

Pendekatan *correlated naïve bayes* pada klasifikasi potensi penyakit jantung

Stevani M Mellisa S Sagarung, Eko Hari Parmadi*

Program Studi Informatika Fakultas Sains Dan Teknologi Universitas Sanata Dharma

*Email: harimbi.parmadi@gmail.com

Abstrak

Penyakit jantung masih menduduki peringkat pertama dan penyebab utama kematian. Kondisi ini dapat dicegah dengan klasifikasi secara dini tentang serangan jantung berdasarkan dataset penyakit jantung yang terdapat pada keaggle.com. Dataset ini memiliki 11 atribut dan 1 label. 11 atribut tersebut meliputi: Age, sex, ChestPain Type, Cholesterol, Fasting BS, Resting BP, Resting ECG, Max HR, Exercise Angina, Oldpeak, ST_Slope. Sedangkan 1 label terdiri dari terdiri dari: satu kelas terkena penyakit jantung (yes) sebanyak 513 record, dan satu kelas tidak terkena penyakit jantung (no) sebanyak 405 record. Melalui tahapan preprocessing berupa *data selection*, dan *data cleaning* terpilihlah 7 atribut berdasarkan perangkikan tertinggi menggunakan *Weka 3.0*. Selanjutnya data dibagi menjadi data training dan data testing sesuai dengan variasi perbandingan 80% : 20%, 70% :30%, 60%: 40% dan 50% ; 50%. Selanjutnya dilakukan pengujian algoritme *correlated naïve bayes* dengan variasi atribut yang ada. Hasilnya diperoleh akurasi tertinggi sebesar 83, 1522% dengan perbandingan data training dan data testing sebesar 80% ; 20%. Sedangkan atribut yang berpengaruh dalam klasifikasi penyakit jantung menggunakan algoritme *correlated naïve bayes* adalah ST Slope, Exercise Angina, Max HR, Chest Pain Type

Kata Kunci: *correlated naïve bayes*; klasifikasi; penyakit jantung

1. Pendahuluan

Berdasarkan data kementerian Kesehatan RI, penyakit jantung masih menjadi penyebab utama kematian di Indonesia. Data Riset Kesehatan Dasar (Riskesdas) 2013 dan 2018 menunjukkan tren peningkatan penyakit jantung yakni 0,5% pada 2013 menjadi 1,5% pada 2018 (Kemkes, 2022). Bahkan data terbaru Organisasi Kesehatan Dunia (WHO) menunjukkan bahwa penyakit jantung koroner dan stroke masih menduduki peringkat pertama dan kedua penyebab kematian utama di dunia. Adapun jumlah kematian akibat penyakit jantung ini secara global mencapai 18,6 juta orang setiap tahunnya. Angka kematian tersebut diperkirakan akan terus meningkat menjadi 20,5 juta orang pada tahun 2020 dan 24,2 juta orang meninggal karena penyakit jantung pada tahun 2030 (Sumartiningtyas, 2022). Maka mendeteksi atau mengklasifikasi sejak dini apakah seseorang berpotensi penyakit jantung atau tidak menjadi suatu hal yang penting.

Jantung merupakan organ vital manusia yang berfungsi memompa darah ke seluruh tubuh. Bagi penderita yang memiliki penyakit jantung sangat berdampak buruk dan dapat berakibat pada kematian. Kondisi tersebut dapat dicegah, salah satunya melalui pemanfaatan data pasien serta melakukan pengecekan klasifikasi secara dini tentang serangan jantung yang dialami seseorang (Riani, dkk., 2019). Salah satu data yang bisa digunakan adalah dataset yang bersumber dari <https://www.kaggle.com/datasets/firdaus9914/penyakit-jantung>. Dataset ini memiliki 11 atribut dan 1 label.

Tujuan dari penelitian ini adalah mengetahui atribut-atribut yang berpengaruh dalam proses klasifikasi serta menghitung akurasi dari klasifikasi potensi penyakit jantung menggunakan pendekatan *correlated naïve bayes*. Dataset yang ada akan mengalami *preprocessing*, selanjutnya dimodelkan ke dalam *correlated naïve bayes*. Data yang telah dibagi menjadi *data testing* dan *data training* selanjutnya dilakukan pengujian kemudian dihitung nilai akurasinya. Algoritma *correlated naïve bayes* dipilih karena berdasarkan hasil dari beberapa penelitian akurasinya lebih tinggi daripada algoritma naïve bayes.

Prediksi penyakit jantung sudah pernah diteliti menggunakan metode *naïve bayes*. Data berasal dari RSUD AWS pada bulan November tahun 2016. Penelitian ini menggunakan 25 *data testing* diperoleh akurasi sebesar 80% dan menggunakan 50 data testing diperoleh akurasi sebesar 78% (Sabransyah, 2017). Peneliti lainnya juga menggunakan algoritma *naïve bayes* untuk klasifikasi penderita penyakit jantung menggunakan *rapid miner*. Penelitian ini menggunakan 100 data dan menghasilkan akurasi sebesar 70% (Maulana & Yahya, 2019).

Penelitian tentang Klasifikasi Sentimen Pergelaran Motogp di Indonesia dengan Metode *Correlated Naïve Bayes* menghasilkan skor akurasi sebesar 82% serta menggunakan sebanyak 1308 data yang diperoleh dari data public yang ada di twitter (Indransyah dkk., 2022). Algoritma *Correlated Naïve Bayes* terbukti mampu meningkatkan akurasi. Hal ini terbukti dari penelitian akurasi metode yang menggunakan metode *Naïve Bayes Classifier* (NBC) adalah 64,33%, sedangkan dari *Correlated Naive Bayes Classifier* (CNBC) adalah 67,15% (Hairani, H, dkk., 2018). Sementara Muktamar, dkk. (2015) juga melakukan penelitian tentang Analisis perbandingan tingkat akurasi algoritma naïve bayes classifier dengan correlated-naïve bayes classifier. Hasilnya algoritma *correlated naïve bayes* memiliki akurasi yang lebih tinggi dibandingkan naïve bayes (Muktamar, dkk., 2015). Algoritme *Correlated Naive Bayes Classifier* memberikan kenaikan akurasi dibandingkan algoritme *Naive Bayes Classifier* dengan data set yang diuji sebesar 14,03% secara significant.

2. Metode Penelitian

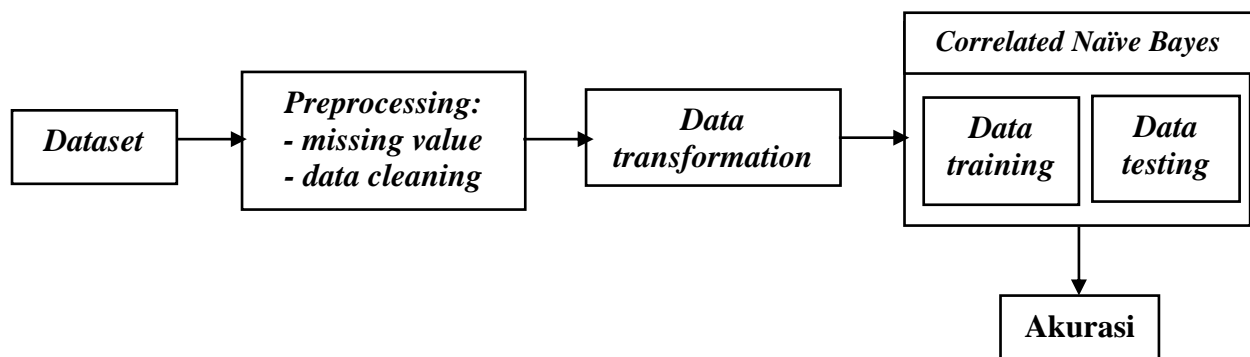
Data penelitian bersumber dari <https://www.kaggle.com/datasets/firdaus9914/penyakit-jantung>. yang diunduh pada tanggal 23 Juni 2022. Data terdiri dari 918 *record*, 11 atribut dan 1 label. Tabel 1 menjelaskan tentang atribut-atribut yang digunakan dalam penelitian ini.

Tabel 1. Keterangan Atribut

Nama Atribut	Keterangan
<i>Age</i>	Usia
<i>Sex</i>	Jenis kelamin
<i>ChestPain Type</i>	Jenis nyeri dada
<i>Resting BP</i>	Tekanan darah pasien ketika istirahat
<i>Cholesterol</i>	Kadar kolesterol dalam darah pasien
<i>Fasting BS</i>	Hasil pemeriksaan gula darah puasa
<i>Resting ECG</i>	Pemeriksaan EKG pada saat pasien kondisi istirahat/berbaring
<i>Max HR</i>	Detak jantung maksimal
<i>Exercise Angina</i>	Keadaan dimana pasien mengalami nyeri dada apabila berolahraga
<i>Oldpeak</i>	Penurunan ST akibat olahraga
<i>ST_Slope</i>	Slope dari puncak ST setelah berolahraga
<i>HeartDisease</i>	Penyakit jantung

Adapun labelnya terdiri dari: terkena penyakit jantung (*yes*) sebanyak 513 *record*, dan tidak terkena penyakit jantung (*no*) sebanyak 405 *record*.

Tahapan Penelitian secara ringkas dapat dilihat pada gambar 1 berikut ini:

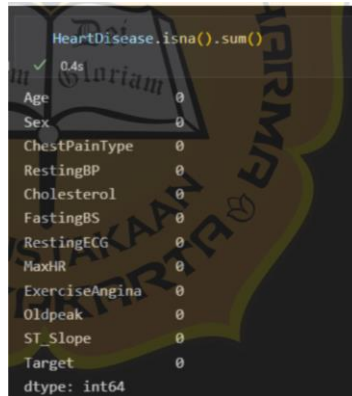


Gambar 1. Gambaran Penelitian

2.1. Preprocessing

Dalam data science, ada dua aspek penting yang perlu diperhatikan yaitu *preprocessing* dan *classification*. *Preprocessing* adalah teknik untuk menyiapkan data agar lebih siap untuk dilakukan lebih lanjut dalam rangka ekstraksi pengetahuan. Salah satu tahapan dalam *preprocessing* adalah: *data cleaning*. Tahap pertama yang perlu dilakukan ketika akan *preprocessing* data adalah *data cleaning* atau membersihkan data. Artinya, data mentah yang telah diperoleh perlu diseleksi kembali.

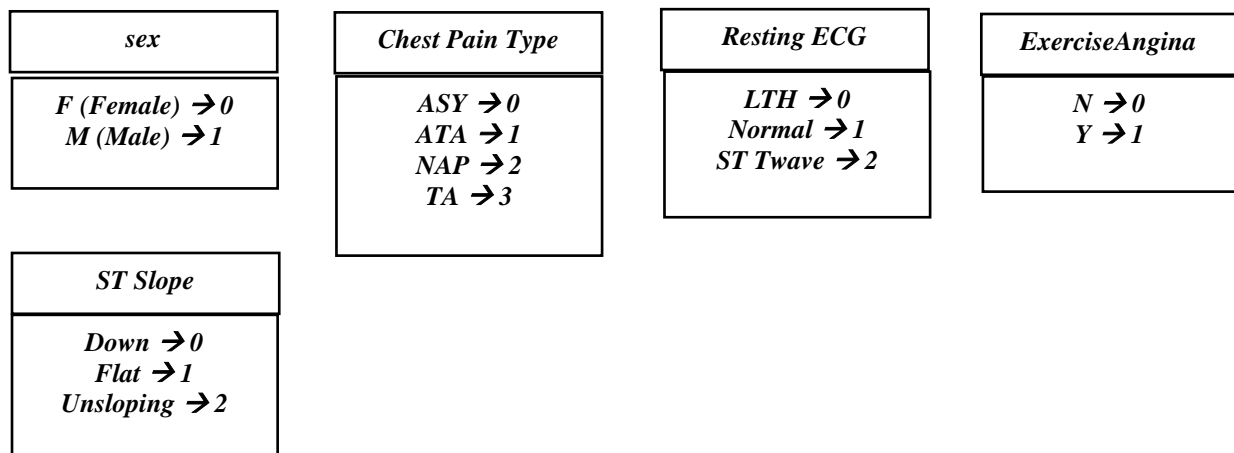
Kemudian, hapus atau hilangkan data-data yang tidak lengkap, tidak relevan, dan tidak akurat. Dengan melakukan tahap ini, Anda akan menghindari kesalahpahaman ketika menganalisis data tersebut. (Suripto, dkk., 2022). Tahapan *cleaning data* ini dapat dilihat pada gambar 2 berikut ini:



Gambar 2. Hasil Program Cek Missing Value

2.2. Data Transformation

Setelah data mengalami *preprocessing*, selanjutnya data yang ada dilakukan transformasi atau perubahan tipe data pada atribut untuk mempermudah penambangan data. Atribut-atribut seperti *sex*, *chest Pain Type*, *Resting ECG*, *Exercise Angina*, dan *ST Slope* yang berbentuk kategorial ditransformasi ke bentuk numerik.



Gambar 3. Transformasi Atribut

Data numerik selanjutnya dilakukan proses normalisasi agar rentang data antar atribut tidak terlalu lebar. Normalisasi *min-max* merupakan perubahan ukuran data dari rentang asli menjadi sebuah nilai numerik dalam rentang 0 sampai 1 (Ambarwari dkk., 2017). Adapun fungsi *min-max* yang digunakan, seperti pada persamaan (1) (Nurjanah dkk., 2017) berikut ini:

$$x_i^l = \text{minbaru}_A + \frac{(x_i - \text{min}_A)}{(\text{max}_A - \text{min}_A)} (\text{maxbaru}_A - \text{minbaru}_A) \dots\dots\dots (1)$$

Keterangan:

- x_i^l = nilai data baru hasil normalisasi
- x_i = nilai data yang akan dinormalisasi
- min_A = nilai data minimum

- max_A = nilai data maksimum
- $maxbaru_A$ = nilai maksimum dalam rentang baru
- $minbaru_A$ = nilai minimum dalam rentangbaru

Gambar 4 dan gambar 5 berikut ini merupakan tampilan data awal sebelum dilakukan transformasi dan sesudah dilakukan transformasi.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	Target
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
2	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
3	39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0
4	45	F	ATA	130	237	0	Normal	170	N	0.0	Up	0
...
913	57	F	ASY	140	241	0	Normal	123	Y	0.2	Flat	1
914	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
915	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
916	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
917	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1

Gambar 4. Data Sebelum Ditransformasi

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	Target
0	0.244898	1	1	0.70	0.479270	0	1	0.788732	0	0.295455	2	0
1	0.183673	1	1	0.65	0.469320	0	2	0.267606	0	0.295455	2	0
2	0.530612	1	2	0.75	0.323383	0	1	0.436620	0	0.295455	2	0
3	0.224490	1	2	0.60	0.562189	0	1	0.774648	0	0.295455	2	0
4	0.346939	0	1	0.65	0.393035	0	1	0.774648	0	0.295455	2	0
...
913	0.591837	0	0	0.70	0.399668	0	1	0.443662	1	0.318182	1	1
914	0.346939	1	3	0.55	0.437811	0	1	0.507042	0	0.431818	1	1
915	0.816327	1	0	0.72	0.320066	1	1	0.570423	0	0.681818	1	1
916	0.591837	1	0	0.65	0.217247	0	1	0.387324	1	0.431818	1	1
917	0.591837	0	1	0.65	0.391376	0	0	0.802817	0	0.295455	1	1

Gambar 5. Data Setelah Ditransformasi

2.3. Correlated Naive Bayes

Naive Bayes Classifier merupakan salah satu algoritme klasifikasi dalam data mining. Sampai saat ini, algoritme *Naive Bayes Classifier* hanya berdasar pada distribusi probabilitas. Algoritme *Naive Bayes Classifier* yang berbasis pada *probability attribute* dan bobot atribut dengan *R Square* dinamakan *Correlated Naive Bayes Classifier*. Algoritme *Correlated Naive Bayes Classifier* memberikan kenaikan akurasi dibandingkan algoritme *Naive Bayes Classifier* (Alfironi, B., 2015).

Nilai R square diperoleh dari rumus (2) (Muktamar, dkk., 2015) berikut ini:

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

$$R \text{ square} = r^2 \dots\dots\dots (3)$$

Keterangan:

- r = Nilai Korelasi antar fitur kelas.
- n = Total data pada dataset.
- X = Variabel X
- $\sum X$ = Total nilai variabel X

- $\sum Y$ = Total nilai variabel Y
- $\sum XY$ = Total perkalian nilai variabel X dengan nilai variabel Y
- $\sum X^2$ = Kuadrat dari total nilai variabel X
- $\sum Y^2$ = Kuadrat dari total nilai variabel Y

Setelah melalui proses transformasi, Langkah berikutnya adalah membuat model *correlated naïve bayes*. Data dibagi menjadi 2 bagian, yaitu data training bagian pertama dan data testing di bagian kedua. Pada penelitian ini, perbandingan data training dan data testing yang digunakan adalah 80%:20%, 70%:30%, 60%:40% dan 50%:50%. Pemilihan atribut pada penelitian ini didasarkan pada pemeringkatan berdasarkan program aplikasi *Weka 3.0* dan dipilih tujuh peringkat tertinggi pertama seperti pada tabel 2.

Tabel 2. Tujuh Peringkat Tertinggi Atribut berdasarkan *Weka 3.0*

Atribut	Peringkat
ST Slope	1
Exercise Angina	2
Max HR	3
Chest Pain Type	4
Oldpeak	5
sex	6
age	7

Atribut-atribut yang sudah diperingkat menggunakan *Weka 3.0* pada tabel 2 inilah yang akan digunakan untuk perhitungan *correlated naïve bayes* menggunakan rumus (4) berikut ini: (Muktamar, dkk., 2015)

$$P(Y|X) = \frac{P(Y) \sum P(X_i | Y) R(X_i | Y)}{P(X)} \dots\dots (4)$$

Keterangan:

- X = Data dengan kelas yang belum diketahui
- Y = Hipotesis data Y merupakan suatu kelas spesifik
- $P(Y|X)$ = Probabilitas data dengan vector X pada kelas Y
- $P(X)$ = Evidence atau probabilitas dari vektor X
- $P(Y)$ = Probabilitas dari kelas Y
- $\sum P(X_i | Y) R(X_i | Y)$ = Jumlah total dari perkalian antara probabilitas kelas Y dari fitur dalam vektor X dengan nilai korelasi fitur pada vektor X terhadap kelas Y

2.4. Confusion Matrix

Proses penghitungan akurasi pada data mining dapat menggunakan *confusion matrix* yang diilustrasikan dengan tabel yang berisi jumlah data uji yang benar dan data uji yang salah diklasifikasikan (Bramer, 2016). Tabel akurasi tersebut dapat dilihat pada gambar 6, berikut ini:

		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negative	True Negative

Gambar 6. *Confusion Matrix* (sumber: <https://www.sciencedirect.com>)

Keterangan:

- True Positive (TP) = Jumlah data positif yang diklasifikasikan menjadi kelas positif
False Positive (FP) = Jumlah data negatif yang diklasifikasikan menjadi kelas positif
False Negative (FN) = Jumlah data positif yang diklasifikasikan menjadi kelas negatif
True Negative (TN) = Jumlah data negatif yang diklasifikasikan menjadi kelas negatif

Kinerja dari model klasifikasi diukur dari nilai akurasi berdasarkan rumus (5), yaitu: (Normawati, D., & Prayogi, S. A., 2021)

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \dots\dots\dots (5)$$

3. Hasil dan Pembahasan

Berdasarkan proses klasifikasi, variasi atribut serta pengujian menggunakan variasi perbandingan data training dan data testing. Adapun hasil pengujian dengan berbagai variasi, dapat dilihat pada tabel 3, tabel 4, tabel 5 dan tabel 6.

Tabel 3. Hasil Akurasi Klasifikasi Menggunakan Perbandingan 80%: 20%

Jumlah Atribut	Nama Atribut	Akurasi (%)
1	ST Slope	80,4348
2	ST Slope, Exercise Angina	81,5217
3	ST Slope, Exercise Angina, Max HR	81,5217
4	ST Slope, Exercise Angina, Max HR, Chest Pain Type	83,1522
5	ST Slope, Exercise Angina, Max HR, Chest Pain Type, Oldpeak	83,1522
6	ST Slope, Exercise Angina, Max HR, Chest Pain Type, Oldpeak, sex	83,1522
7	ST Slope, Exercise Angina, Chest Pain Type, Oldpeak, Sex, age	83,1522

Tabel 4. Hasil Akurasi Klasifikasi Menggunakan Perbandingan 70%: 30%

Jumlah Atribut	Nama Atribut	Akurasi (%)
1	ST Slope	78,2609
2	ST Slope, Exercise Angina	79,7101
3	ST Slope, Exercise Angina, Max HR	79,7101
4	ST Slope, Exercise Angina, Max HR, Chest Pain Type	79,7101
5	ST Slope, Exercise Angina, Max HR, Chest Pain Type, Oldpeak	81,8841
6	ST Slope, Exercise Angina, Max HR, Chest Pain Type, Oldpeak, sex	81,8841
7	ST Slope, Exercise Angina, Chest Pain Type, Oldpeak, Sex, age	81,8841

Tabel 5. Hasil Akurasi Klasifikasi Menggunakan Perbandingan 60%: 40%

Jumlah Atribut	Nama Atribut	Akurasi (%)
1	ST Slope	79,3478
2	ST Slope, Exercise Angina	80,7065
3	ST Slope, Exercise Angina, Max HR	80,4348
4	ST Slope, Exercise Angina, Max HR, Chest Pain Type	80,4348
5	ST Slope, Exercise Angina, Max HR, Chest Pain Type, Oldpeak	82,3370
6	ST Slope, Exercise Angina, Max HR, Chest Pain Type, Oldpeak, sex	82,0652
7	ST Slope, Exercise Angina, Chest Pain Type, Oldpeak, Sex, age	82,0652

Tabel 6. Hasil Akurasi Klasifikasi Menggunakan Perbandingan 50%: 50%

Jumlah Atribut	Nama Atribut	Akurasi (%)
1	ST Slope	79,7386
2	ST Slope, Exercise Angina	80,3922
3	ST Slope, Exercise Angina, Max HR	79,9564
4	ST Slope, Exercise Angina, Max HR, Chest Pain Type	79,9564

Jumlah Atribut	Nama Atribut	Akurasi (%)
5	ST Slope, Exercise Angina, Max HR, Chest Pain Type, Oldpeak	81,2636
6	ST Slope, Exercise Angina, Max HR, Chest Pain Type, Oldpeak, sex	82,5708
7	ST Slope, Exercise Angina, Chest Pain Type, Oldpeak, Sex, age	82,5708

Dari hasil pengujian berbagai variasi data training dan data testing serta variasi atribut maka diperoleh akurasi optimal pada pengujian menggunakan variasi data training dan data testing sebesar 80%: 20%. Sedangkan atribut yang digunakan adalah sebanyak 4 atribut yaitu: ST Slope, Exercise Angina, Max HR, Chest Pain Type dengan akurasi sebesar 83,1522 %. Penambahan atribut menjadi 5, 6, atau 7 menghasilkan nilai akurasi yang sama.

4. Kesimpulan

Berdasarkan pembahasan dan hasil penelitian diatas dapat disimpulkan bahwa:

- 1) Atribut-atribut yang berpengaruh pada klasifikasi penyakit jantung menggunakan pendekatan *correlated naïve bayes* adalah ST Slope, Exercise Angina, Max HR, Chest Pain Type
- 2) Akurasi tertinggi pada klasifikasi penyakit jantung menggunakan pendekatan *correlated naïve bayes* sebesar 83, 1522% dengan perbandingan data training dan data testing sebesar 80% : 20% dan 4 atribut.

Daftar Pustaka

- Alfironi, B. (2015). *Correlated Naive Bayes Classifier* (Doctoral dissertation, Universitas Gadjah Mada).
- Ambarwari, A., Adrian, Q. J., & Herdiyeni, Y. (2020). Analisis pengaruh data scaling terhadap performa algoritme machine learning untuk identifikasi tanaman. *J. Rekayasa Sist. dan Teknol. Inf*, 4(1), 117-112.
- Bramer, M. (2007). *Principles of data mining* (Vol. 180, p. 2). London: Springer.
- Hairani, H., Nugraha, G. S., Abdillah, M. N., & Innuddin, M. (2018). Komparasi akurasi metode *correlated naïve Bayes classifier* dan *naïve Bayes classifier* untuk diagnosis penyakit diabetes. *InfoTekJar: Jurnal Nasional Informatika Dan Teknologi Jaringan*, 3(1), 6-11.
- Indransyah, R., Chrisnanto, Y. H., & Sabrina, P. N. (2022). Klasifikasi Sentimen Pergelaran Motogp di Indonesia Menggunakan Algoritma *Correlated Naïve Bayes Clasifier*. *Infotech journal*, 8(2), 60-66.
- Maulana, D., & Yahya, R. (2019). Implementasi Algoritma *Naïve Bayes* Untuk Klasifikasi Penderita Penyakit Jantung Di Indonesia Menggunakan *Rapid Miner*. *Jurnal SIGMA*, 10(2), 191-197.
- Muktamar, B. A., Setiawan, N. A., & Adji, T. B. (2015). Pembobotan Korelasi pada *Naive Bayes Classifier*. *SEMNASSTEKNOMEDIA ONLINE*, 3(1), 2-1.
- Muktamar, B. A., Setiawan, N. A., & Adji, T. B. (2015). Analisis perbandingan tingkat akurasi algoritma *naïve bayes classifier* dengan *correlated-naïve bayes classifier*. *Semnasteknomedia Online*, 3(1), 2-1.
- Normawati, D., & Prayogi, S. A. (2021). Implementasi *Naïve Bayes Classifier* Dan *Confusion Matrix* Pada Analisis Sentimen Berbasis Teks Pada Twitter. *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, 5(2), 697-711.
- 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*, 1(12), 1750-1757.
- Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer (2022). [cited: 28 Juni 2023]. Available at: <https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20220929/0541166/penyakit-jantung-penyebab-utama-kematian-kemenkes-perkuat-layanan-primer/>
- Riani, A., Susianto, Y., & Rahman, N. (2019). Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Menggunakan Metode *Naive Bayes*. *Journal of Innovation Information Technology and Application (JINITA)*, 1(01), 25-34.

- Sabransyah, M., Nasution, Y. N., & Amijaya, F. D. T. (2017). Aplikasi Metode Naive Bayes dalam Prediksi Risiko Penyakit Jantung. *EKSPONENSIAL*, 8(2), 111-118.
- Sumartiningsih H.K.N. (2022). Angka Kematian Penyakit Jantung Masih Tinggi di Indonesia Jadi Tantangan Perki [cited: 28 juni 2023]. Available from : <https://www.kompas.com/sains/read/2022/08/04/163200923/angka-kematian-penyakit-jantung-masih-tinggi-di-indonesia-jadi-tantangan?page=all>
- Suripto, Rr Nurul Rahmanita dan Ajeng Sekar Kirana. (2022). Teknik pre-processing dan classification dalam data science [cited: 28 juni 2023]. Available from: <https://mie.binus.ac.id/2022/08/26/teknik-pre-processing-dan-classification-dalam-data-science/#:~:text=Melalui%20data%20preprocessing%2C%20memungkinkan%20proses,sudah%20melalui%20beberapa%20tahap%20pembersihan.>