Perbandingan Metode Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Morfologi Gen Sel Darah Putih

Muhammad Nur'adli Hasbi Gumay [1], Yuni Widiastiwi[2], Mayanda Mega Santoni[3], Yulnelly[4]
Informatika / Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Jakarta Selatan, DKI Jakarta 12450
muhammadnhg@upnvj.ac.id[1], widiastiwi@yahoo.com[2], megasantoni@upnvj.ac.id[3], yulnelly@upnvj.ac.id[4]

Abstrak. Dibidang kesehatan, mendiagnosis penyakit leukemia merupakan hal yang sulit karena masih didiagnosis secara manual dengan bantuan dokter. Diagnosis manual tersebut dapat mengalami kesalahan yang disebabkan oleh kelalaian manusia. Dari permasalahan tersebut, maka dibutuhkan diagnosis jenis penyakit leukemia menggunakan kecanggihan teknologi yaitu *machine learning* untuk mengatasi permasalahan tersebut. Dalam penelitian ini, *machine learning* tersebut mengolah data yang berasal dari jenis leukemia yaitu *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL) berdasarkan ciri morfologi gen sel darah putih tersebut. Metode pengklasifikasian data yang digunakan untuk penelitian ini yaitu K-Nearest Neighbor (K-NN) dan Naïve Bayes yang kemudian kedua metode klasifikasi tersebut dibandingkan untuk melihat metode klasifikasi yang terbaik. Penelitian ini menggunakan praproses *data cleaning*, seleksi fitur, dan *scaling* untuk meningkatkan nilai akurasi. Hasil dari penelitian ini adalah metode klasifikasi K-Nearest Neighbors (K-NN) merupakan klasifikasi yang terbaik dengan nilai akurasi yang menggunakan kurva ROC/AUC bernilai 0.952 jika dibandingkan dengan metode klasifikasi Naïve Bayes yaitu 0.912.

Kata Kunci: Perbandingan, Naïve Bayes, K-Nearest Neighbors, Leukemia, *Machine Learning*

1 Pendahuluan

Leukemia atau kanker darah merupakan penyakit sel darah yang disebabkan oleh pertumbuhan tidak normal pada sel darah putih (leukosit), dimana sel darah putih muda tidak menjadi matang seperti seharusnya. Penyakit ini dapat menyerang bagian tubuh manusia. Bagian tubuh dari penderita akan terinfeksi mengalami gangguan pada sel darah putihnya sehingga menyebabkan sistem imun mengalami gangguan. Penyakit leukimia memiliki dua klasifikasi yaitu klasifikasi *granulocytes* dan klasifikasi *monocytes*.

Dibidang kesehatan, mendiagnosis penyakit leukemia merupakan suatu hal yang sulit karena kurangnya peralatan yang dapat mendeteksi penyakit leukimia dan masih didiagnosis secara manual dengan bantuan dokter. Diagnosis manual tersebut dapat mengalami kesalahan yang disebabkan oleh kelalaian manusia [1]. Dari masalah tersebut, diagnosis jenis penyakit leukemia dapat dibantu oleh kecanggihan teknologi yaitu kecerdasan buatan. Kecerdasan buatan telah dimanfaatkan oleh banyak orang untuk membuat keputusan melalui komputer otomatis [2]. Dalam kasus ini, kecerdasan buatan tersebut akan mengolah data yang berasal dari ciri morfologi gen sel darah putih. Dengan adanya teknologi komputer otomatis, kecerdasan buatan dapat mendiagnosis penyakit leukemia menggunakan teknik klasifikasi seperti metode Naïve Bayes dan metode K-Nearest Neighbor (KNN).

Berdasarkan penelitian sebelumnya dan permasalahan yang ada, pada penelitian bermaksud untuk membandingkan metode Naïve Bayes dan metode K-NN untuk klasifikasi ciri morfologi gen sel darah putih. Dengan menggunakan dua metode tersebut, nantinya akan melihat nilai hasil tingkat akurasi dari kedua metode tersebut dan setelah itu dilakukan perbandingan dari hasil akurasi kedua metode tersebut untuk menentukan nilai optimal dan metode terbaik untuk mengolah data ini.

2 Landasan Teori

2.1 *Acute Myeloid Leukemia* (AML)

Leukemia mieloid akut (AML) adalah penyakit heterogen dan kompleks yang ditandai dengan proliferasi sel yang cepat, perjalanan klinis yang agresif, dan umumnya kematian yang tinggi. Leukemia mieloid akut (AML) ditandai dengan peningkatan jumlah sel mieloid di sumsum dan terhentinya pematangannya, seringkali mengakibatkan insufisiensi hematopoietik (granulositopenia, trombositopenia, atau anemia), dengan atau tanpa leukositosis [3].

2.2 *Acute Lymphoblastic Leukemia* (ALL)

Leukemia limfoblastik akut (ALL) adalah kanker yang paling umum dan juga terjadi pada orang dewasa. Meskipun hasil dari rejimen kemoterapi multi-agen telah sangat meningkat, toksisitas tinggi dan kambuh pada banyak pasien memerlukan pengembangan pendekatan terapeutik baru [4]. Kemajuan dalam profil molekuler dan sitogenetik telah mengidentifikasi berbagai kelainan genetik, termasuk mutasi gen, translokasi kromosom dan aneuploidy, yang telah memberikan pemahaman yang lebih komprehensif tentang biologi dan patogenesis AL.

2.3 Naïve Bayes

Merupakan suatu teknik yang menggunakan prediksi dari probabilitas yang sederhana. Prediksi tersebut berdasarkan dari penerapan teorema Bayes yang sangat kuat di asumsi independensinya. Selain itu, Naïve Bayes menggunakan fitur yang independen di dalam modelnya. Independen tersebut memiliki arti yaitu kuat pada fiturnya. Selain itu, independen di dalam model Naïve Bayes merupakan data yang terdapat di model tersebut tidak berkaitan dan tidak ada hubungannya dengan data yang lainnya di dalam atribut yang lain dan di kasus yang sama [5].

2.4 K-Nearest Neighbors (KNN)

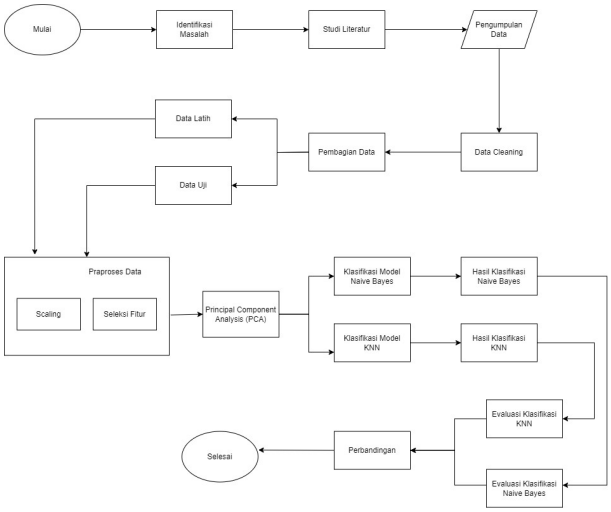
Metode ini merupakan mengklasifikasikan sebuah data baru yang memiliki tujuan yaitu untuk mencari data yang similaritas atau yang mirip dan setelah itu metode ini akan mengambil label data atau kelas yang similaritas tersebut. Kemudian, data yang mirip (similaritas) tersebut dihitung dengan menggunakan metrik yang diukur berdasarkan jarak [6].

2.5 Kurva *Receiver Operating Characteristic* (ROC)

Kurva *Receiver Operating Characteristic* (ROC) merupakan representasi grafis berdasarkan interaksi antara sensitivitas dan 1-spesifisitas. pada penelitian medis kurva *Receiver Operating Characteristic* (ROC) dipakai untuk mendeskripsikan keakuratan diagnostik dan mimilih nilai cut-off yang optimal. Keakuratan penaksiran didapatkan dari wilayah pada bawah kurva ROC dan optimal *cut-off* dipergunakan untuk mengidentifikasi syarat positif dan negatif dalam penaksiran. banyak penelitian sudah memakai kurva ROC menggunakan metode empiris untuk mendeskrisikan keakurasiannya. Analisis *Receiver Operating Characteristic* (ROC) digunakan untuk mendeskripsikan, mengklasifikasikan dan mengatur beberapa jenis yang telah ditentukan oleh model statistik sesuai dengan cara kerjanya. ROC berkembang pada saat perang dunia kedua yang digunakan untuk mengetahui akurasi dalam membedakan sinyal-sinyal yang tertangkap oleh radar. Analisis ROC menyebar dalam penggunaannya untuk mendeskripsikan kebiasaan sistem diagnosis. Pada bidang medis, analisis *Receiver Operating Characteristic* (ROC) ini digunakan untuk pengambilan keputusan dengan hasil kurva ROC dalam pengujian diagnostik [7].

3. Metode Penelitian

3.1 Tahap Penelitian



Gambar. 1. Alur penelitian

1. Identifikasi Masalah

Pada tahap identifikasi masalah akan melakukan identifikasi terhadap permasalahan yang ada yaitu melakukan diagnosis terhadap jenis penyakit leukemia. Permasalahan yang muncul ketika mendiagnosis jenis penyakit leukemia yaitu masih secara manual sehingga memerlukan waktu yang lebih lama. Selain itu, permasalahan lainnya ketika mendiagnosis jenis penyakit leukemia secara manual yaitu memiliki resiko terjadinya kesalahan oleh manusia. Oleh karena itu penelitian ini dibuat berdasarkan atas permasalahan yang sudah disebutkan sebelumnya.

2. Studi Literatur

Studi Literatur merupakan pencarian referensi atau sumber berupa tulisan yang relevan terkait penelitian yang digunakan sebagai dasar pengetahuan. Tujuan dari studi literatur yaitu untuk mempelajari dan mengetahui dari studi dan penelitian sebelumnya yang berkaitan dengan penelitian ini. Dalam tahap ini akan mendapatkan ilmu yang baru dan juga di dalam tahap ini akan mempelajari tentang hal-hal mengenai *Acute Lymphoblastic Leukemia* (ALL), *Acute Myeloid Leukemia* (AML) dan juga metode yang digunakan di penelitian ini yaitu klasifikasi menggunakan K-Nearest Neighbor (K-NN) dan Naïve Bayes serta informasi lainnya yang berkaitan dengan penelitian yang dilakukan.

3. Pengumpulan Dataset

Data pada penelitian ini terdapat pada studi *proof-of-concept* yang diterbitkan oleh Golub et al. "*Genomic evolution of cancer models: perils and opportunities. Nature Reviews Cancer*" [8], dataset pada penelitian ini menunjukkan bagaimana kasus-kasus baru kanker dapat diklasifikasikan oleh pemantauan ekspresi gen (melalui microarray DNA) dan dengan demikian memberikan pendekatan umum untuk mengidentifikasi kelas kanker baru. Data ini digunakan untuk mengklasifikasikan pasien dengan *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL). Terdapat tiga dataset yang berisi dataset train yang berjumlah 38 sampel, dataset test yang berjumlah 34 sampel, dan *dataset target* yang berisi 72 sampel yang digunakan dalam penelitian ini. *Dataset train* dan *test* berisi *Gene Description*, *Gene Accession Number* dari *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL), sedangkan dataset target berisi data pasien yang telah diklasifikasikan ke dalam penyakit kanker sel darah putih yaitu *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL). Dataset ini berisi pengukuran yang sesuai dengan sampel *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL) dari Sumsum Tulang dan Darah Perifer. Nilai intensitas telah diskalakan ulang sehingga intensitas keseluruhan untuk setiap chip setara. Himpunan data ini telah dikonversi ke *Comma Separated Value Files* (CSV). Dataset pada penelitian ini dapat diakses pada *UCI Machine Learning Repository*.

4. Data Cleaning

Data Cleaning merupakan pembersihan data yang dilakukan melalui suatu proses untuk memastikan integritas, konsistensi, dan utilitas data yang ada dalam dataset. Tahapan *data cleaning* yaitu mendeteksi kesalahan atau kerusakan pada data, kemudian memperbaiki atau menghapus data tersebut jika tidak digunakan, sehingga hasil data tersebut dapat dilakukan langkah pembagian data dan selanjutnya akan dilakukan praproses data agar dapat dilakukan model klasifikasi menggunakan algoritma K-Nearest Neighbor (KNN) dan Naïve Bayes.

5. Pembagian Data

Setelah dilakukan tahap *data cleaning*, maka selanjutnya adalah membagi dataset menjadi data *train* sebanyak 38 sampel (53%) dan data test sebanyak 34 sampel (47%). Data *train* dapat digunakan untuk membuat model dan data test digunakan untuk tahap klasifikasi dengan menggunakan metode Naïve Bayes dan K-Nearest Neighbors. Namun, *dataset train* dan *dataset test* tersebut keduanya akan dilakukan praproses data terlebih dahulu.

6. Praproses Data

Tahap praproses data merupakan tahap untuk menjadikan data mentah menjadi data yang dapat dengan mudah dipahami dan dapat digunakan untuk proses selanjutnya. Tahap scaling pada penelitian ini menggunakan *scikit-learn* yang dimana memiliki fitur untuk *preprocessing* yaitu *StandardScaler*. Seleksi fitur berfungsi untuk mengurangi dimensi data, sehingga fitur yang tidak berguna akan diabaikan dan akan memilih fitur yang terbaik dari fitur awalnya. Pada penelitian ini, proses seleksi fitur yang digunakan yaitu menggunakan library *SelectFromModel*.

7. Principal Component Analysis (PCA)

Setelah dilakukan tahap praproses data, maka hasil dari seleksi fitur yang menggunakan *Logistic Regresion* akan digunakan untuk proses *Principal Component Analysis* (PCA). Proses PCA digunakan untuk menunjukkan hasil berupa plot yang dimana plot tersebut merupakan bukti bahwa dataset train yang sudah dilabeli hasil klasifikasi sudah terpisah menjadi *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL) atau belum. Pada proses ini, nilai komponen PCA yang digunakan yaitu *n_components*=3.

8. Klasifikasi

Setelah dilakukan proses *Principal Component Analysis* (PCA), data train akan dilakukan pemodelan menggunakan metode Naïve Bayes dan K-NN. Setelah model dibentuk, model akan diuji menggunakan data test untuk dilakukan proses prediksi menggunakan algoritma Naïve Bayes dan K-Nearest Neighbors, kemudian akan dikalkulasi untuk mendapatkan nilai akurasi model tersebut. Hasil akurasi setiap model kemudian akan digunakan sebagai perbandingan, sehingga dapat terlihat model klasifikasi apa yang memiliki nilai akurasi tertinggi.

9. Evaluasi

Hasil dari pelatihan dataset yang digunakan akan dibandingkan dengan hasil pengujian. Setelah itu hasil dari pengujiannya akan dibandingkan dengan nilai yang sebenarnya. Proses itu bertujuan untuk mendapatkan *Confusion Matrix*. *Confusion Matrix* akan memberikan hasil recall, akurasi dan precision. Nilai akurasi pada Confusion Matrix dapat dihitung menggunakan persamaan 1

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN}. \qquad (1)$$

Akurasi bisa didapatkan dari jumlah data yang berhasil dikategorikan sesuai kategorinya (TP). Kemudian ditambah dengan jumlah data negatif benar (TN) dibagi dengan jumlah semua data yang ada. Kemudian untuk precision dapat digunakan untuk menghitung pola positif yang berhasil diprediksi dengan benar (TP) dari total pola prediksi dalam kelas positif baik itu data yang berhasil diprediksi dengan benar (TP) dengan data yang berhasil diprediksi bukan kelas positif (FP). Untuk menghitung precision dapat menggunakan persamaan 2 berikut.

$$Precision = \frac{TP}{TP+FP} . \quad (2)$$

Kemudian untuk Recall dapat digunakan untuk melihat ukuran keberhasilan dengan membagi data yang berhasil diprediksi dengan benar (TP) kemudian dibagi dengan jumlah dari data yang berhasil diprediksi negatif salah (FN) dan data yang berhasil diprediksi dengan benar (TP). Untuk menghitung nilai Recall dapat menggunakan persamaan 3 berikut.

$$Recall = \frac{TP}{TP+FN} . \quad (3)$$

4 Hasil dan Pembahasan

4.1 Pengumpulan Data

Dataset yang digunakan merupakan dataset yang didapatkan dari Golub et al. dan dataset ini menunjukkan bagaimana kasus-kasus baru kanker dapat diklasifikasikan oleh pemantauan ekspresi gen (melalui *microarray DNA*) dan dengan demikian memberikan pendekatan umum untuk mengidentifikasi kelas kanker baru dan menugaskan tumor ke kelas yang diketahui. Data ini digunakan untuk mengklasifikasikan pasien dengan *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL). Terdapat tiga dataset yang berisi *dataset train* yang berjumlah 38 sampel (53%), *dataset test* yang berjumlah 34 sampel (47%), dan *dataset target* yang berisi 72 sampel yang digunakan dalam penelitian ini. Jumlah sampel pada dataset merupakan jumlah pasien. Atribut Dataset train dan dataset test berisi *Gene Description*, *Gene Accession Number* dari *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL), sedangkan atribut pada dataset target berisi data pasien beserta jenis cancer yang telah diklasifikasikan ke dalam penyakit kanker sel darah putih yaitu *Acute Myeloid Leukemia* (AML) dan *Acute Lymphoblastic Leukemia* (ALL). Kumpulan dataset yang digunakan pada penelitian ini telah dikonversi ke dalam bentuk file *Comma Separated Value Files* (CSV).

4.2 *Data Cleaning*

Deteksi untuk melakukan *data cleaning* ini adalah deteksi oleh produsen *DNA Microarray* yang digunakan dalam penelitian ini (Nilai-nilai ini terkait dengan *p-values*). Oleh karena itu pada tahap ini untuk mendapatkan hasil yang terbaik dilakukan pengecualian baris untuk semua nilai *Absent* (A), yaitu baris yang tidak dapat digunakan, sehingga di dalam penelitian akan menghapus nilai *Absent* (A) pada kolom call dengan melakukan filter pada kolom *call* tersebut. Tujuan dari penghapusan pada kolom *call* yang bernilai *Absent* (A) yaitu agar mengetahui *Gene Description* apa yang akan digunakan pada penelitian ini. Hasil dari menghapus nilai Absent (A) pada kolom *call* yaitu didapatkan 1802 *Gene Description* yang bernilai *Absent* (A) pada kolom *call* dan 5328 bernilai P (*Present*) maupun M (*Marginal*).

4.3 Pembagian Data

Dataset yang telah dilakukan proses *cleaning* dibagi sesuai dengan data train dan data test. Kemudian dataset tersebut hanya menggunakan kolom jenis *Gene Description* sehingga kolom 'dataset', kolom 'patient' dan kolom 'cancer' tidak digunakan. Hal itu dikarenakan, dalam proses ini hanya membutuhkan kolom yang dapat diproses untuk tahap selanjutnya yaitu tahap praproses data. Hasil pembagian dataset dapat dilihat pada Tabel 1.

Tabel 1. Hasil Pembagian Dataset

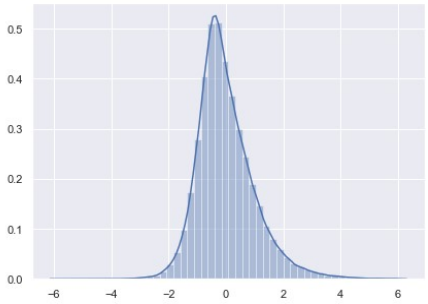
	Data *Train*	Data *Test*
Baris	38	34
Kolom	5327	5327

4.4 Praproses Data

4.4.1 Scaling

Pada proses *scaling* hanya *dataset train* yang dijadikan parameter dalam proses *scaling* dan akan diuji dengan metode seleksi fitur. Pada proses scaling menggunakan *scikit-learn* yang dimana memiliki fitur untuk

preprocessing yaitu *StandardScaler* yang membuat tiap fitur memiliki nilai rata-rata 0 dan variansi 1. Pada saat membandingkan data dalam unit yang berbeda, dapat dibantu dengan menggunakan penskalaan. Untuk menghasilkan hasil yang konsisten di semua kumpulan data, perlu mengubah data sehingga variansnya adalah satu dan meannya nol.



Gambar. 2. Hasil setelah proses Scaling

4.4.2 Seleksi Fitur

Metode seleksi fitur yang digunakan yaitu *SelectFromModel* dan *Logistic Regression*. Dalam prosesnya akan menggunakan *dataset train* yang sebelumnya telah diproses menggunakan metode scaling.

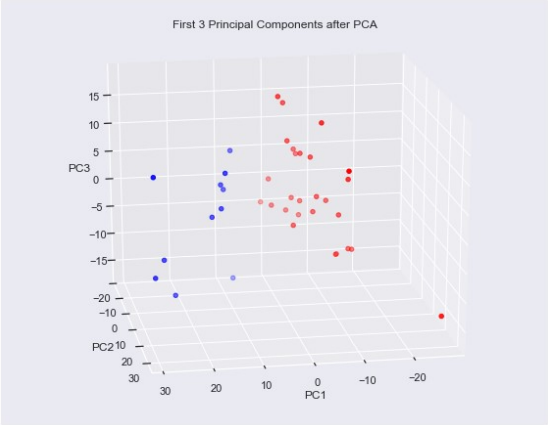
Tabel 2. Jumlah Dataset Train Hasil Logistic Regresion

	Dataset Train setelah proses seleksi fitur menggunakan *Logistic Regression*
Baris	38
Kolom	1262

Pada Tabel 2 terlihat bahwa jumlah data train berkurang dari dataset train awal. Hal itu dikarenakan sudah terjadi proses seleksi fitur dengan menghapus dan mengurangi sejumlah data sehingga dapat mempersingkat waktu komputasi. Selain itu, seleksi fitur berfungsi untuk meningkatkan akurasi klasifikasi dari algoritma yang digunakan.

4.5 Principal Component Analysis (PCA)

Merupakan salah satu cara untuk mengurangi kompleksitas komputasi dengan cara mengubah data ke dalam dimensi lebih kecil. Jika dapat mengurangi jumlah dimensi input, maka kompleksitas model akan berkurang, dan akan lebih mudah untuk dipahami. Ide dasarnya adalah mencari bentuk data (pada dimensi lebih kecil) yang memiliki nilai varians tinggi untuk setiap fitur, karena dengan varians tinggi berarti kemampuan fitur yang tinggi untuk diklasifikasi.



Gambar. 3. Persebaran Dataset setelah menggunakan PCA

Plot pada Gambar 3 menunjukkan pemisahan yang relatif dari dua jenis kanker sel darah putih yaitu *Acute Lymphoblastic Leukemia* (ALL) dan *Acute Myeloid Leukemia* (AML) sepanjang 3 komponen pertama seperti yang didefinisikan pada saat menggunakan PCA. Bahkan hanya menggunakan PC1 dan PC2 sudah akan memisahkan data nya. Titik yang berwarna merah pada plot Gambar 3 menunjukkan dataset *Acute Lymphoblastic Leukemia* (ALL) dan yang bewarna biru menunjukkan dataset *Acute Myeloid Leukemia* (AML).

4.6 Klasifikasi

4.6.1 Naïve Bayes

Klasifikasi Naïve Bayes memiliki keuntungan yaitu pada Naïve Bayes hanya membutuhkan data yang kecil dari penelitian ini yang nantinya digunakan untuk menentukan nilai estimasi parameter apa saja yang dibutuhkan dan diperlukan untuk proses klasifikasi. Untuk menghitung dan menghasilkan nilai akurasi, dataset train akan dilakukan prediksi terhadap dataset test. Selain itu, pada penelitian ini juga akan mencari nilai precision, recall, f1-score, support pada model klasifikasi Naïve Bayes. Untuk nilai dari confusion matrix menggunakan klasifikasi Naïve Bayes dapat dilihat pada Tabel 3.

Tabel 3. Nilai Confusion Matrix metode Naïve Bayes

		Nilai Aktual	
	Values	*Positive*	*Negative*
Nilai Prediksi	*Positive*	16	4
	Negative	0	14

Tabel Error! No text of specified style in document.**.** Hasil Model Klasifikasi menggunakan Naïve Bayes

	precision	**recall**	**f1-score**	**support**
ALL	1.00	0.80	0.89	20
AML	0.78	1.00	0.88	14
accuracy			0.88	34
macro avg	0.89	0.90	0.88	34
weighted avg	0.91	0.88	0.88	34

Dari Tabel 4, hasil untuk nilai akurasi metode klasifikasi Naïve Bayes adalah sebesar 88% dan untuk nilai precision nya menunjukkan bahwa nilai untuk *Acute Lymphoblastic Leukemia* (ALL) sebesar 100 % dan hasil tersebut lebih besar jika dibandingkan dengan *Acute Myeloid Leukemia* (AML)

4.6.2 KNN

Metode K-Nearest Neighbors digunakan pada penelitian ini untuk mencari nilai akurasi dan dibandingkan dengan metode klasifikasi Naïve Bayes. Sama seperti metode klasifikasi sebelumnya yaitu Naïve Bayes, dataset yang telah dikumpulkan, kemudian dipisah menjadi dataset train dan test dan melalui beberapa proses lainnya seperti praproses data dan diuji dengan Model *Selection Kneighbors Classifier* (n_neighbors=5).

Tabel 5. Nilai Confusion Matrix metode KNN

		Nilai Aktual	
	Values	*Positive*	*Negative*
Nilai Prediksi	*Positive*	20	0
	Negative	7	7

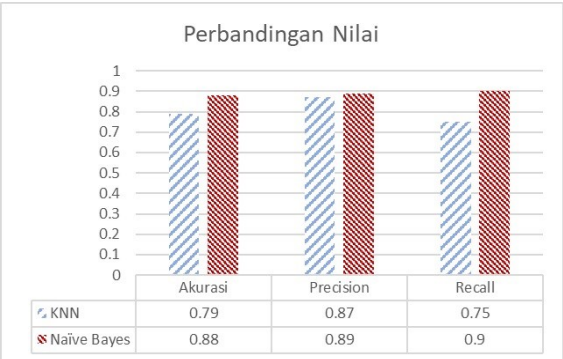
Tabel 6. Hasil Model Klasifikasi menggunakan KNN

	precision	**recall**	**f1-score**	**support**
ALL	0.74	1.00	0.85	20
AML	1.00	0.50	0.67	14
accuracy			0.79	34
macro avg	0.87	0.75	0.76	34
weighted avg	0.85	0.79	0.78	34

Pada Tabel 6 merupakan hasil dari model klasifikasi menggunakan K-Nearest Neighbors dan didapatkan nilai akurasi adalah 79%. Kemudian untuk nilai *precision* dari metode klasifikasi ini menunjukkan bahwa pada Dataset *Acute Myeloid Leukemia* (AML) bernilai 100 % dan nilai tersebut lebih besar dibandingkan pada *Acute Lymphoblastic Leukemia* (ALL).

4.6.3 Perbandingan Metode Klasifikasi Naïve Bayes dan KNN

Setelah dilakukannya pengujian dengan menggunakan kedua metode klasifikasi yaitu K-Nearest Neighbor dan Naïve Bayes, maka selanjutnya yaitu melakukan perbandingan untuk hasil nilai akurasi, *precision*, dan *recall* yang ditunjukkan pada Gambar 3.

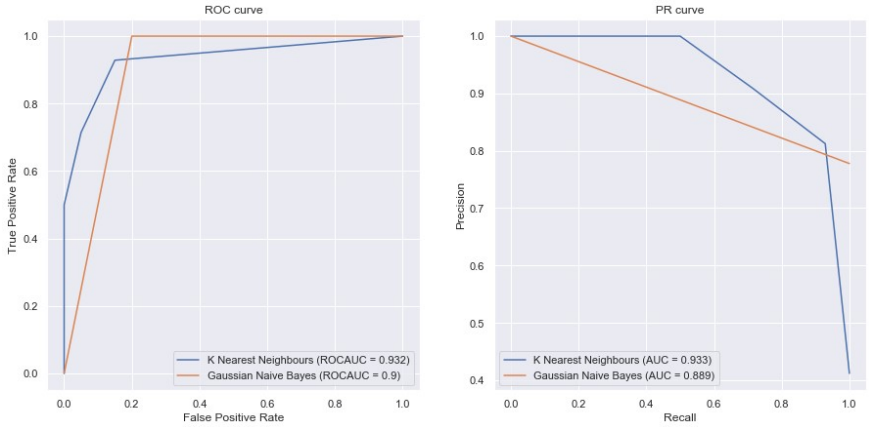


Gambar. 4. Perbandingan Nilai KNN dan Naïve Bayes

Berdasarkan Gambar 3, didapatkan hasil yaitu metode Naïve Bayes merupakan metode yang memiliki nilai akurasi lebih baik dibandingkan dengan metode K-Nearest Neighbor, akan tetapi di dalam penelitian ini untuk menentukan dan mencari metode klasifikasi apa yang terbaik perlu dilakukan pengujian menggunakan metode ROC/AUC.

4.7 Evaluasi

Setelah mendapatkan nilai akurasi dan nilai *confusion matrix*, maka dilakukan proses *Receiver Operating Characteristic* (ROC) untuk mendeskripsikan korelasi antara kelas yang diamati dan kelas yang diprediksi. Keakuratan pembagian terstruktur mengenai ROC ditentukan menggunakan cara menghitung luas wilayah pada bawah kurva ROC. Kurva ROC artinya kurva yang menyajikan gambaran performansi berasal dari binary classifier system dalam membentuk sebuah prediksi. Plot pada Gambar 4 adalah gambaran untuk nilai *true positive rate* dan *false positive rate*, dan bentuk grafik yang terdapat nilai *Recall* dan *Precision*.



Gambar. 5. AUROC/PR CURVES Naïve Bayes dan KNN

Dari penjelasan di tahap evaluasi ini, dapat disimpulkan bahwa untuk nilai akurasi pengujian terhadap dua metode klasifikasi tersebut menghasilkan metode klasifikasi Naïve Bayes menghasilkan nilai akurasi yang lebih besar yaitu 88%. Akan tetapi, untuk grafik AUC (*Area under the Curve of*) menghasilkan bahwa nilai metode klasifikasi menggunakan K-Nearest Neighbors lebih besar baik untuk nilai pada grafik *False Positive Rate* dan *True Positive Rate* dan grafik untuk *Recall* dan *Precision*. Oleh karena itu, pada penelitian ini yang menggunakan dataset *acute myeloid leukemia* (AML) dan *acute lymphoblastic leukemia* (ALL) jika menggunakan metode klasifikasi K-Nearest Neighbors merupakan metode klasifikasi yang terbaik. Hal itu ditunjukkan dengan nilai AUC (*Area under the Curve of*) yang semakin besar menunjukkan bahwa variabel yang diteliti semakin baik dalam memprediksi kejadian.

5 Kesimpulan

Berdasarkan hasil pembahasan dari penelitian yang dilakukan, maka dapat disimpulkan sebagai berikut:

1. Dataset Leukemia yang digunakan pada penelitian ini terbukti berhasil dan dapat digunakan untuk metode klasifikasi K-Nearest Neighbors dan Naïve Bayes.
2. Metode klasifikasi yang digunakan pada penelitian ini terbukti berhasil untuk dibandingkan dan mencari nilai akurasi tertinggi sehingga mendapatkan hasil akhir yaitu metode klasifikasi K-Nearest Neighbors merupakan metode klasifikasi yang terbaik untuk dataset yang digunakan. Hal itu dibuktikan dengan nilai pada grafik AUC pada klasifikasi menggunakan K-Nearest Neighbors adalah bernilai 0.933 pada grafik *Recall Precision* dan bernilai 0.932 pada grafik *false positive rate* dan *true positive rate* jika dibandingkan dengan metode klasifikasi Naïve Bayes. Karena jika nilai yang didapatkan pada grafik AUC semakin besar menunjukkan bahwa variabel yang diteliti semakin baik dalam memprediksi kejadian.

6 Referensi

[1] Saputra, G. D., & Syidada, S. (2017). Pengenalan Sel Acute Lymphoblastic Leukemia (All) Dengan Menggunakan Metode Jaringan Syaraf Tiruan. Melek IT Information Technology Journal, 3(2), 23-26.

[2] M. M. Kini, S. H. Devi, P. G. Desai, and N. Chiplunkar, "Text Mining Approach to Classify Technical Research Documents using Naïve Bayes," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 7, pp. 386–391, 2015.

[3] Winer, E. S., & Stone, R. M. (2019). Novel therapy in Acute myeloid leukemia (AML): moving toward targeted approaches. Therapeutic advances in hematology.

[4] Mohseni, M., Uludag, H., & Brandwein, J. M. (2018). Advances in biology of acute lymphoblastic leukemia (ALL) and therapeutic implications. American journal of blood research, 8(4), 29.

[5] Fadlan, C., Ningsih, S., & Windarto, A. P. (2018). Penerapan Metode Naïve Bayes Dalam Klasifikasi Kelayakan Keluarga Penerima Beras Rastra. JUTIM (Jurnal Teknik Informatika Musirawas), 3(1), 1-8.

[6] Mustafa, M. S., & Simpen, I. W. (2019). Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba. In SISITI: Seminar Ilmiah Sistem Informasi dan Teknologi Informasi (Vol. 8, No. 1).

[7] Nugroho, D. D., & Nugroho, H. (2020). Analisis Kerentanan Tanah Longsor Menggunakan Metode Frequency Ratio di Kabupaten Bandung Barat, Jawa Barat. Geoid, 16(1), 8-18.

[8] Golub, T. R., Ben-David, U., & Beroukhim, R., (2019). Genomic evolution of cancer models: perils and opportunities. Nature Reviews Cancer, 19(2), 97-109.