–278.

Widiastuti, n a, Santosa, S., & c supriyanto. (2014). Algoritma klasifikasi data mining naive bayes berbasis particle swarm optimization untuk deteksi penyakit jantung. *Nature Methods*, *7*(1), 34. https://doi.org/10.1038/nmeth.f.284