Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 233 – 237
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

Perbandingan Algoritma *K-Nearest Neighbor (Knn)* Dan *Gaussian Naive Bayes (Gnb)* Dalam Klasifikasi *Breast Cancer Coimbra*

Comparison Between K-Nearest Neighbor (Knn) And Gaussian Naive Bayes (Gnb) Algorithm In The Coimbra Breast Cancer Classification

Johan Taruna Wijaya[1], Hardian Oktavianto[2*], Habibatul Azizah Al Faruq[3]
[1]Mahasiswa Fakultas Teknik, Universitas Muhammadiyah Jember
email: johantaruna007@gmail.com
[2,] Dosen Fakultas Teknik, Universitas Muhammadiyah Jember *Koresponden Author
email: hardian@unmuhjember.ac.id[2]
[3]Dosen Fakultas Teknik, Universitas Muhammadiyah Jember
email: habibatulazizah@unmuhjember.ac.id[3,]

ABSTRAK

Kanker payudara didefinisikan sebagai suatu penyakit *neoplasma* ganas yang berasal dari *parenchyma* dan menghasilkan frekuensi kematian yang menjadi penyebab utama kekhawatiran di dunia. Kanker payudara merupakan kanker kedua yang paling banyak diderita dan penyebab kelima kematian kanker di seluruh dunia dengan presentase 6,4% dari semua penyebab kematian. Pada penelitian ini dilakukan klasifikasi terhadap data kanker payudara, dimana data tersebut terdapat adalah data darah pengidap kanker payudara. Metode yang digunakan pada klasifikasi ini adalah *K-Nearest Neighbor* (KNN) dan *Gaussian Naive Bayes* (GNB). Pengujian akurasi pada penelitian ini menggunakan *Cross Validation* dan evaluasi data ujia dengan *Confusion Matrix*. Dari penelitian ini didapatkan hasil pada 116 data darah kanker payudara, metode KNN menghasilkan akurasi 86,9% lebik baik dari pada GNB, dan untuk presisi dan recall, metode KNN menghasilkan presisi sebesar 87,3%, dan *recall* sebesar 86,7%, pengujian pada metode KNN menggunakan nilai K=4.

Kata Kunci : Klasifikasi data, Kanker Payudara, *K-Nearest Neighbor, Gaussian Naive Bayes*

ABSTRACT

Breast cancer is defined as a malignant neoplasm that originates from the parenchyma and produces a death frequency which is a major cause of concern in the world. Breast cancer is the second most common cancer and the fifth leading cause of cancer deaths worldwide, accounting for 6.4% of all causes of death. In this study, the classification of breast cancer data was carried out, where the data contained was blood data for people with breast cancer. The methods used in this classification are K-Nearest Neighbor (KNN) and Gaussian Naive Bayes (GNB). Testing the accuracy in this study using Cross Validation and evaluation of the test data with the Confusion Matrix. From this study, it was found that on 116 breast cancer blood data, the KNN method produced an accuracy of 86.9% better than GNB, and for precision and recall, the KNN method produced a precision of 87.3%, and a recall of 86.7%, testing on the KNN method uses the value of K = 4.

Keywords : Data classification, Breast Cancer, K-Nearest Neighbor, Gaussian Naive Bayes

Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 233 – 237
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

1. PENDAHULUAN

Kanker payudara *(Breast Cancer)* didefinisikan sebagai suatu penyakit *neoplasma* ganas yang berasal dari *parenchyma* dan menghasilkan frekuensi kematian yang menjadi penyebab utama kekhawatiran di dunia. Menurut data statistik Globocan (2015), kanker payudara merupakan kanker kedua yang paling banyak diderita dan penyebab kelima kematian kanker di seluruh dunia dengan presentase 6,4% dari semua penyebab kematian. Pada tahun 2017, sekitar 252.710 wanita terkena diagnosa kanker payudara dan 40.610 diantaranya mengalami keadaan kritis dan hampir menyebabkan kematian (Harbeck dan Gnant, 2019). Kanker payudara terbagi menjadi jinak dan ganas, para ahli dan dokter akan memberikan perwatan yang berbeda pada setiap pasien. Kunci untuk bertahan hidup penderita kanker payudara adalah dengan mendeteksi kanker payudara sedini mungkin, sebelum sel kanker menyebar ke bagian-bagian tubuh lainnya.

Untuk mengetahui seseorang terkena penyakit kanker payudara atau tidak sangatlah susah, oleh karna itu, pemeriksaan penting dilakukan dengan cara cek darah guna memastikan apakah seseorang terkena kanker payudara atau tidak. Dari dataset konsultasi rutin dan analisis darah (Patricio, dkk., 2018), menurut Adi dan Sari (dalam Artha., dkk. 2019), dengan bantuan *data mining*, dataset tersebut bisa diklasifikasikan, kemudian hasilnya dapat membantu dokter untuk mendiagnosis pasien apakah positif kanker payudara atau negatif.

Seiring dengan perkembangan ilmu pengetahuan dan teknologi informasi, kehadiran *machine learning* di bidang komputer telah menarik banyak perhatian. *Machine learning* memainkan peran luas dalam pengembangan terutama dalam pengembangan data analitik (Alarifi dan Young., 2018). Ada salah satu metode yang ada pada *machine learning* yaitu klasifikasi. Untuk klasifikasi itu sendiri banyak jenis algoritma yang dapat digunakan seperti *K-Nearest Neighbor* dan *Naive Bayes*.

Pada penelitian sebelumnya yang dilakukan oleh (Eyupoglu, 2017) dengan judul "*Cancer Classification Using K-Nearest Neighbors Algorithm*", dengan menggunakan algoritma KNN atau bisa disebut dengan K-NEAREST NEIGHBORS menghasilkan akurasi 97% yang termasuk dalam *Good Classification.*

Pada penelitian yang dilakukan oleh (Kamel, dkk., 2019) dengan judul "Cancer Classification Using Gaussian Naive Bayes Algorithm", dengan menggunakan algoritma *Gaussian Naive Baye*s menghasilkan akurasi 98% yang termasuk dalam *Good Classification*. Dengan demikian bahwa metode algoritma *Gaussian Naive Bayes* ini akurat dalam melakukan klasifikasi kanker payudara.

Berdasarkan penelitian sebelumnya yang telah dilakukan oleh Kamel, dkk. (2019) dan Eyupoglu (2017) yang membuktikan bahwa KNN dan GNB memiliki tingkat akurasi tinggi. Dikarenakan masih belum ada penelitian yang membahas tentang perbandingan KNN dan GNB dengan menggunakan dataset *Breast Cancer Coimbra* maka dilakukanlah penelitian ini.

Berdasarkan uraian di atas, maka dalam penelitian ini akan dilakukan perbandingan antara metode KNN dan *GNB* untuk mengklasifikasikan Breast Cancer Coimbra dengan judul "Perbandingan Algoritma *K-Nearest Neighbor* (*K-NN*) dan *Gaussian Naive Bayes* (*GNB*) dalam Klasifikasi *Breast Cancer Coimbra*".

Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 233 – 237
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

2. TINJAUAN PUSTAKA

A. K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (K-NN**)** berfungsi dengan cara mencari jumlah k ditiap pola (diantara semua pola latih yang ada di semua kelas) yang terdekat pada pola masukan, kemudian menentukan kelas keputusan berdasarkan jumlah pola terbanyak di antara k pola tersebut (voting) (Suyanto, 2018). Dekat atau jauh lokasinya (jarak) dapat dihitung melalui salah satu dari besaran jarak yang telah ditentukan, Akan tetapi dalam penerapannya jarak *Euclidean* sangat sering digunakan dikarenakan memiliki tingkat akurasi dan juga *productivity* yang tinggi (Asiyah dan Fithriasari, 2016). Rumus jarak *Euclidean* ditunjukan oleh **Persamaan 1** berikut:

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^{p}(x_{ip} - x_{jp})^2} \quad (1)$$

d(x_i, x_j) merupakan jarak *euclidean* dari data uji dengan data latih sedangkan x_{ip} dan x_{jp} merupakan data *testing* ke-i dan data *training* ke-j.

B. Gaussian Naive Bayes

Saat menangani data kontinu, distribusi Gaussian digunakan untuk mendistribusikan nilai kontinu yang terkait dengan setiap kelas. Data pelatihan dibagi menjadi beberapa kelas, dan mean dan deviasi standar setiap kelas dapat dihitung, (Kamel, dkk, 2019). Oleh karena itu untuk memperkirakan probabilitas data kontinu ditetapkan persamaan sebagai berikut:

$$P(X = x|C = c) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Dimana :
x = variabel
c = kelas
μ = *mean*
σ = *deviasi stander*

C. Akurasi, Presisi, dan *Recall*

Menurut Suyanto (2018) Akurasi menyatakan persentase dari jumlah *tuple* dalam data uji yang diklasifikasikan dengan benar oleh model klasifikasi. Ia juga berpendapat bahwa presisi adalah ukuran kepastian, yaitu berapa persentase *tuple* yang dilabeli sebagai positif adalah benar pada kenyataanya, dan *recall* adalah ukuran kelengkapan yaitu berapa persentase *tuple* positif yang dilabeli sebagai positif. Rumus akurasi, presisi, dan *recall* ditunjukan pada **Tabel 1** berikut:

Tabel 1 Rumus Akurasi, Presisi, dan *Recall*

Akurasi	$\frac{TP + TN}{TP + TN + FP + FN}$
Presisi	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$

Sumber: Jurnal Puspitasari dkk 2018

3. METODE PENELITIAN

A. Pengumpulan data

Proses pengambilan data dilakukan secara manual pada UCI *Machine Learning Repository* tersedia di https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra diakses 18 April 2020, dengan mengunduh *file Excel*, dengan isi data yaitu 116 *record* dan terdapat dua kelas, kemudian *Excel* tersebut ditaruh di *folder* tertentu untuk memudahkan pemanggilan *file* data ke *Jupyter Notebook*.

B. Penerapan Algoritma

Pada proses ini algoritma yang digunakan yaitu *K-Nearest Neighbor* (*KNN*) dan *Gaussian Naive Bayes* (*GNB*). Untuk *KNN* sendiri data diuji untuk mendapatkan pemodelan dalam proses klasifikasi ke dalam 2 kelas dengan mencari jarak *Euclidean* pada data *training* dan data *testing*. Dan untuk *GNB* data diuji dengan

Jurnal Smart Teknologi
Vol. 3, No. 3, Maret 2022, Halaman 233 – 237
ISSN: 2774-1702, http://jurnal.unmuhjember.ac.id/index.php/JST

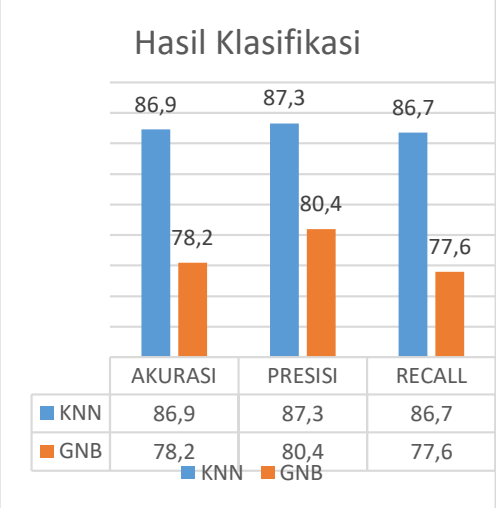
mencari nilai probabilitas dan nilai data continu.

C. Validasi dan Evaluasi

Untuk mendapatkan hasil akurasi yang optimal proses validasi menggunakan *K-Fold Cross Validation* dan penentuan nilai K yang digunakan pada penelitian ini yaitu 2,3,4, dan 5 karena untuk menghasilkan partisi data yang seimbang. Proses evaluasi dilakukan dengan menghitung nilai TP, FP, TN, FN sehingga menghasilkan nilai akurasi, *precision*, dan *recall*.

4. PEMBAHASAN DAN HASIL

pada algoritma *KNN* didapatkan hasil terbesar dengan akurasi sebesar 86,9%, presisi sebesar 87,3%, dan *recall* sebesar 86,7% pada percobaan ke 5. Pada algoritma *GNB* mendapatkan hasil terbesar dengan akurasi sebesar 78,2%, presisi sebesar 80,4%, dan *recall* sebesar 77,6% pada percobaan ke 5. Berikut merupakan gambar perbandingan kinerja kedua algoritma ditunjukan pada **Gambar 1.**



Gambar 1. Hasil perbandingan kinerja algoritma.
Sumber: Hasil Perhitungan

5. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat diambil kesimpulan sebagai berikut:

1. Hasil akurasi paling tinggi dalam klasifikasi *Breast Cancer* adalah algoritma KNN didapatkan hasil sebesar 86,9% pada *Kfold* 5, dan akurasi paling tinggi algoritma GNB yaitu sebesar 78,2% pada *Kfold* 5.
2. Hasil presisi paling tinggi dalam klasifikasi *Breast Cancer* adalah algoritma KNN didapatkan hasil sebesar 87,3% pada *Kfold* 5, dan akurasi paling tinggi algoritma GNB yaitu sebesar 80,6% pada *Kfold* 3.
3. Hasil *recall* paling tinggi dalam klasifikasi Breast Cancer adalah algoritma KNN didapatkan hasil sebesar 86,7% pada *Kfold* 5, dan akurasi paling tinggi algoritma GNB yaitu sebesar 77,6% pada *Kfold* 5.
4. Secara keseluruhan kinerja dari dua algoritma ini dalam mengklasifikasikan *Breast Cancer* sudah cukup baik, terbukti dengan menghasilkan nilai akurasi, presisi, dan *recall* yang cukup tinggi, Tetapi KNN mampu menghasilkan nilai presisi dan *recall* lebih tinggi dari pada GNB.

B. Saran

Berdasarkan penelitian yang telah dilakukan, beberapa saran yang dapat dikembangkan untuk penelitian selanjutnya adalah sebagai berikut:

1. Diharapkan untuk peneliti selanjutnya agar menggunakan data *Breast Cancer* terbaru untuk dapat mendapatkan hasil yang maksimal.
2. Untuk penelitian selanjutnya bisa menggunakan algoritma lainnya contohnya *SVM*, *Random Forest* dan *Decision Tree*.

3. Untuk penelitian selanjutnya bisa menggunakan *Kfold* yang lebih beragam

6. DAFTAR PUSTAKA

Alarifi, G. S., Young, & Hana, S. 2018. *Using Multiple Machine Learning Algorithms to Predict Autism in Children Int'l Conf.* Artificial Intelligence | ICAI'18 |.

Asiyah, S. N. 2016. *Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K-Nearest Neighbor*. Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember. Tersedia di https://www.semanticscholar.org/paper/Klasifikasi-Berita-Online-Menggunakan-Metode-Vector-Asiyah-Fithriasari/55bcac61894b644f0ea5b8bda67feda2f34a5e24. Diakses 9 Juni 2020

Artha, D. T., Adinugroho, S., & Adikara, P. P. 2019. *Klasifikasi Pengidap Kanker Payudara Menggunakan Metode Voting Based Extreme Learning Machine (V-ELM).* Universitas Brawijaya. Fakultas Ilmu Komputer. Program Studi Teknik Informatika. Tersedia di http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/4629/2152/. Diakses 27 Mei 2020

Eyupoglu, C. 2019. *Cancer Classification Using K-Nearest Neighbour Algoritm.* Turkey: Istanbul Commerce University, Department of Computer Engineering. Tersedia di https://www.researchgate.net/publication/320547279_Breast_Cancer_Classification_Using_k-Nearest_Neighbors_Algorithm. Diakses 27 Mei 2020

Kamel, H., Abdulah, D., Mieee, Jamal M.Al-Tuwaijari 2019. *Cancer Classification Using Gaussian Naive Bayes Algorithm.* Iraq: Department of Computer Science College of Science, University of Diyala. Tersedia di https://ieeexplore.ieee.org/document/8950650. Diakses 27 Mei 2020

Harbeck & Gnant. December 2019. *Breast Cancer*. Austria: *Medical University of Vienna. NATURE REVIEWS. DISEASE PRIMERS*. Article citation ID: 2019 5:66, Tersedia di https://www.nature.com/articles/s41572-019-0111-2