BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penelitian ini menggunakan sumber pustaka yang berhubungan dengan kasus yang diteliti, antara lain :

Tabel 2. 1 Perbandingan

Penulis	**Judul Penelitian**	**Metode**	**Objek**	**Kriteria**
Faza, A. (2019)	Sistem Klasifikasi Stadium Karies Gigi menggunakan *naïve Bayes Classifier*	*Naïve Bayes*	Stadium Karies Gigi	Uji coba prediksi kategori kelas dengan data pengujian (*testing*) sebanyak 3 kali dengan uji yaitu 20,40 dan 60 record, diperoleh rata-rata tingkat presentasi ketelitian sebesar 82,66%.
Rizkyani, Erisa, Aliffiyanti Iskandar, Nur, Chamidah, Nurul (2021)	Klasifikasi dalam Mendeteksi Penyakit Kanker Payudara dengan Menggunakan *Metode Random Forest* dan *Adaboost*	*Random Forest, Adoboost*	Penyakit kanker Payudara	Berdasarkan penelitian yang telah dilakukan, diperoleh nilai akurasi dengan menggunakan *Random Forest* sebesar 95% dan nilai akurasi dengan menggunakan *Adaboost* sebesar 70%.
Ramadhan l, Kurniawati K (2020)	*Data Mining* untuk Klasifikasi Penderita Kanker Payudara Berdasarkan Data dari *University Medical Center* Menggunakan	*Naïve Bayes*	Kanker Payudara	Akurasi terhadap klasifikasi data yang sudah dilakukan sebelumnya menggunakan algoritma *Naive Bayes*, diperoleh nilai akurasi terhadap pasien yang mengalami kambuh atau tidak sebesar 71,43%

Hadi Kristanto, W. and Fitri, V.A. (2018)	Penerapan *Algoritma Naïve Bayes Berbasis Particle Swarm Optimization (Pso)* Dalam Menangani Kasus Kanker Payudara.	*Naïve Bayes*	Kanker payudara	Hasil penelitian ini menunjukan bahwa pengguna *PSO* dalam algoritma *Naïve Bayes* dapat meningkatkan akurasi klasifikasi pada kasus kanker payudara 96.76%.
Rekayasa K, Nugraheni A, Dias Ramadhani R et al. (2022)	Perbandingan Performa Antara Algoritma Naive Bayes dan K-Nearest Neighbour Pada Klasifikasi Kanker payudara	*Naïve Bayes* dan *K-Nearest Neighbour*	Kanker Payudara	Hasil dari penelitian ini menunjukkan bahwa algoritma Naïve Bayes memiliki rata-rata akurasi tertinggi 69.12%, presisi sehat 64.90%, presisi sakit 83%, recall sehat 88%, recall sakit 61.11% dan AUC 0.8. Sedangkan untuk hasil rata-rata tertinggi algoritma K-Nearest Neighbour adalah 76.83% untuk akurasi, presisi sehat 76%, presisi sakit 80.21%, 74.18% untuk recall sehat, recall sakit 80.81%.
Meilani N, Nurdiawan O (2023)	*Data Mining* Untuk Klasifikasi Penderita kanker Payudara Menggunakan Algoritma *K-Nearest Neighbor*	*K-Nearest Neighbor*	Kanker Payudara	Hasil Menunjukkan Persentase yang cukup baik yaitu menghasilkan *accuracy* sebesar 72,62% dengan nilai K = 5, pada dataset pasien kanker prediksi *no-recurrence-events* (tidak kambuh) lebih mendominasi.
Novayanti Klau, (2024)	Klasifikasi Dalam Mendeteksi Penyakit Kanker Payudara Menggunakan *Metode Naïve Bayes*	*Naïve Bayes*	Kanker Payudara	Rencana Penelitian ini akan melakukan klasifikasi pada penyakit kanker payudara dengan *confusion matrix* untuk mengetahui seberapa baik akurasi yang didapatkan dari model *naïve bayes*

Metode *Naïve Bayes* dapat digunakan untuk memprediksi kategori kelas Karies gigi pada poli RSUD dr. Soehadi Prijonegoro-Sragen. Berdasarkan hasil percobaan, Program diujikan pada data pelatihan (*training*) dengan jumlah data 200 record. Kemudian uji coba prediksi kategori kelas dengan data pengujian (*testing*) sebanyak 3 kali dengan uji yaitu 20,40 dan 60 record, diperoleh rata-rata tingkat persentase ketelitian sebesar 82,66%. (Faza, A. 2019)

Penelitian ini dilakukan dalam mengklasifikasi pendeteksian penyakit kanker dengan menggunakan Metode *Random Forest* dan *Adaboost*. Dalam kasus Pendeteksian Penyakit Kanker Payudara terdapat penelitian yang menggunakan metode *random forest* dan *Adaboost*, penelitian ini memperoleh nilai akurasi dengan menggunakan Random Forest sebesar 95% dan nilai akurasi dengan menggunakan *Adaboost* sebesar 70%. Hal ini menunjukkan bahwa pendeteksian penyakit kanker payudara menggunakan random forest lebih baik digunakan dibandingkan dengan menggunakan Adaboost. (Rizkyani, Erisa, Aliffiyanti Iskandar, Nur, Chamidah, Nurul (2021)

Penelitian menggunakan *data mining* untuk klasifikasi penderita kanker payudara, kelas data terbagi menjadi 2 kelompok yaitu kelas kambuh dan tidak kambuh. Berdasarkan perhitungan data mining menggunakan algoritma *Naive Bayes*, dapat dikatakan bahwa kelas pasien “Kambuh” lebih besar dibandingkan kelas pasien “Tidak Kambuh”. Akurasi terhadap klasifikasi data yang sudah dilakukan sebelumnya menggunakan algoritma *Naive Bayes*, diperoleh nilai akurasi terhadap pasien yang mengalami kambuh atau tidak sebesar 71,43% dimana

hal ini juga bisa disebabkan oleh kekurang kompleksan data sehingga model dapat memprediksi secara akurat. Ramadhan l, Kurniawati K (2020)

Penelitian ini bertujuan untuk meningkatkan kinerja algoritma *Naïve Bayes* dalam mengklasifikasi kanker payudara dengan menggabungkannya dengan Teknik optimasi yaitu *Particle Swarm Optimization* (PSO). PSO digunakan untuk memperbaiki parameter-parameter yang digunakan oleh *Naïve Bayes,* sehingga meningkatkan kemampuan algoritma dalam menganalisis dan mengklasifikasikan data kanker payudara. Hasil penelitian menunjukan bahwa pengguna PSO dalam algoritma *Naïve Bayes* dapat meningkatkan akurasi klasifikasi pada kasus kanker payudara 96.76%. Hadi Kristanto, W. and Fitri, V.A. (2018)

Dalam proses penelitian menggunakan dataset penyakit kanker payudara dari dataset Breast Cancer Coimbra tahun 2018 UCI Machine Learning Repository dengan total 116 data, sedangkan untuk perhitungan kelayakan metode menggunakan Confusion Matrix (Akurasi, Presisi, dan Recall) dan kurva ROC-AUC. Tujuan dari penelitian ini adalah membandingkan performansi algoritma Naïve Bayes dan K-Nearest Neighbour. Pada pengujian menggunakan algoritma Naïve Bayes dan algoritma K-Nearest Neighbour, terdapat beberapa skenario pengujian yaitu, pengujian data sebelum dan sesudah normalisasi, pengujian model berdasarkan perbandingan data training dan data testing, pengujian model berdasarkan nilai K pada K-Nearest Neighbour, dan pengujian model berdasarkan pemilihan atribut terkuat dengan uji korelasi Pearson. Hasil dari penelitian ini menunjukkan bahwa algoritma Naïve Bayes memiliki rata-rata akurasi tertinggi 69.12%, presisi sehat 64.90%, presisi sakit 83%, recall sehat 88%, recall sakit

61.11% dan AUC 0.82 yang termasuk kategori good classification. Sedangkan untuk hasil rata-rata tertinggi algoritma K-Nearest Neighbour adalah 76.83% untuk akurasi, presisi sehat 76%, presisi sakit 80.21%, 74.18% untuk recall sehat, recall sakit 80.81% dan AUC 0.91 yang termasuk kategori excellent classification. Rekayasa K, Nugraheni A, Dias Ramadhani R et al. (2022)

Dalam penelitian ini menggunakan algoritma *K-Nearest Neighbor*. *K-Nearest Neighbor* adalah algoritma untuk melakukan klasifikasi terhadap objek dengan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Alat yang digunakan untuk mencari nilai akurasi adalah *rapidminer* versi 10. Metode yang digunakan dalam penelitian ini adalah metode *knowledge Discovery In.* Penelitian ini digunakan untuk mengambil keputusan berdasarkan hasil yang didapat untuk menentukan kebijakan yang akan diambil dalam penanganan pasien kanker payudara. Hasil Menunjukkan *Persentase* yang cukup baik yaitu menghasilkan *accuracy* sebesar 72,62% dengan nilai K = 5, pada dataset pasien kanker prediksi *no-recurrence-events* (tidak kambuh) lebih mendominasi. Meilani N, Nurdiawan O (2023)

2.1 Dasar Teori

2.2.1 Kanker Payudara

Kanker Payudara Merupakan penyakit ke-2 terbanyak yang menjadi penyebab kematian kanker setelah kanker paru-paru pertama. Kanker payudara terjadi saat sel- sel pada jaringan payudara mulai memiliki pertumbuhan yang tidak dapat dikendalikan serta dapat mengganggu jaringan sehat yang ada. Kanker

payudara umumnya dibedakan ke dalam dua jenis yaitu kanker jinak dan kanker ganas. Untuk kanker jinak adalah kanker yang kondisinya masih dalam tahap awal sehingga jenis kanker ini masih bisa ditangani oleh tenaga medis dan bahkan sangat besar potensi penyembuhannya. Sedangkan kanker ganas adalah jenis kanker yang sangat berbahaya apabila tidak ditangani dengan baik akan berpotensi pada kematian.(Farahdiba and Nugroho, 2016)

2.2.2 Klasifikasi

Klasifikasi merupakan suatu proses menilai objek model atau fungsi yang mendeskripsikan dan membedakan data ke dalam label yang ada. Dalam klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkan objek ke dalam salah satu kelas yang sudah didefinisikan sebelumnya.

2.2.3 Naïve Bayes

Naïve Bayes merupakan sebuah pengklasifikasian *probabilistic* sederhana yang digunakan untuk menghitung sekumpulan probabilitas dengan menjumlahkan *frekuensi* dan kombinasi nilai dari dataset. Algoritma menggunakan *teorema Bayes* dan mengasumsikan semua atribut *independen* atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan *Naïve Bayes* merupakan pengklasifikasian dengan metode *probabilitas statistic* yang dikemukakan oleh *ilmuwan Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Alfa, 2015).

Naïve Bayes memiliki Persamaan *Teorema Bayes* adalah (Alfa, 2015):

$$P(Y|X_{1\ldots}X_n) = \frac{P(X|Y)*P(X1\ldots Xn|Y)}{P(X1\ldots Xn)}$$

Dimana:

1. P(Y|X) : Probabilitas Kelas Y berdasarkan kondisi Kelas X.
2. P(Y) : Probabilitas hipótesis Y (prior dari kelas Y).
3. P(X|Y) : Probabilitas X berdasarkan kondisi pada hipotesis Y.
4. P(X) : Probabilitas Dari Data X.

Untuk menjelaskan metode *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok dengan sampel yang dianalisis. Dimana: Variabel Y merepresentasikan kelas, sementara *variable* Xl…Xn merepresentasikan atribut-atribut yang dibutuhkan untuk melakukan klasifikasi.

2.2.4 Preprocessing Data

Preprocessing data merupakan tahapan dalam data mining untuk mengolah mempersiapkan data agar sesuai dengan format atau struktur data yang dibutuhkan.

2.2.5 Confusion Matrix

Confusion Matrix merupakan Pengukuran *performa* untuk masalah klasifikasi *machine learning* yang dimana output dapat berupa dua kelas atau lebih tabel, dengan 4 kombinasi yang berbeda dari nilai prediksi dan nilai aktual. Ada 4 istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu *True Positif, True Negatif, False Positif,* dan *False negatif.*

Tabel 2. 2 Penjelasan Confusion Matrix

Nilai Aktual	Nilai Prediksi		Total
	P	N	
P	TP (True Positive)	FN (False Negative)	TP+FN
N	FP (False Neggatif)	TN (True Negative)	FP+TN
Total	TP+FP	FN+TN	

True positive atau TP adalah jumlah data positif yang telah diklasifikasi sebagai *positive*. *True negative* atau TN adalah jumlah data negatif yang terklasifikasi sebagai *negative*. *False Positive* atau FP adalah jumlah data negatif yang terklasifikasi sebagai positive atau salah prediksi. *False Negative* atau FN adalah jumlah data positif yang terklasifikasi sebagai *negative* atau salah prediksi.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

$$\text{Precision} = \frac{TP}{TP+FP} \text{ X } 100\%$$

$$\text{Recall} = \frac{TP}{TP+FN} \text{ X } 100\%$$

Keterangan

1. Accuracy : jumlah prediksi benar positif dan negatif dengan total keseluruhan data.
2. Precision : jumlah prediksi benar positif dengan keseluruhan hasil yang diprediksi benar positif.
3. Recall : jumlah prediksi benar positif dengan keseluruhan data yang benar positif.
4. F1-Score : Jumlah nilai rata-rata dari precision dan recall.

2.2.6 Python

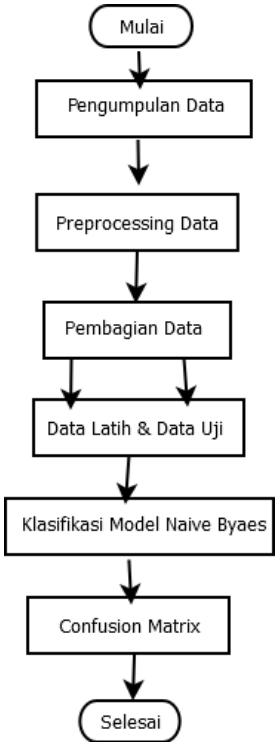
Bahasa pemrograman *python* merupakan salah satu Bahasa pemrograman komputer yang biasa dipakai untuk membangun *situs, software*/aplikasi termasuk *machine learning* dan analisis data. Bahasa pemrograman ini memungkinkan analisis data untuk melakukan perhitungan *statistic*, membuat visualisasi data serta algoritma *machine learning*. Bahasa pemrograman ini termasuk dalam bahasa tujuan umum. Artinya ia bisa digunakan untuk membuat berbagai pemrograman yang berbeda, karna *python* bersifat relatif dan mudah untuk dipelajari. (dicoding.com)

2.2.7 Streamlit

Streamlit adalah sebuah *framework* open- *source* yang mudah digunakan untuk membangun antarmuka pengguna (UI) interaktif untuk aplikasi data analisis. *Framework* ini dirancang khusus untuk mempermudah pengembangan aplikasi *web* dengan menggunakan bahasa pemrograman *python.* (dqlab.id)

2.2.8 Alur Penelitian

Pada gambar 2.1 merupakan proses tahapan yang akan dilakukan dalam penelitian.



Gambar 2. 1 Tahapan Penelitian

Gambar 2.1 merupakan proses tahapan penelitian dimulai dari pengumpulan data. Data diperoleh dari kaggle.com/datasets/yasserhessein/breast-cancer-coimbra-data-set/data sebanyak 116 data, 9 *variabel* dan 2 kelas yaitu: 2 (pasien kanker payudara), 1 (Kesehatan yang Terkontrol). Selanjutnya dilakukan proses *preprocessing* data dengan memeriksa jumlah nilai yang hilang (*missing values*) pada setiap kolom, dan memeriksa apakah terdapat data yang dobel(duplikat data) yang terdapat dalam data tersebut. Selanjutnya dari 116 data dilakukan Pembagian data yaitu data *training* 80% dan data *testing* 20%. Kemudian melakukan proses pelatihan dan pengujian dengan klasifikasi Model *Naive Bayes* untuk menentukan kelas sehingga mendapatkan hasil klasifikasi.Tahap terakhir yang dilakukan yaitu evaluasi, dimana menggunakan *confusion matrix* untuk mendapatkan hasil dari *accuracy presisi*, dan *recall.*