e-ISSN: 2460-870X | p-ISSN: 2085-1367

Komparasi Algoritma Decision Tree, Naive Bayes, dan K-Nearest Neighbors dalam Klasifikasi Kanker Payudara

Comparison of Decision Tree Algorithms, Naive Bayes, and K-Nearest Neighbors in Breast Cancer Classification

Muhammad Abdul Jabbar[a,1], Erfan Hasmin[b,2], Sunardi[c,3], Cucut Susanto[d,4], Wilem Musu[e,5]
[a,b,c,d]Universitas Dipa Makassar; Jl. Perintis Kemerdekaan KM. 9 Makassar, Indonesia
jabbar.kbj14@gmail.com[1], erfan.hasmin@undipa.ac.id[2], sunardi@undipa.ac.id[3], cucut@undipa.ac.id[4], wilem.musu@undipa.ac.id[5]

ABSTRAK

Kanker payudara merupakan tipe kanker yang umumya terbentuk di sel-sel payudara dan sel-sel kanker tersebut tumbuh diluar kendali. Kanker payudara dapat terjadi pada semua gender. Di Indonesia, jumlah kasus kanker payudara sampai menempati urutan pertama dibandingkan jenis kanker yang lain dan menjadi salah satu penyumbang kematian pertama. Berdasarkan jumlah kasus kematian tersebut dan mengingat kanker payudara tidak memandang gender, seharusnya baik pria maupun wanita sadar dengan kesehatan mereka dengan cara melakukan tindakan seperti deteksi dini dan menghindari risiko yang menyebabkan terjadinya kanker. Data yang digunakan dalam penelitian ini berasal dari UCI Machine Learning Repository. Tujuan dari penelitian ini yaitu dapat membuat Artificial Intelligence yang bisa mendeteksi kanker payudara secara dini dengan memanfaatkan beberapa algoritma data mining yang ada dan membandingkan tiga algoritma data mining dalam melakukan klasifikasi terhadap kanker payudara. Pada penelitian ini algoritma yang digunakan dalam melakukan perbandingan adalah algoritma Decision Tree, Naive Bayes dan K-Nearest Neighbors dengan menggunakan 2 metode cross-validation, Hold-Out dan K-Fold. Hasil dari pengujian menunjukkan bahwa algoritma K-Nearest Neighbors selalu menghasilkan performa akurasi yang sangat baik dibanding algoritma Naive Bayes dan Decision Tree, yaitu 98% pada metode Hold-Out dan 96% pada metode K-Fold, sedangkan Naive Bayes 95% pada metode Hold-Out dan 95% pada metode K-Fold, dan Decision Tree 94% pada metode Hold-Out dan 93% pada metode K-Fold.

Kata Kunci : Kanker payudara, Decision Tree, Naive Bayes, KNN

ABSTRACT

Breast cancer is a type of cancer that generally forms in breast cells and the cancer cells grow out of control. Breast cancer can occur in all genders. In Indonesia, the number of breast cancer cases ranks first compared to other types of cancer and is one of the first contributors to deaths. Based on the number of deaths and considering that breast cancer does not look at gender, both men and women should be aware of their health by taking measures such as early detection and avoiding the risks that cause cancer. The data used in this study came from the UCI Machine Learning Repository. The purpose of this study is to create Artificial Intelligence that can detect breast cancer early by utilizing several existing data mining algorithms and comparing three data mining algorithms in classifying breast cancer. In this study, the algorithms used in making comparisons were the Decision Tree, Naive Bayes, and K-Nearest Neighbors algorithms using 2 methods of cross-validation, Hold-Out, and K-Fold. The results of the test show that the K-Nearest Neighbors algorithm always produces excellent accuracy performance compared to the Naive Bayes and Decision Tree algorithms, namely 98% in the Hold-Out method and 96% in the K-Fold method, while Naive Bayes is 95% in the Hold-Out method and 95% in the K-Fold method, and the Decision Tree is 94% in the Hold-Out method and 93% in the K-Fold method.

Keywords : Breast Cancer, Decision Tree, Naive Bayes, KNN

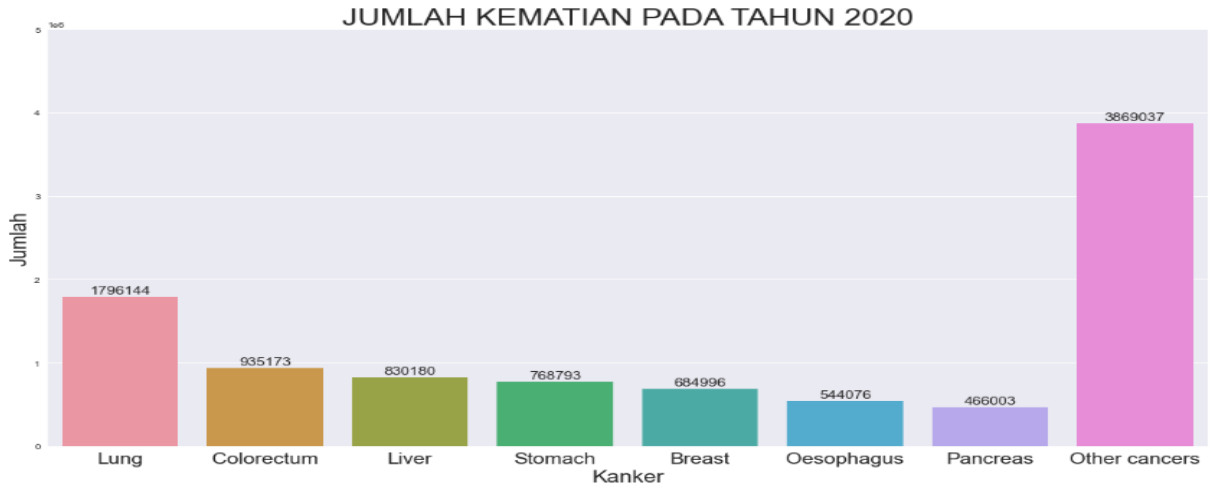
Info Artikel :
Disubmit: 15 September 2022 *Direview: 18 Oktober 2022* *Diterima: 29 November 2022*

DOI: https://www.doi.org/10.22303/csrid.14.3.2022.258-270

e-ISSN: 2460-870X | p-ISSN: 2085-1367

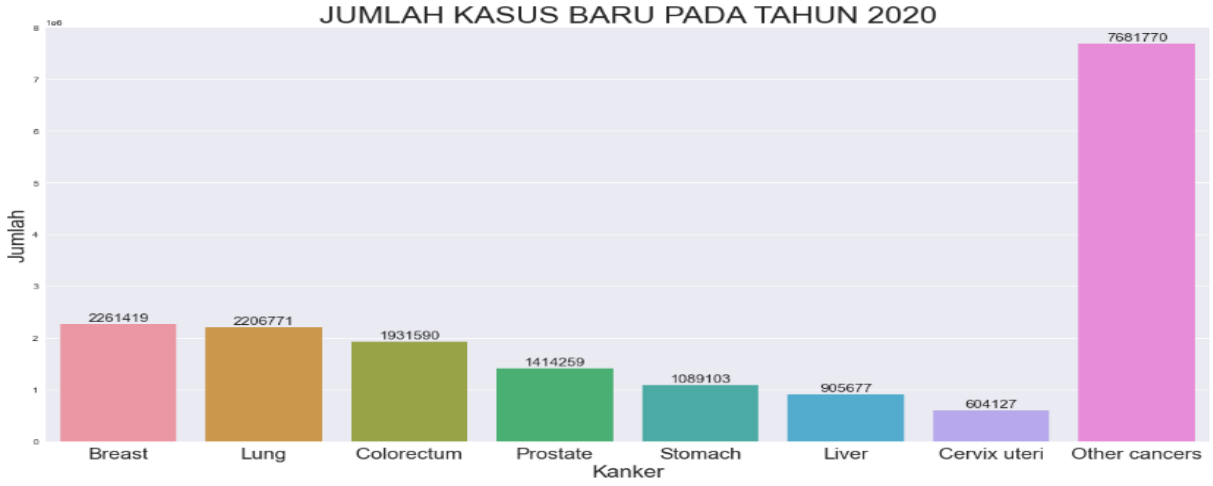
1. PENDAHULUAN

Kanker payudara merupakan tipe kanker yang umumya terbentuk di sel-sel payudara dan sel-sel kanker tersebut tumbuh diluar kendali. Kanker payudara dapat terjadi pada semua gender, tetapi kanker ini umumnya lebih sering terjadi pada wanita. Di Australia kanker payudara merupakan tipe kanker yang sangat umum terjadi pada wanita dan kanker paling umum kedua yang menjadi penyebab kematian pada wanita setelah kanker paru-paru [1]. Sedangkan di Indonesia sendiri, jumlah kanker payudara ini sampai menempati urutan pertama dibandingkan jenis kanker yang lain dan menjadi salah satu penyumbang kematian pertama [2]. Pada tahun 2020 Global Cancer Observatory (GLOBOCAN) mencatat angka kematian pada kanker payudara sampai menembus 684.996 jiwa dari total 9.958.133 kematian di seluruh dunia terhadap kanker.



Gambar 1. Grafik Kematian Kanker Tahun 2020

Gambar 1 merupakan grafik jumlah kasus kematian berdasarkan jenis kanker di tahun 2020. Dapat dilihat bahwa pada tahun 2020 kanker payudara memiliki angka kematian yang tidak sedikit, yaitu 684.996 jiwa, jumlah kasus kematian ini lebih banyak dibandingkan jenis kanker oesophagus dan pankreas.



Gambar 2. Grafik Kasus Baru Kanker Tahun 2020

e-ISSN: 2460-870X | p-ISSN: 2085-1367

Selain itu, pada tahun yang sama GLOBOCAN juga mencatat kasus baru sebanyak 2.261.419 jiwa[3], Gambar 2 merupakan grafik dari jumlah kasus baru kanker payudara yang muncul di tahun 2020. Jumlah kasus baru ini lebih banyak dibandingkan 6 jenis kanker lainnya.

Dari angka kematian tersebut, seharusnya baik pria maupun wanita sadar dengan kesehatan mereka dengan cara melakukan tindakan seperti deteksi dini dan menghindari risiko yang menyebabkan terjadinya kanker. Permasalahan dalam penelitian ini yaitu bagaimana membuat (*Artificial Intelligence*) AI yang dapat mendeteksi kanker payudara secara dini dan menganalisis perbandingan performa dari beberapa algoritma data mining dalam melakukan klasifikasi kanker payudara. Deteksi dini dapat dilakukan menggunakan algoritma data mining, dengan menggabungkan algoritma data mining dengan teknologi kecerdasan buatan, maka akan lebih mudah mendeteksi kanker payudara yang diderita secara dini, hal ini dapat diwujudkan dengan cara memanfaatkan algoritma dari data mining dengan membuat sebuah model *machine learning*, dimana membuat sebuah mesin itu belajar dengan menggunakan algoritma-algoritma data mining yang ada.

Machine learning merupakan bagaimana sebuah mesin bisa menjadi pintar dengan belajar dari data-data yang telah ada sebelumnya dan dari hasil belajar tersebut dapat dilakukan regresi, klasifikasi, dan klustering. Penelitian mengenai *machine learning* dapat menggunakan algoritma *supervised learning* maupun *unsupervised learning*. Pada penelitian ini, penulis menggunakan algoritma *supervised learning* untuk mendeteksi kanker payudara, algortima yang digunakan adalah Decision Tree, Naive Bayes dan K-Nearest Neighbors, kelebihan dari ketiga algoritma tersebut yaitu selain pada penelitian sebelumnya akurasi yang dihasilkan sangat baik, algoritma tersebut juga memiliki komputasi yang murah jika ingin membuat model *machine learning*. Selain itu penulis menggunakan bahasa pemrograman python untuk membuat model *machine learning* dan melakukan komparasi, untuk IDE sendiri penulis menggunakan jupyter notebook.

Penelitian menggunakan teknik *machine learning* maupun hanya menggunakan algoritma data mining telah banyak dilakukan dan menjadi metode yang sering digunakan dalam hal regresi, klasifikasi maupun klustering, seperti penelitian yang dilakukan oleh Alireza Osareh dan Bita Shadgar dengan menggunakan algoritma SVM, K-Nearest Neighbors dan Probabilitas Neural Networks tentang prediksi kanker payudara menggunakan 2 dataset menghasilkan akurasi terbaik pada SVM dengan akurasi pada masing-masing dataset sebesar 98.80% dan 96.33%[4]. Pada penelitian sebelumnya yang dilakukan oleh Wahyu Ananda, M Fauzan dan M Safii menggunakan algoritma Naive Bayes tentang prediksi jumlah panen kelapa sawit menghasilkan tingkat akurasi sebesar 100% dengan menggunakan bantuan tools rapid miner dan didapatkan akurasi yang sama dengan menggunakan perhitungan manual[5]. Penelitian sebelumnya juga yang dilakukan oleh Anik Andriani tentang prediksi penyakit diabetes menggunakan algoritma Decision Tree menghasilkan akurasi untuk data training sebesar 87.86% dan akurasi untuk data testing sebesar 84.29%[6]. Pada penelitian sebelumnya yang dilakukan oleh Yogiek Indra Kurniawan dan Tiyssa Indah Barokah tentang klasifikasi pengajuan kartu kredit dengan menggunakan algoritma K-Nearest Neighbors menghasilkan akurasi sebesar 93%[7]. Penelitian sebelumnya juga dilakukan oleh Wilem Musu menggunakan algoritma Decision Tree dalam menguji akurasi dari setiap proporsi pembagian data, hasil dari pengujian membuktikan bahwa proporsi pembagian data dapat mempengaruhi akurasi dari model [8]. Pada penelitian sebelumya yang dilakukan oleh M.Faizal Kurniawan dan Ivandari menggunakan tiga algoritma dan menggunakan bantuan tools rapid miner menghasilkan algoritma Naive Bayes mendapatkan akurasi yang paling tinggi dibandingkan dengan algoritma K-Nearest Neighbors dan Decision Tree C4.5 yaitu 95.85%.[9].

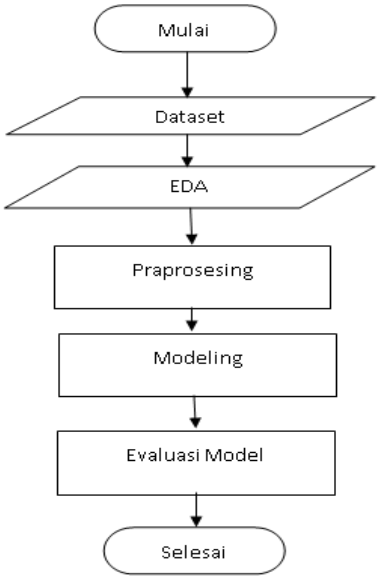
Komparasi algoritma data mining pada penelitian ini memiliki beberapa tujuan seperti membandingkan algoritma mana yang memiliki tingkat akurasi yang paling baik dalam klasifikasi kanker payudara, hal ini dapat membantu para statistikawan agar mengetahui metode klasifikasi mana yang memiliki nilai akurasi tinggi, membantu programmer dalam membuat *Artificial Intelligence* yang

e-ISSN: 2460-870X | p-ISSN: 2085-1367

dapat mendeteksi masalah-masalah dalam bidang medis seperti kanker payudara dengan menggunakan data gambar digital FNA, serta membantu meningkatkan kualitas pelayanan medis dalam mendeteksi kanker, khususnya kanker payudara.

2. METODE DAN DATA

A. Flowchart Penelitian



Gambar 3. Flowchart Penelitian

B. Data

Dataset diperoleh dari *University of California, Irvine Machine Learning Repository* yang merupakan kumpulan himpunan data, teori domain, dan generator data yang dipakai dalam analisis algoritma *machine learning* oleh komunitas *machine learning*[10]. Nama datasetnya sendiri yaitu "*Breast Cancer Wisconsin (Diagnostic)"*.

Feature-Feature diperoleh dari gambar digital *Fire Needle Aspirate* (FNA) dari tiap pasien yang menjadi acuan dalam menentukan tipe kanker. Setiap *features* dibagi mejadi 3 kategori (kecuali id dan diagnosis) yaitu *mean*, *worst* dan *standard error*, jadi total keseluruhan *features* adalah 32 sedangkan jumlah keseluruhan datanya adalah 569. Tabel 1 merupakan daftar *features* yang terdapat dalam dataset.

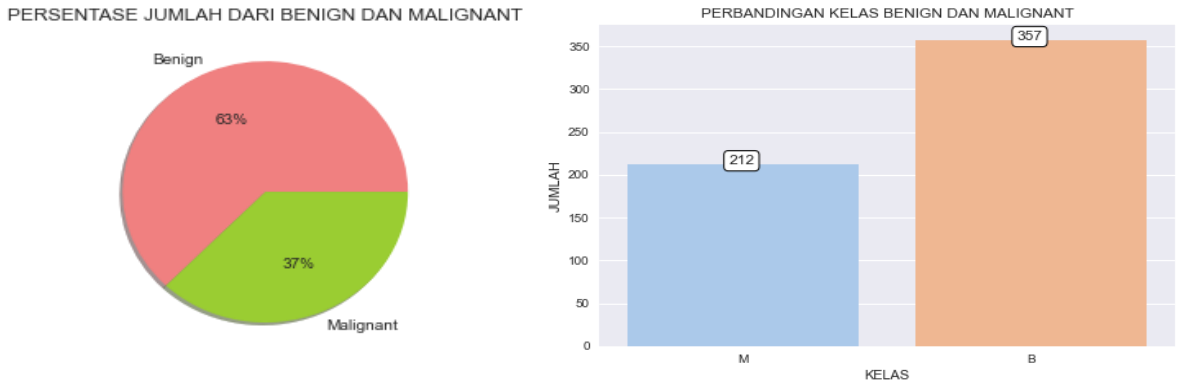
e-ISSN: 2460-870X | p-ISSN: 2085-1367

Tabel 1. Daftar Features

Nama Feature	Deskripsi
Id	Nomor identifikasi pemeriksaan
Diagnosis	M = Malignant (Ganas), B = Benign (Jinak)
Radius	Rerata jarak dari pusat ke titik pada keliling
Texture	Standard deviasi nilai gray-scale
Perimeter	Keliling
Area	Luas
Smoothness	Variasi lokal dalam panjang radius
Compactness	Keliling^2 / Luas – 1.0
Concavity	Keparahan bagian cekung dari kontur
Concave points	Jumlah bagian cekung dari kontur
Symmetry	Simetri
Fractal Dimension	Dimensi Fraktal

C. Exploratory Data Analysis

Proses menemukan pengetahuan dari data merupakan definisi dari analisis data, sebelum menerapkan beberapa model *machine learning* ke dalam dataset perlu untuk memahami masalah dari data, menangani data yang hilang, menghapus data yang *duplicate*, melakukan visualisasi data, dan memilih model *machine learning* [11]. Pada tahap EDA ini hal yang juga dapat dilakukan adalah melihat apakah tipe data pada setiap *featuresnya* sudah sesuai atau tidak dan memahami gambaran umum dari sebuah dataset dengan melakukan visualisasi.



Gambar 4. Grafik Jumlah Benign dan Malignant

Gambar 4 merupakan gambaran umum dari dataset. Dari grafik visualisasi yang ditampilkan, dapat dilihat persentase dari 2 kelas yaitu benign dan malignant cukup jauh dimana benign sebanyak 63% dan malignant sebanyak 37%. Jadi dapat dikatakan bahwa dataset yang digunakan memiliki kelas yang tidak seimbang.

D. Praprosesing

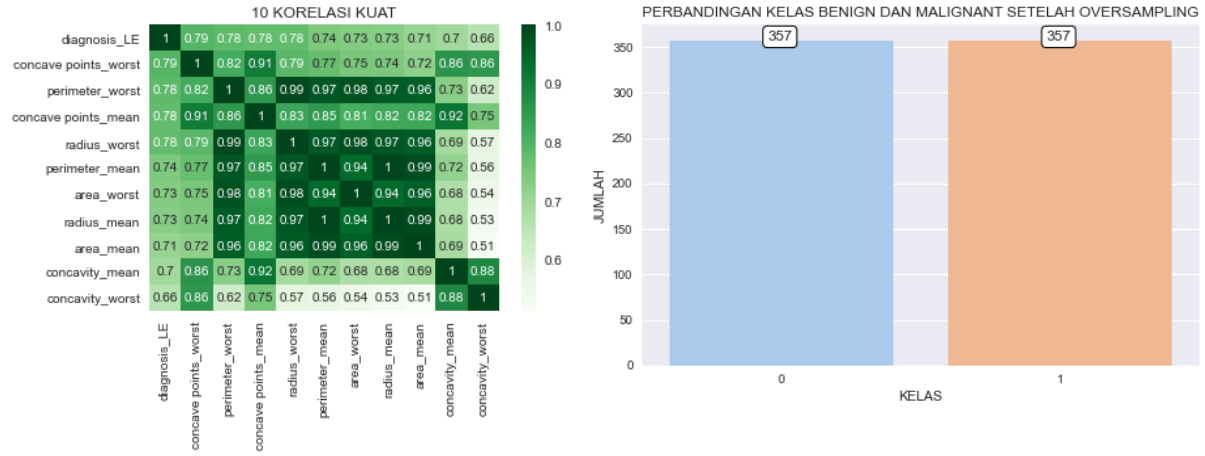
Pada tahap praprosesing data, hal yang dilakukan adalah mempersiapkan data sebaik mungkin sebelum dipelajari oleh mesin. Pada tahap ini jika terdapat *missing value* hal yang dapat dilakukan yaitu menghapus data tersebut atau mengisi dengan nilai baru seperti nilai *mean*, *median* atau *mode*. Selain itu dapat dilakukan transformasi data yang bersifat kategorik, transformasi data kategorik dapat dilakukan dengan 2 cara, yaitu menggunakan *Label Encoder* atau *One-Hot Encoder*. Selain mentransformasi data kategorik, hal yang juga dapat dilakukan adalah melakukan transformasi pada data numerik

DOI: https://www.doi.org/10.22303/csrid.14.3.2022.258-270

e-ISSN: 2460-870X | p-ISSN: 2085-1367

menggunakan *Standard Scaler* atau *Min Max Scaler*. Selain itu hal yang dapat dilakukan adalah menangani data yang tidak seimbang. Yang terakhir melakukan *feature selection*, yaitu menghapus *feature* yang tidak penting dan memilih *feature* yang dapat meningkatkan akurasi dari model, salah satu cara dalam memilih *feature* yaitu menggunakan metode *pearson correlation*.

Pada dataset *Breast Cancer Wisconsin (Diagnostic)* ini tidak terdapat *missing value* dan tipe data pada setiap *features* sudah sesuai. Pada tahap praprosesing penulis melakukan transformasi data kategorik pada *feature* diagnosis dengan menggunakan metode *Label Encoder* dan transformasi data numerik menggunakan *Standard Scaler*, Penulis juga melakukan *oversampling* data menggunakan *SMOTE*. *Feature* yang dipakai penulis pada penelitian ini hanya 10, karena 10 *features* tersebut memiliki korelasi kuat dengan target.



Gambar 5. Grafik Korelasi dan Jumlah Kelas Setelah Oversampling

Gambar 5 merupakan grafik korelasi dan jumlah kelas setelah dilakukan teknik oversampling. Dari gambar yang ditampilkan dapat dilihat grafik setelah dilakukan teknik *oversampling* pada dataset yang sebelumnya memiliki kelas yang tidak seimbang dan pada gambar disebelah kiri dapat dilihat 10 daftar *features* yang memiliki korelasi kuat dengan target.

E. Modeling
F.

1) Data Mining
Data mining adalah proses penerapan teknologi, metodologi statistika dan matematika untuk memilah-milah data untuk menemukan hubungan, pola, dan tren baru yang signifikan. [12]. Data mining juga dapat didefinisikan sebagai proses eksplorasi dan analisis dengan cara otomatis atau semi-otomatis dari sejumlah besar data untuk menemukan pola dan aturan yang bermakna[13]. Dalam hal ini data mining disimpulkan sebagai proses menemukan sebuah pola yang terdapat dalam dataset dengan menggunakan teknik statistik maupun matematis.

2) Machine Learning
Machine learning merupakan bagaimana sebuah mesin dapat menjadi pintar dengan cara belajar dari data-data yang telah ada sebelumnya dan dari hasil belajar tersebut dapat dilakukan regresi, klasifikasi, klustering dan lain sebagainya. Proses belajar tersebut dapat menggunakan algoritma data mining seperti Decision Tree, Naive Bayes ataupun KNN. Dengan mengembangkan mesin komputer yang belajar dari data yang ada dan membuat keputusan tanpa harus dilatih lagi menggunakan *machine learning*, salah satu bidang yang berada di bawah payung *Artificial Intelligence* (AI) adalah *machine learning*[14].

e-ISSN: 2460-870X | p-ISSN: 2085-1367

3) Algoritma Decision Tree

Decision Tree adalah algoritma yang digunakan dalam membuat model keputusan menggunakan struktur pohon atau struktur yang hirarki[15]. Pohon pada decision tree memiliki *root node* dan *node*, *root node* merupakan puncak dari pohon sedangkan *node* merupakan percabangan dari *root node* itu sendiri. Pada setiap *node* decision tree terdapat proses pembuatan keputusan yang menghasilkan dua cabang yaitu "ya" atau "tidak", pembuatan keputusan sendiri dilakukan dengan menguji suatu variabel, proses pengujian ini terus berlanjut hingga *node* paling bawah atau disebut dengan *leaf node*[14]. Dalam proses decision tree hal yang paling pertama dilakukan adalah memilih *root node*, salah satu cara dalam pemilihan *root node* ini dapat dilakukan dengan menghitung nilai *gain* pada setiap atribut, sebelum melakukan perhitungan pada nilai *gain* perlu dilakukan perhitungan pada nilai *entropy* terlebih dahulu. Berikut formula *entropy* pada algoritma decision tree :

Formula Entropy[14]:

$$Entropy(S) = -\sum_{i=1}^{n} P_i * \log_2(P_i) \quad (1)$$

Keterangan :
n = Jumlah kelas S
P = Proporsi nilai-nilai masuk ke dalam kelas di tingat i

4) Algoritma Naive Bayes

Naive Bayes merupakan algoritma yang memakai prinsip probabilitas dalam menciptakan model prediksi klasifikasi[14]. Naive Bayes merupakan salah satu algoritma yang metode pembelajarannya bersifat *supervised*, dimana pada saat proses pembelajaran dibutuhkan data latih untuk dapat mengambil keputusan. Nilai probabilitas dari setiap kelas target yang ada akan dihitung terhadap input yang diberikan pada tahap klasifikasi. Kelas target yang memiliki probabilitas paling besarlah yang menjadi kelas pada data inputan tersebut[16]. Keunggulan dari Naive Bayes adalah cepat dan efektif dalam mengolah data dalam jumlah besar[14]. Dibawah ini merupakan formula persamaan Naive Bayes:

Formula Naive Bayes[14] :

$$P(J \mid K) = \frac{P(K \mid J)P(J)}{P(K)} = \frac{P(J \cap K)}{P(K)} \quad (2)$$

Keterangan :
P(J|K) = Probabilitas J terjadi bila K terjadi
P(K|J) = Probabilitas K terjadi bila J terjadi
P(J) = Probabilitas J terjadi
P(K) = Probabilitas K terjadi
P(J∩K) = Probabilitas P(J) dan P(K) terjadi secara bersamaan

5) Algoritma K-Nearest Neighbors

KNN adalah algoritma yang sifat pembelajarannya *semi-supervised* dimana membutuhkan data *training* dan nilai k yang telah ditentukan sebelumnya [17]. Dalam KNN, k merupakan jumlah tetangga yang diambil dalam membuat sebuah keputusan[9]. Algoritma KNN memiliki prinsip kerja yaitu mencari jarak tetangga terdekat antara data yang akan dievaluasi dengan data latih[18]. Berikut merupakan persamaan dalam menentukan jarak *Euclidean* pada K-Nearest Neighbors:

e-ISSN: 2460-870X | p-ISSN: 2085-1367

Formula Euclidean[19]:

$$Euclidean = \sqrt{\sum_{i=1}^{p} (X_{2i} - X_{1i})^2} \quad (3)$$

Keterangan :
p = dimensi data
X1 = Data train
X2 = Data test

G. Evaluasi Model

Metode yang umumnya dipakai dalam mengukur performa suatu model adalah confusion matrix, confusion matrix menghasilkan *binary classification* yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN)[14]. Hasil dari confusion matrix ini nantinya dapat dijadikan sebagai acuan dalam menentukan akurasi dari model.

Tabel 2. Confusion Matrix

		Prediksi	
		Negative	**Positive**
Kenyataan	**Negative**	TN	FP
	Positive	FN	TP

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

3. HASIL DAN PEMBAHASAN

Klasifikasi kanker payudara dalam penelitian ini menggunakan algoritma Decision Tree, Naive Bayes dan K-Nearest Neighbors dengan jumlah K=5. Pada penelitian ini penulis menggunakan bahasa pemrograman python dan menggunakan IDE jupyter notebook dalam melakukan penelitian. Pada metode Hold-Out penulis melakukan 9 skenario dalam pembagian data yang nantinya hasil akurasi terbaik dari setiap algoritma tersebut dibandingkan. selain itu karena pada dasarnya metode HoldOut data yang diambil pada saat pembagian itu acak, penulis menggunakan parameter *random state* pada *scikit learn* agar proses pengacakan datanya konsisten. Sedangkan pada metode K-Fold penulis menggunakan K sebanyak 5.

A. Hold-Out Cross Validation

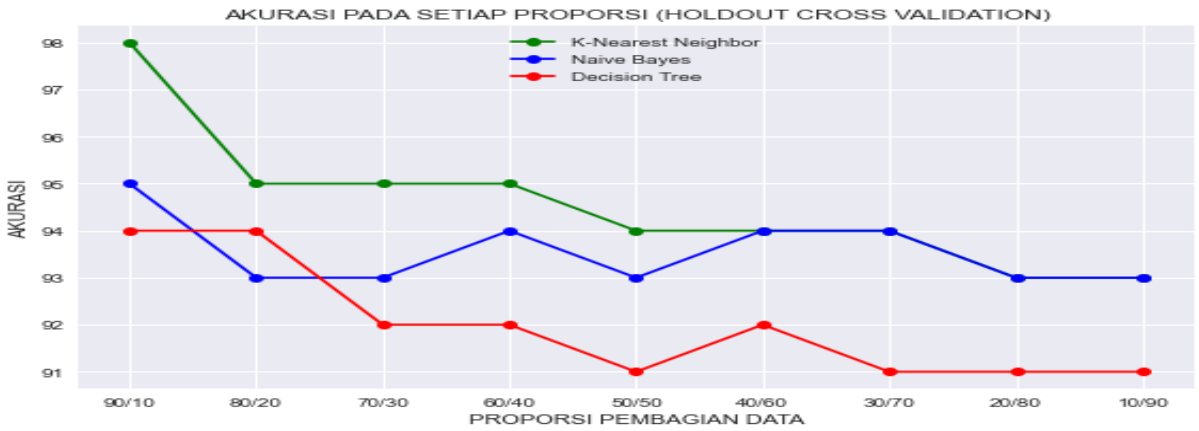
Tabel 3. Akurasi dan Proporsi Data pada Setiap Skenario

Skenario	**Proporsi**			**Akurasi**		
	Train	**Test**	**Random State**	**Decision Tree**	**Naive Bayes**	**KNN**
1	90%	10%	0	94.44%	95.83%	98.61%
2	80%	20%	0	94.41%	93.71%	95.10%
3	70%	30%	0	92.56%	93.95%	95.81%

e-ISSN: 2460-870X | p-ISSN: 2085-1367

4	60%	40%	0	92.66%	94.41%	95.80%
5	50%	50%	0	91.88%	93.84%	94.68%
6	40%	60%	0	92.07%	94.41%	94.64%
7	30%	70%	0	91.40%	94.20%	94.40%
8	20%	80%	0	91.08%	93.88%	93.88%
9	10%	90%	0	91.29%	93.93%	93.78%

Tabel 3 merupakan tabel akurasi dan pembagian proporsi data pada setiap skenario yang dilakukan. Dapat dilihat bahwa pembagian proporsi data dapat mempengaruhi akurasi pada setiap algoritma, Algoritma K-Nearest Neighbors selalu memiliki akurasi yang paling baik dibanding algoritma Decision Tree dan Naive bayes dimana akurasi yang paling tinggi pada algortima K-Nearest Neigbors yaitu 98.61% dengan proporsi pembagian data 90/10 dan akurasi yang paling rendah berada pada 93.78% dengan proporsi pembagian data 10/90, sedangkan untuk algoritma Decision Tree akurasi yang paling tinggi mencapai 94.44% dengan proporsi 90/10 dan akurasi yang paling rendah yang didapatkan adalah 91.08% dengan proporsi 10/90, sedangkan untuk algoritma Naive Bayes akurasi yang paling tinggi mencapai 95.83% dengan proporsi 90/10 dan akurasi yang paling rendah yaitu 93.71% dengan proporsi 80/20.



Gambar 6. Grafik akurasi metode hold-out

Dari percobaan 9 skenario dengan menggunakan metode Hold-Out, dapat dilihat pada Gambar 6 grafik dari ketiga algoritma memiliki akurasi terbaik pada skenario ke-1 dengan proporsi 90/10 (90% data *training* dan 10% data *testing*). Pada metode Hold-Out ini, algoritma K-Nearest Neighbors memiliki akurasi yang paling baik dibandingkan kedua algoritma lainnya, dengan akurasi terbaik yang didapatkan algoritma K-Nearest Neighbors sebesar 98%, Naive Bayes sebesar 95% dan Decision Tree sebesar 94%.

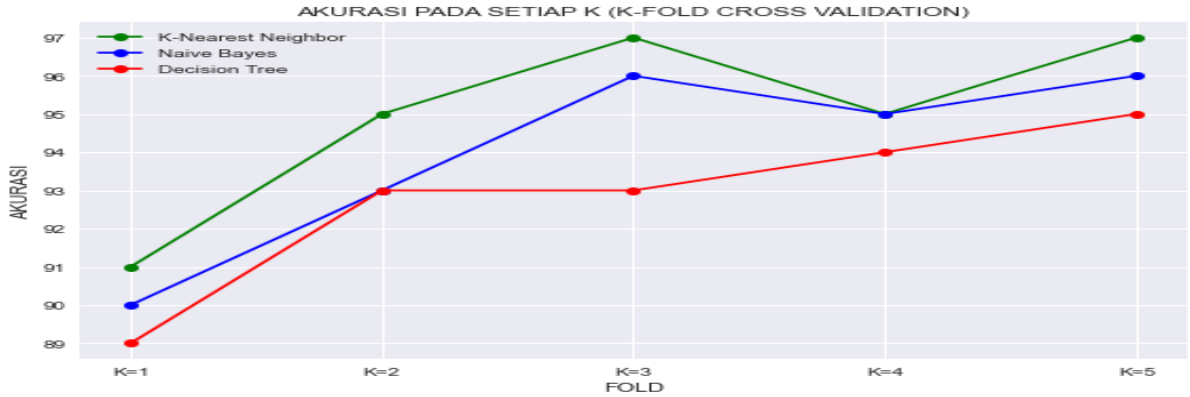
B. K-Fold Cross Validation

Tabel 4. Akurasi K-Fold

Algoritma	**Akurasi**
Decision Tree	93%
Naive Bayes	95%
KNN	96%

e-ISSN: 2460-870X | p-ISSN: 2085-1367

Pada metode K-Fold, algoritma K-Nearest Neighbors memiliki nilai rata-rata akurasi yang paling baik dibanding algoritma Naive Bayes dan Decision Tree dimana K-Nearest Neighbors memiliki akurasi sebesar 96%, Naive Bayes sebesar 95% dan Decision Tree sebesar 93%. Meskipun ada perbedaan akurasi dari ketiga algoritma yang digunakan, tetapi perbedaan akurasi yang dihasilkan dari algoritma Decision Tree, Naive Bayes dan K-Nearest Neighbors tidak terlalu signifikan.

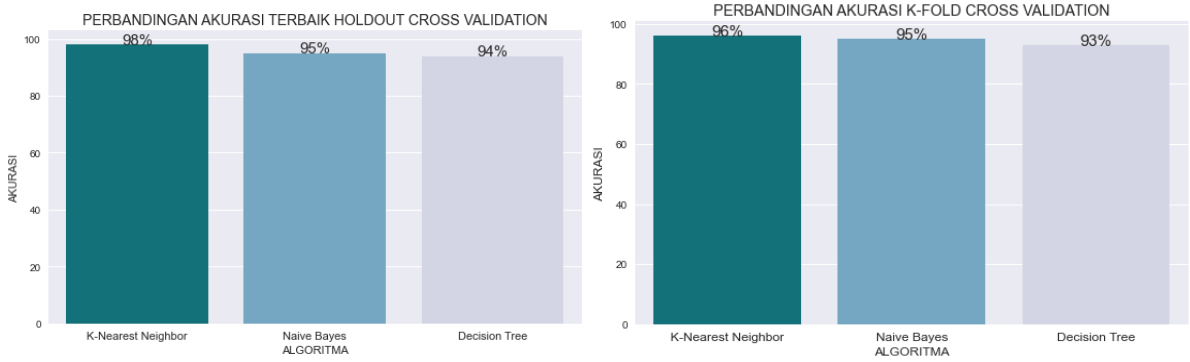


Gambar 7. Grafik Akurasi Setiap K Metode K-Fold

Berdasarkan Gambar 7, ketiga algoritma memiliki akurasi paling besar pada K=5 dan paling rendah pada K=1. Akurasi maksimum yang didapatkan pada algoritma K-Nearest Neighbors yaitu sebesar 97%, Naive Bayes sebesar 96% dan Decision Tree sebesar 95%, sedangkan akurasi minimum yang didapatkan pada algoritma K-Nearest Neighbors yaitu sebesar 91%, Naive Bayes sebesar 90% dan Decision Tree sebesar 89%.

C. Rangkuman Akurasi

Dari beberapa percobaan yang dilakukan, didapatkan hasil rangkuman akurasi dari metode Hold-Out dan K-Fold.



Gambar 8. Komparasi Akurasi dari Metode Hold-Out dan K-Fold

Gambar 8 merupakan rangkuman akurasi dari percobaan yang dilakukan pada ketiga algoritma. Dari percobaan yang dilakukan menggunakan metode Hold-Out dan K-Fold Cross Validation, algortima K-Nearest Neighbors selalu memiliki performa akurasi yang sangat baik dibanding algoritma Naive Bayes dan Decision Tree, dimana akurasi yang didapatkan algoritma K-Nearest Neighbors pada metode Hold-Out sebesar 98% dan metode K-Fold sebesar 96%, sedangkan akurasi yang didapatkan algoritma Naive Bayes pada metode Hold-Out sebesar 95% dan pada metode K-Fold sebesar 95%, sedangkan akurasi yang didapatkan algoritma Decision Tree pada metode Hold-Out sebesar 94% dan pada metode K-Fold sebesar 93%.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

Tabel 5. Confusion Matrix Skenario Terbaik Algoritma Decision Tree Metode Hold-Out

	Negative	**Positive**
Negative	35	2
Positive	2	33

Tabel 6. Confusion Matrix Skenario Terbaik Algoritma Naive Bayes Metode Hold-Out

	Negative	**Positive**
Negative	36	1
Positive	2	33

Tabel 7. Confusion Matrix Skenario Terbaik Algoritma KNN Metode Hold-Out

	Negative	**Positive**
Negative	37	0
Positive	1	34

Tabel 5, 6, dan 7 merupakan confusion matrix dari ketiga algoritma pada metode Hold-Out Cross Validation. Dikarenakan kasus yang dihadapi merupakan kasus yang berhubungan dengan medis atau berkaitan dengan nyawa, maka dari itu penulis lebih fokus membahas confusion matrix yang memiliki False Negative seminimum mungkin, dalam hal ini algoritma K-Nearest Neighbors memiliki False Negative yang paling sedikit yaitu hanya 1, sedangkan algoritma Decision Tree dan Naive Bayes masing-masing memiliki False Negative sebanyak 2, jadi dapat disimpulkan bahwa pada confusion matrix dari ketiga algoritma, algoritma K-Nearest Neighbors memiliki performa yang sangat baik dalam melakukan klasifikasi dibandingkan agoritma Decision Tree dan Naive Bayes.

4. KESIMPULAN

Berdasarkan pengujian yang telah penulis lakukan didapatkan beberapa kesimpulan.

- Pada metode Hold-Out ketiga algoritma memiliki akurasi tertinggi pada skenario ke-1 dengan proporsi data 90/10, algoritma K-Nearest Neighbors memiliki performa yang sangat baik dimana akurasi terbaik yang didapatkan sebesar 98%, sedangkan algoritma Naive Bayes akurasi terbaik yang didapatkan sebesar 95% dan Decision Tree akurasi terbaik yang didapatkan sebesar 94%.
- Pada metode K-Fold, algoritma K-Nearest Neighbors memiliki akurasi tertinggi dibanding algoritma Naive Bayes dan Decision Tree, dimana akurasi K-Nearest Neighbors memilki akurasi sebesar 96%, Naive Bayes sebesar 95%, dan Decision Tree sebesar 93%.
- Berdasarkan confusion matrix metode Hold-Out pada ketiga algoritma, algoritma K-Nearest Neighbors memiliki performa yang baik dalam melakukan klasifikasi dan memiliki False Negative yang paling rendah dibandingkan algortima Decision Tree dan Naive Bayes yaitu K-Nearest Neighbors sebanyak 1, Decision Tree sebanyak 2 dan Naive Bayes sebanyak 2.

DOI: https://www.doi.org/10.22303/csrid.14.3.2022.258-270

- Dari 2 metode Cross Validation yang telah dilakukan yaitu Hold-Out dan K-Fold, K-Nearest Neighbors selalu memiliki performa yang sangat baik dibanding algoritma Naive Bayes dan Decision Tree.

Pada penelitian ini penulis hanya menggunakan 3 algoritma dalam melakukan perbandingan performa klasifikasi kanker payudara dan hanya menggunakan dataset yang berasal dari 1 sumber, maka dari itu saran dalam penelitian selanjutnya yaitu dapat menguji coba menggunakan dataset kanker payudara yang lain dengan menggunakan metode yang sama serta melakukan komparasi dengan menambahkan algoritma yang berbeda seperti algoritma jaringan syaraf tiruan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih ke Universitas Dipa Makassar yang telah membantu untuk memafisilatasi tulisan ini untuk terbit ke CSRID. Terima kasih juga kepada para pembimbing yang telah membimbing kami sehingga selesai penelitian ini.

REFERENSI

[1] C. Council, “Breast cancer | Causes, Symptoms & Treatments | Cancer Council,” 2020. https://www.cancer.org.au/cancer-information/types-of-cancer/breast-cancer (accessed May 10, 2022).

[2] K. K. R. Indonesia, “Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan,” 2022. https://www.kemkes.go.id/article/view/22020400002/kanker-payudaya-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan.html (accessed May 10, 2022).

[3] G. C. Observatory, “Cancer Today,” 2020. https://gco.iarc.fr/ (accessed May 11, 2022).

[4] A. Osareh and B. Shadgar, “Machine learning techniques to diagnose breast cancer,” in *2010 5th International Symposium on Health Informatics and Bioinformatics*, 2010, pp. 114–120. doi: 10.1109/HIBIT.2010.5478895.

[5] W. Ananda, M. Safii, and M. Fauzan, “Prediksi Jumlah Hasil Panen Sawit Menggunakan Algoritma Naive Bayes,” *TIN Terap. Inform. Nusant. Vol*, vol. 1, no. 10, pp. 513–519, 2021.

[6] A. Andriani, “Sistem prediksi penyakit diabetes berbasis decision tree,” *J. Bianglala Inform.*, vol. I, no. 1, pp. 1–10, 2013.

[7] Y. I. Kurniawan and T. I. Barokah, “Klasifikasi Penentuan Pengajuan Kartu Kredit Menggunakan K-Nearest Neighbor,” *J. Ilm. Matrik*, vol. 22, no. 1, pp. 73–82, 2020, doi: 10.33557/jurnalmatrik.v22i1.843.

[8] W. Musu, A. Ibrahim, and Heriadi, “Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4 . 5,” *Pros. Semin. Ilm. Sist. Inf. Dan Teknol. Inf.*, vol. X, no. 1, pp. 186–195, 2021.

[9] F. Kurniawan and Ivandari, “Komparasi Algoritma Data Mining Untuk Klasifikasi Penyakit Kanker Payudara,” *IC-Tech*, vol. XII, no. 1, pp. 1–8, 2017, [Online]. Available: http://jurnal.stmik-wp.ac.id

[10] Arthur Asuncion and D. Newman, “About,” 2017. https://archive.ics.uci.edu/ml/about.html (accessed May 14, 2022).

[11] G. R. Shinde, S. Majumder, H. R. Bhapkar, and P. N. Mahalle, “Exploratory Data Analysis BT - Quality of Work-Life During Pandemic: Data Analysis and Mathematical Modeling,” G. R. Shinde, S. Majumder, H. R. Bhapkar, and P. N. Mahalle, Eds. Singapore: Springer Singapore, 2022, pp. 97–105. doi: 10.1007/978-981-16-7523-2_7.

[12] G. A. Marcoulides, *Discovering Knowledge in Data: an Introduction to Data Mining*, vol. 100, no. 472.

e-ISSN: 2460-870X | p-ISSN: 2085-1367

2005. doi: 10.1198/jasa.2005.s61.

[13] M. A. Berry and G. Linoff, “Data mining techniques - for marketing, sales, and customer support,” 1997.

[14] D. M. S. Kurniawan, *Pengenalan Machine Learning Python*. Jakarta: PT ELEX MEDIA KOMPUTINDO, 2020.

[15] N. Jayanti, S. Puspitodjati, and T. Elida, “Teknik Klasifikasi Pohon Keputusan Untuk Memprediksi Kebangkrutan Bank Berdasarkan Rasio Keuangan Bank,” *Proceeding, Semin. Ilm. Nas. Komput. dan Sist. Intelijen (KOMMIT 2008) ISSN 1411-6286*, no. Kommit, pp. 101–107, 2008.

[16] D. Sartika, D. I. Sensuse, U. Indo, G. Mandiri, and F. I. Komputer, “Perbandingan Algoritma Klasifikasi Naive Bayes , Nearest Neighbour , dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian,” vol. 1, no. 2, pp. 151–161, 2017.

[17] K. Chomboon, P. Chujai, P. Teerarassammee, K. Kerdprasop, and N. Kerdprasop, “An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm,” pp. 280–285, 2015, doi: 10.12792/iciae2015.051.

[18] T. Rismawan, A. W. Irawan, W. Prabowo, and S. Kusumadewi, “Sistem Pendukung Keputusan Berbasis Pocket Pc Sebagai Penentu Status Gizi Menggunakan Metode Knn (K-Nearest Neighbor),” *Teknoin*, vol. 13, no. 2, pp. 18–23, 2008, doi: 10.20885/teknoin.vol13.iss2.art5.

[19] I. A. Nikmatun and I. Waspada, “Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor,” *J. SIMETRIS*, vol. 10, no. 2, pp. 421–432, 2019.