Analisis Kinerja Algoritma *Naïve Bayes* dan *k-NN* untuk Memprediksi Penyakit Kanker Paru

Moh. Naezer dan Raden Supriyanto

Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
E-mail: mohnaezer35@gmail.com, supriyanto.r@gmail.com

Abstrak

Penyakit kanker merupakan penyakit pembunuh nomor satu di dunia. Kanker paru merupakan salah satu diantara jenis kanker diseluruh dunia. Dalam bidang kesehatan, diagnosa kanker merupakan masalah yang menantang dan banyak penelitian yang difokuskan untuk meningkatkan kinerja guna mendapatkan hasil terbaik. Berdasarkan permasalahan tersebut salah satu teknik yang dapat digunakan dalam pendeteksian adalah data mining dengan teknik *classification*. Pada penelitian ini, melakukan analisis kinerja *data mining* menggunakan algoritma *Naïve Bayes* dan k-NN untuk memprediksi penyakit kanker paru. Dalam perhitungan algoritma k-NN menggunakan nilai k = 3. Pengujian dilakukan dengan menggunakan perbandingan 75% (972 data) sebagai data training : 25% (326 data) sebagai data testing. Berdasarkan hasil penelitian dan pengujian bahwa algoritma *Naïve Bayes* memiliki akurasi lebih baik dibandingkan dengan k-NN. Algoritma *Naïve Bayes* berdasarkan perhitungan dengan tools memiliki akurasi akurasi tertinggi sebesar 98.8%, sedangkan algoritma k-NN memiliki akurasi 83.7%..

Kata kunci : *Naïve Bayes*, k-NN, Kanker Paru, Klasifikasi.

Pendahuluan

Saat ini penyakit kanker merupakan penyakit pembunuh nomor satu di dunia. Hal ini dapat dilihat dari data pasien yang mengidap penyakit kanker dari beberapa rumah sakit. Kanker paru merupakan salah satu diantara berbagai jenis kanker diseluruh dunia adalah salah satu tumor paling ganas. Menurut penelitian yang dilakukan [1] bahwa setiap tahun, sekitar 1,2 juta kanker paru telah didiagnosis pada jutaan orang dan hampir 1,1 juta orang telah terdiagnosis. Dalam bidang kesehatan, diagnosa kanker merupakan masalah yang menantang dan banyak penelitian yang difokuskan untuk meningkatkan kinerja guna mendapatkan hasil yang memuaskan. Berdasarkan permasalahan tersebut salah satu teknik yang bisa digunakan dalam melakukan pendeteksian adalah data mining. Teknik data mining yang bisa digunakan yaitu teknik *classification.* Pemanfaatan teknik data mining untuk memprediksi penyakit kanker paru telah dilakukan oleh peneliti terdahulu. Beberapa penelitian terdahulu dilakukan terkait dengan prediksi kanker paru menggunakan algoritma klasifikasi data mining: Penelitian yang dilakukan oleh [2] mengimplementasikan algoritma Decision tree, *k-Nearest Neighbour , Logistic Regression, Random Forest* dan *Support Vector*. Dalam hasil penelitian tersebut bahwa tingkat prediksi yang lebih baik dilakukan dengan algortima *K-Nearest Neighbour* dan *Logistic Regression* dibandingkan dengan *Decision tree, Random Forest* dan *Support Vector Machine*. Penelitian dilakukan oleh [3]. Peneliti menganalisis dan memprediksi kanker paru menggunakan algoritma klasifikasi seperti *Naive Bayes* dan algoritma J48. Tujuan utama dari penelitian tersebut adalah untuk memberikan peringatan dini kepada pengguna dan analisis kinerja algoritma klasifikasi. Hasil dari penelitian tersebut menunjukkan bahwa pengunaan algoritma *Naïve Bayes* lebih baik dengan akurasi 0,91 dibandingkan dengan algoritma J48 dengan akurasi 0,85. Penelitian dilakukan oleh [4]. Peneliti menganalisis prediksi kanker paru-paru dengan menggunakan algoritma klasifikasi seperti *Naive Bayes*, Network Bayesian dan Algoritma J48. Tujuan utama dalam penelitian tersebut yaitu untuk memberikan peringatan sebelumnya kepada pengguna dan analisis kinerja algoritma klasifikasi. Hasil penelitian menggunakan Weka dengan beberapa teknik klasifikasi data mining dan ditemukan bahwa algoritma *Naive Bayes* memberikan kinerja yang lebih baik dari pada algoritma klasifikasi lainnya seperti *Bayesian Network* dan J48. Penelitian dilakukan oleh [5]. Dalam penelitian tersebut klasifikasi datanya yaitu kumpulan data pasien be-

DOI : http://dx.doi.org/10.32409/jikstik.22.2.3352

dah toraks (kanker paru) yang mencakup 470 kasus dengan 14 atribut yang telah dikumpulkan. Tujuan dari penelitian tersebut adalah untuk melihat penyebab sesak nafas kanker paru-paru. Dari hasil penelitian tersebut menunjukkan bahwa Naïve Bayes adalah yang terbaik dalam operasi toraks untuk memprediksi kelangsungan hidup setelah satu tahun operasi toraks.

Selanjutnya penelitian dilakukan oleh [6]. Peneliti mengembangkan sistem prediksi kanker untuk memprediksi kanker paru-paru berdasarkan gejalanya. Data yang terkumpul diproses sebelumnya dengan algoritma *data mining* seperti *Decision Tree, Logistic Regression, Random Forest* dan *Support Vector Machines* yang digunakan untuk mengukur kinerja. Tujuan utama dari penelitian tersebut adalah untuk memprediksi jenis kanker dan terapi yang disarankan untuk pasien dengan menggunakan algoritma Random Forest. Penelitian dilakukan oleh [7]. Dalam penelitian tersebut menggunakan algortima *Naïve Bayes, Naïve Bayes+Bagging*, SVM, *Decision Tree* dan k-NN. Hasil eksperimen pada penelitian tersebut bahwa algoritma *Naïve Bayes* dengan metode Bagging (NB+BG) memberikan hasil kinerja paling tinggi dibandingkan algoritma lain seperti SVM, *Decision Tree* dan k-NN. Penelitin dilakukan oleh [8]. Tujuan dari penelitian tersebut yaitu untuk membantu ahli onkologi dan praktisi medis dalam mendiagnosis pasien dengan menganalisis data yang tersedia dan informasi yang relevan. Dalam penelitian tersebut, tiga teknik data mining yang digunakan secara bersama yaitu metode AHP, *Rule Based Association* dan *Naïve Bayes Classifier* digunakan untuk mengusulkan skema diagnosis medis untuk memprediksi kanker. Penelitian dilakukan oleh [9]. Dalam penelitian tersebut, peneliti mengusulkan dua langkah proses untuk mendiagnosis adanya kanker baik kanker jinak ataupun kanker ganas. Pada langkah pertama, fitur diekstraksi dengan menggunakan *Gray Level Co-occurrence Matrix* (*GLCM) GLCM. Gray Level Co-occurrence Matrix (GLCM)* adalah suatu metode yang digunakan untuk analisis tekstur/ekstraksi ciri. Kemudian pada langkah kedua, file sel kanker paru diklasifikasikan baik kanker jinak atau kanker ganas dengan menggunakan *Nearest Neighbour Classifier*. Hasil percobaan menunjukkan bahwa kinerja pendekatan menggunakan *Nearest Neighbour Classifier* terbukti lebih baik dengan akurasi 98,76% dibandingkan dengan klasifikasi SVM dan *Random Forest.* Penelitian dilakukan oleh [10]. Dalam penelitian tersebut, prediksi dilakukan dengan menganalisa kondisi pasien sebelum dan sesudah operasi. Penelitian tersebut mengkombinasikan teknik *boosting AdaBoost* sebagai optimasi level algoritma untuk memprediksi harapan hidup pasien kanker paru-paru pasca operasi bedah toraks. Berdasarkan hasil eksperimen pada penelitian tersebut, maka dapat ditarik kesimpulan bahwa algoritma AdaBoost dapat meningkatkan performa algoritma *k-nearest neighbor* dalam memprediksi harapan hidup pasien pasca operasi bedah toraks sebesar 7.45% dari 77.66% menjadi 85.11%.

Kemudian penelitian dilakukan oleh [11]. Penelitian tersebut melakukan studi komparatif terhadap sekumpulan data antara dua alat ekstraksi data Weka dan Rattle R untuk klasifikasi menggunakan algoritma decision tree. Peneliti menemukan bahwa kedua metode dapat menghasilkan model *decisionTree* dalam waktu yang lebih singkat. Peneliti dapat melihat dengan jelas bahwa Weka lebih baik dalam hal akurasi daripada Rattle R. Penelitian dilakukan oleh [12]. Dalam penelitian tersebut menggunakan dataset *thoracic surgery* dengan menggunakan algoritma *k- Nearest Neighbor.* Berdasarkan hasil pengujian dan perbandingan dari kedua model yang diusulkan, algoritma k-NN dengan optimasi fitur menggunakan metode *Forward Selection* memiliki nilai akurasi lebih baik sebesar 85.74% dibandingkan dengan algoritma k-NN tanpa seleksi fitur sebesar 83.40%. Penelitian dilakukan oleh [13]. Dalam penelitian tersebut digunakan algoritma klasifikasi *Naïve Bayes*, algoritma SVM (*Support Vector Machine*), algoritma k-NN dan algoritma J48. Dari penelitian ini, berdasarkan hasilnya bahwa tidak ada pengklasifikasian tunggal yang lebih baik dari yang lain. Penelitian dilakukan oleh [14]. Dalam penelitian tersebut menggunakan dua metode yaitu metode *Genetic Algorithm* dan *Naïve Bayes Classifier.* Tujuan dari penelitian tersebut yaitu memberikan penanganan dini untuk menekan tingkat kematian pasien kanker paru-paru pasca operasi toraks. Klasifikasi naïve bayes mampu memberikan hasil performansi untuk harapan hidup pasca operasi toraks 17,872%. Sedangkan seleksi fitur *genetic algorithm* mampu memberikan hasil 85,319%. Penelitian dilakukan oleh [15]. Dalam penelitian tersebut algoritma GA (*Genetic Algorithm*) diusulkan untuk menyelesaikan masalah pengoptimalan hasil. Algoritma GA adalah salah satu algoritma yang digunakan untuk mengoptimalisasi hasil akhir berdasarkan inputan data acak, untuk menghasilkan file dari model pelatihan.

Pada penelitian ini, peneliti melakukan analisis kinerja *data mining* menggunakan algoritma klasifikasi untuk memprediksi penyakit kanker paru untuk menjadi sebuah informasi dan pengetahuan bagi pihak terkait. Tujuan dalam penelitian ini adalah menentukan diagnosa dan hasil akurasi menggunakan teknik data mining klasifikasi berdasarkan karakteristik yang terdapat dalam *dataset. Dataset* bersumber dari Kaggle.com (https://www.kaggle.com/ravikiran90/lung-cancer-classification). Dataset tersebut memiliki 11 variabel namun dalam penelitian ini variabel yang digunakan hanya 8 variabel yaitu Age, Smokes, Smokes (years), Smokes (packs/year), AreaQ, Alkhol, Family History dan Result. Jumlah total dataset yaitu 1298 data yang terdiri dari 2 hasil proses yaitu 0 / No (Tidak bere-

siko penyakit kanker paru) dan 1 / Yes (Beresiko penyakit kanker paru). Hasil penelitian ini diharapkan dapat membantu pemerintah, tenaga medis maupun masyarakat dalam mengambil langkah terhadap deteksi dini penyakit kanker paru yang mana salah satu penyakit dengan jumlah resiko kematian tertinggi.

Metode Penelitian

Identifikasi Masalah

Permasalahan yang timbul yaitu pada saat mendiagnosa pasien yang beresiko penyakit kanker paru dengan gejala-gejala yang timbul pada *dataset* yang diteliti. Target dari *dataset* tersebut untuk membedakan pasien yang beresiko penyakit kanker paru dengan pasien yang sehat (tidak beresiko penyakit kanker paru).

Tujuan Penelitian

Tujuan penelitian ini yaitu melakukan prediksi mengenai pasien yang beresiko penyakit kanker paru dengan metode klasifikasi dengan menggunakan model *Naïve Bayes Classifie*r dan KNN.

Studi Pustaka dan Pengumpulan Data

Studi pustaka bertujuan untuk mengetahui teori yang mendukung dalam penelitian yang dikerjakan, serta menjadi referensi terkait permasalahan yang dialami. Dalam tahap pengumpulan data, peneliti mengambil data sekunder bersumber dari Kaggle.com (https://www.kaggle.com/ravikiran90/lung-cancer classification).

Analisis Kebutuhan Sistem

Spesifikasi perangkat keras (hardware) yaitu Processor Intel Core i3-5005u, 2.0Ghz, RAM 4Gb, Flashdisk 4Gb, Harddisk 500 GB. Sedangkan spesifikasi perangkat lunak (software) yaitu Microsoft Windows 10 64bit, Microsoft Office 2010, Orange Tools Data Mining versi 3.25.0.

Dataset

Dataset bersumber dari *Kaggle.com.* Dataset ini memiliki 11 variabel dengan jumlah total 1298 data yang terdiri dari 2 hasil proses yaitu 0 / No (Tidak beresiko penyakit kanker paru) dan 1 / Yes (Beresiko penyakit kanker paru). Dalam penelitian ini variabel yang digunakan hanya delapan (8) variabel yaitu *Age, Smokes, Smokes (years), Smokes (packs/year), AreaQ, Alkhol, Family History* dan *Result,* sedangkan tiga (3) variabel sisanya tidak digunakan yaitu *Name, Member_ID* dan *Diagnosis.* Delapan (8) variabel yang digunakan dapat dilihat dalam Tabel 1.

Tabel 1: Variabel *Dataset*

Variabel	Nilai
Age	18 – 77
Smokes	0 – 34
Smokes (years)	0 – 37
Smokes (packs/year)	0 – 37
AreaQ	1 – 10
Alkhol	0 – 8
Family History	0 – 1
Result	0 – 1

Berikut keterangan dari setiap variabel pada *dataset* berdasarkan pada Tabel 1 :

Age : pasien berusia dari 18 tahun hingga 77 tahun; *Smokes* : nilai kriteria perokok pasien dari 0% hingga 34%;

Smokes *(years)* : nilai kriteria perokok pasien berdasarkan tahun dari 0% hingga 37%;

Smokes *(packs/years)* : nilai kriteria perokok berdasarkan berapa pack untuk setiap tahun dari 0% hingga 37%;

AreaQ : nilai pemeriksaan penyebaran kanker pada area paru dalam persentase dari 1% hingga 10%;

Alkhol : nilai pemeriksaan pasien terkait kandungan kadar alkohol dari 0% hingga 8%.

Family *History* : riwayat keluarga pasien apakah termasuk keluarga pengidap penyakit kanker paru atau bukan. Nilai 0 menunjukkan bukan keluarga pengidap penyakit kanker paru. Sedangkan nilai 1 menunjukkan keluarga pengidap penyakit kanker paru;

Result : Hasil deteksi (0 : No / tidak terdeteksi penyakit kanker paru; 1 : Yes / terdeteksi penyakit kanker paru);

Preprocessing Data

Pada tahap ini *dataset* akan dilakukan pengecekan dan pembersihan pada dataset sehingga fitur yang dilakukan uji coba hanya yang relevan untuk penelitian. Berikut tahapan preprocessing data:

Data Cleaning

Data Cleaning digunakan untuk membersihkan data yang tidak relevan, menghapus/mengisi data kosong sesuai dengan rata-rata dan data *noise.* Dalam penelitian ini ditemukan adanya 23 data yang nilai variabelnya kosong. Nilai vaiabel yang kosong tersebut dapat dilihat pada Tabel 2.

Tabel 2: *Nilai Variabel Kosong*

Age	31
Smokes	20
Smokes (Years)	?
Smokes (packs/years)	?
AreaQ	9
Alkhol	4
Family History	1
Result	0

Pada Tabel 2 nilai variabel *Smokses (Years)* dan *Smokes (pack/year)* ditemukan nilai variabel kosong (?), untuk mengatasinya dapat menggunakan *Data Cleaning* yaitu dengan cara nilai variabel kosong tersebut diisi sesuai dengan rata-rata dari nilai variabel yang tidak kosong. Nilai rata-rata variabel *Smokes (years)* yaitu 1,9537. Sedangkan nilai rata-rata variabel *Smokes (packs/years)* yaitu 1,07623.

Discretization

Diskterisasi yaitu proses pengkategorian atau pengelompokkan nilai berdasarkan masing-masing atribut. Berikut hasil dari proses diskterisasi:

1. *Age*

   Pada atribut usia dikategorikan menjadi dua (2) kelompok, yaitu pasien yang berusia lebih dari sama dengan 18 tahun dan kurang dari sama dengan 77 tahun. Dari dua (2) kelompok usia tersebut dapat dihitung usia pasien yang rentan terdiagnosa penyakit kanker paru yaitu berkisar antara usia 48 tahun.

2. *Smokes*

   Pada atribut ini dikategorikan menjadi dua (2) kelompok, yaitu pasien yang memiliki nilai kriteria perokok dari sama dengan 0% dan kurang dari sama dengan 34%. Dari dua (2) kelompok nilai kriteria perokok pasien yang rentan terdiagnosa penyakit kanker paru yaitu berkisar antara 17%.

3. *Smokes* (years)

   Pada atribut ini dikategorikan menjadi dua (2) kelompok, yaitu pasien yang memiliki nilai kriteria perokok dari sama dengan 0% dan kurang dari sama dengan 37%. Dari dua (2) kelompok nilai kriteria perokok pasien yang rentan terdiagnosa penyakit kanker paru yaitu berkisar antara 18% - 19%.

4. *Smokes (packs/year)*

   Pada atribut ini dikategorikan menjadi dua (2) kelompok, yaitu pasien yang memiliki nilai kriteria perokok dari sama dengan 0% dan kurang dari sama dengan 37%. Dari dua (2) kelompok nilai kriteria perokok pasien yang rentan terdiagnosa penyakit kanker paru yaitu berkisar antara 18% - 19%.

5. *AreaQ*

   Pada atribut ini dikategorikan menjadi dua (2) kelompok, yaitu pasien yang memiliki nilai pemeriksaan penyebaran kanker pada area paru dari sama dengan 0% dan kurang dari sama dengan 10%. Dari dua (2) kelompok nilai pemeriksaan penyebaran kanker pada area paru yang rentan terdiagnosa penyakit kanker paru berkisar antara 5%.

6. *Alkhol*

   Pada atribut ini dikategorikan menjadi dua (2) kelompok, yaitu pasien yang memiliki nilai pemeriksaan pasien terkait kandungan kadar alkohol dari sama dengan 0% dan kurang dari sama dengan 8%. Dari dua (2) kelompok nilai pemeriksaan pasien terkait kandungan kadar alcohol yang rentan rentan terdiagnosa penyakit kanker paru berkisar antara 4%.

7. *Family History*

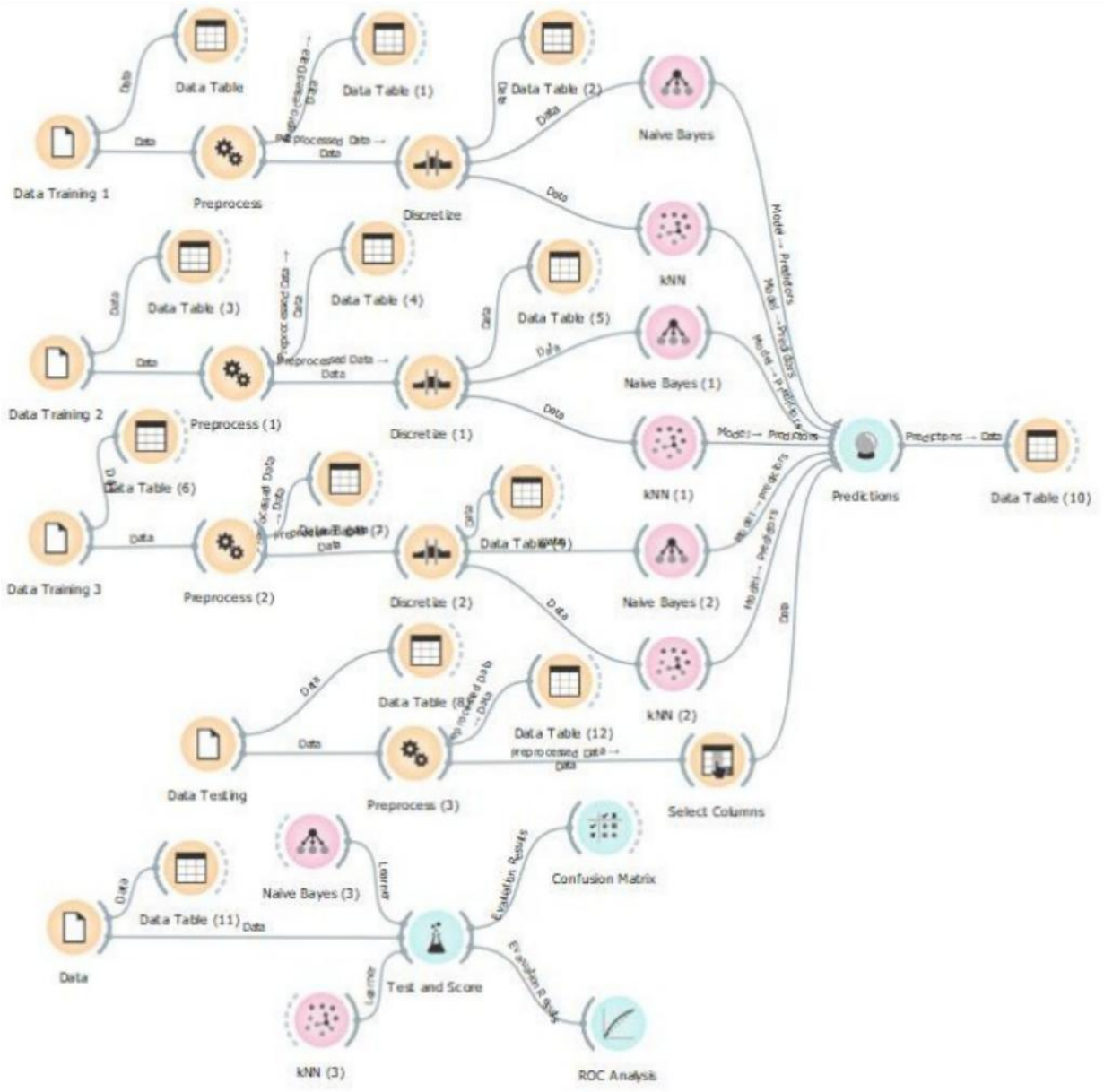
   Pada atribut ini dikategorikan menjadi dua (2) kelompok yaitu pasien yang tidak memiliki riwayat keluarga pengidap penyakit kanker paru bernilai 0 dan pasien yang memiliki riwayat keluarga pengidap penyakit kanker bernilai 1.

8. *Result*

   Pada atribut ini dikategorikan menjadi dua (2) kelompok yaitu pasien yang diprediksi bernilai 0 (Sehat) dan pasien yang diprediksi bernilai 1 (Sakit).

Pelatihan dan Pengujian Model Klasifikasi

Pada pelatihan dan pengujian menggunakan dua model yaitu *Naïve Bayes Classifier* dan *k-Nearest Neighbor* (k-NN), kemudian dibandingkan dari dua model tersebut mana yang memiliki akurasi terbaik dalam pengujian dataset untuk memprediksi penyakit kanker paru. Pengujian dilakukan dengan menggunakan perbandingan 75% (972 data) sebagai data training : 25% (326 data) sebagai data testing. Hasil penelitian nantinya akan diuji dengan confusion matrix dan analisa kurva ROC untuk mengetahui tingkat akurasi dan laju kesalahan yang terdapat dalam setiap pengujian model, lihat Gambar 1.



Gambar 1: Pelatihan dan Pengujian Model Algoritma

Hasil Pengujian dan Analisa Perbandingan Model Klasifikasi

Naïve Bayes Classifier

Pengujian *dataset* dengan model klasifikasi pertama menggunakan model *Naïve Bayes Classifier*. Model ini biasa disebut dengan teknik *Bayes* yang merupakan sebuah teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema *Bayes* dengan asumsi yang *independent* (tidak ketergantungan). Hipotesis dalam teorema *Bayes* merupakan label kelas yang menjadi pemetaan dalam klasifikasi, sedangkan bukti menjadi fitur pendukung. Klasifikasi *Naïve Bayes* bekerja berdasarkan teori probabilitas yang memandang semua fitur dari berbagai data sebagai bukti dalam probabilitas. *Dataset* yang dilakukan pengujian berdasarkan sumber berjumlah 1298 data. Pengujian dilakukan dengan perbandingan 75% (972 data) sebagai data training dan 25% (326 data) sebagai data testing.

k-Nearest Neighbor (k-NN)

Pengujian dataset dengan model klasifikasi kedua menggunakan model *k-Nearest Neighbor* (k-NN). *k-Nearest Neighbor* (k-NN) merupakan salah satu algoritma yang digunakan dalam masalah pengklasifikasian. Algoritma *k-Nearest Neighbor* termasuk dalam algoritma *supervised learning* dimana hasil dari *instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori k-tetangga terdekat. Tujuan dari algoritma ini adalah untuk mengklasifikasikan obyek baru berdasarkan atribut dan sampel dari data training. Algoritma *k-Nearest*

Neighbor menggunakan *Neighborhood Classification* sebagai nilai prediksi dari nilai *instance* yang baru. Prinsip kerja k-NN ialah mencari jarak terdekat antar data yang akan dievaluasi dengan tetangga terdekat dalam data training. Algoritma k-Nearest Neighbor (k-NN) adalah salah satu algoritma paling sederhana untuk memecahkan masalah klasifikasi dan sering menghasilkan hasil yang kompetitif dan signifikan. Untuk menghitung jarak menggunakan jarak *Euclidean*. Dalam penelitian ini menggunakan nilai k = 3. *Dataset* yang dilakukan pengujian berdasarkan sumber berjumlah 1298 data. Pengujian dilakukan dengan perbandingan 75% (972 data) sebagai data training dan 25% (326 data) sebagai data testing.

Hasil dan Pembahasan

Hasil dari setiap model algoritma yang dibuat dapat menampilkan hasil prediksi, nilai akurasi, nilai *recall*, nilai *precision* serta kurva ROC (*Receiver Operating Characteristic*) pada kedua algoritma tersebut.

Hasil dari prediksi hanya ada 4 (empat) kasus yang terjadi :

1. TP (*True Positive*) : kasus dimana pasien diprediksi (Positif) terdiagnosa penyakit kanker paru, memang benar (*True*) terdiagnosa penyakit kanker paru.
2. TN (*True Negative*) : kasus dimana pasien diprediksi tidak (*Negatif*) terdiagnosa penyakit kanker paru dan sebenarnya pasien tersebut memang (*True*) tidak terdiagnosa penyakit kanker paru.
3. FP (*False Positive*) : kasus dimana pasien yang diprediksi (*Positif*) terdiagnosa penyakit kanker paru, ternyata tidak terdiagnosa penyakit kanker paru. Prediksinya salah (*False*).
4. FN (*False Negative*) : kasus dimana pasien yang diprediksi tidak terdiagnosa penyakit kanker paru (*Negatif*), tetapi ternyata sebenarnya (*True*) terdiagnosa penyakit kanker paru.

Pengukuran Performance

1. *Accuracy*

   Merupakan rasio prediksi Benar (positif dan negatif) dengan keseluruhan data. Akurasi dapat menjawab pertanyaan "Berapa persen pasien yang benar diprediksi terdiagnosa penyakit kanker paru dan tidak terdiagnosa penyakit kanker paru dari keseluruhan pasien".

   Akurasi (CA) = (TN+TP) / (TN+TP+FP+FN)

2. *Precision*

   Merupakan rasion prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Precision dapat menjawab pertanyaan "Berapa persen pasien yang benar terdiagnosa penyakit kanker paru dari keseluruhan pasien yang diprediksi terdiagnosa penyakit kanker paru".

   Precision = (TP) / (TP+FP)

   Precision = (TN) / (TN+FN)

3. *Recall*

   Merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Recall dapat menjawab pertanyaan "Berapa persen pasien yang diprediksi terdiagnosa penyakit kanker paru dibandingkan dengan keseluruhan pasien yang sebenarnya terdiagnosa penyakit kanker paru.

   Recall = (TP) / (TP+FN)

   Recall = (TN) / (TN+FP)

Kurva ROC (*Receiver Operating Characteristic*)

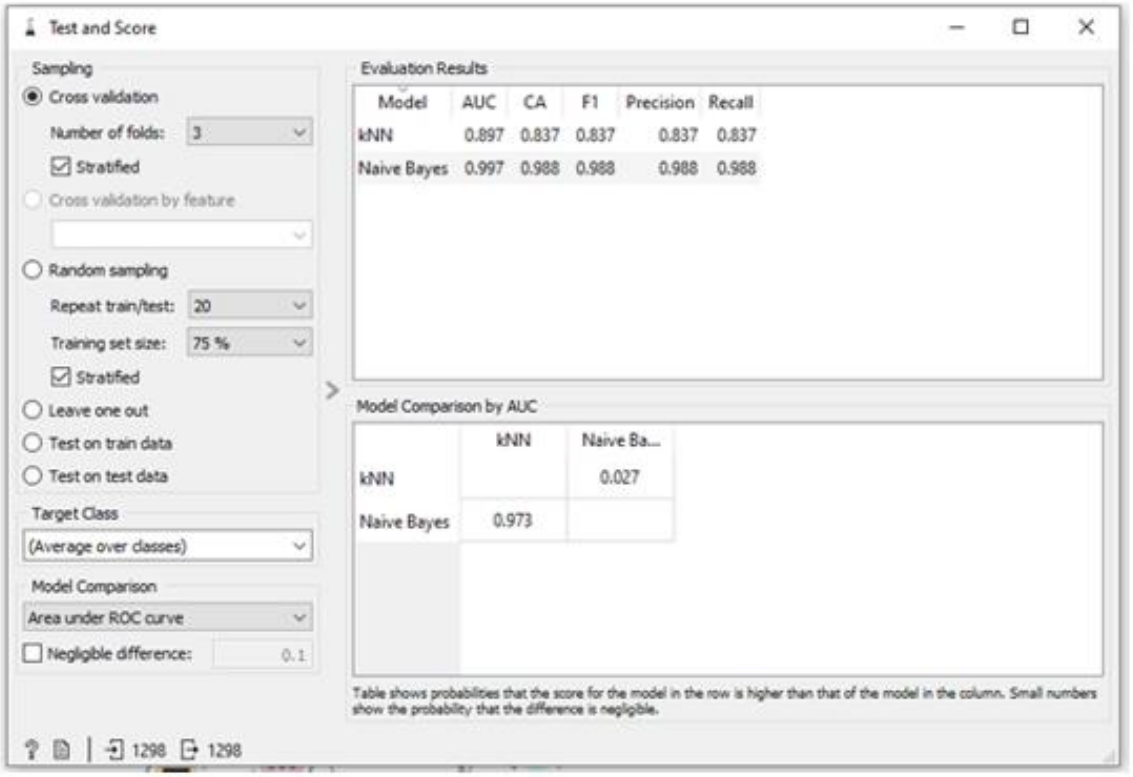
ROC (*Receiver Operating Characteristic*) dan *Area Under Curve* (AUC) merupakan pasangan yang tidak dapat dipisahkan satu sama lain. Meotde ROC adalah suatu metode statistika yang merupakan hasil tarik ulur antara nilai sensitivitas dengan spesifisitas pada berbagai alternatif titik potong yang disajikan dalam bentuk grafik. Sementara AUC adalah hasil wilayah yang dihasilkan oleh kurv ROC. Area yang berada dibawah kurva merupakan wilayah yang menunjukkan tingkat keakuratan dari model prediksi dan dihitung dengan metode perhitungan yang disebut *Area Under Curve* (AUC). Rentang nilai AUC dapat dilihat pada Tabel 3

Tabel 3: *Rentang Nilai AUC*

AUC	Kelas
0.90 – 1.00	*Excellent Classification*
0.80 – 0.90	*Good Classification*
0.70 – 0.80	*Fair Classification*
0.60 – 0.70	*Poor Classification*
<0.60	*Failure*

Hasil Akurasi Pengujian Algoritma *Naïve Bayes Classifier* dengan Perbandingan 75% sebagai *Data Training* : 25% sebagai *Data Testing*

Setelah dilakukan pengujian dan pemodelan dengan menggunakan *tools Orange Data Mining* dengan Algoritma *Naïve Bayes Classifier* menampilkan hasil syang disajikan pada Gambar 2 dan 3.



Gambar 2: *Hasil Test&Score Naïve Bayes Classifier*

Gambar 3: *Hasil Confussion Matrix Naïve Bayes Classifier*

Dari total 1298 data yang dilakukan pengujian, terdapat 677 pasien tidak terdiagnosa penyakit kanker paru. Dari data tersebut diprediksi pasien yang tidak terdiagnosa penyakit kanker paru (sehat) sebanyak 667 data dan hasilnya sesuai, serta 10 data pasien diprediksi tidak terdiagnosa penyakit kanker paru (sehat) namun hasilnya terdiagnosa penyakit kanker paru (sakit). Sedangkan dari 621 data pasien terdiagnosa penyakit kanker paru. Dari data tersebut 6 pasien diprediksi terdiagnosa penyakit kanker paru (sakit) namun hasilnya tidak terdiagnosa penyakit kanker paru (sehat) dan 615

data terdiagnosa penyakit kanker paru (sakit) dan hasilnya sesuai.

Keterangan persamaan confussion matrix sebagai berikut :

TP (True Positive) : 615

FP (False Positive) : 6

TN (True Negative) : 667

FN (False Negative) : 10

Precision (Hasil = Sakit) $=\frac{615}{615+6} = \frac{615}{621} = 0.990 = 99\%$

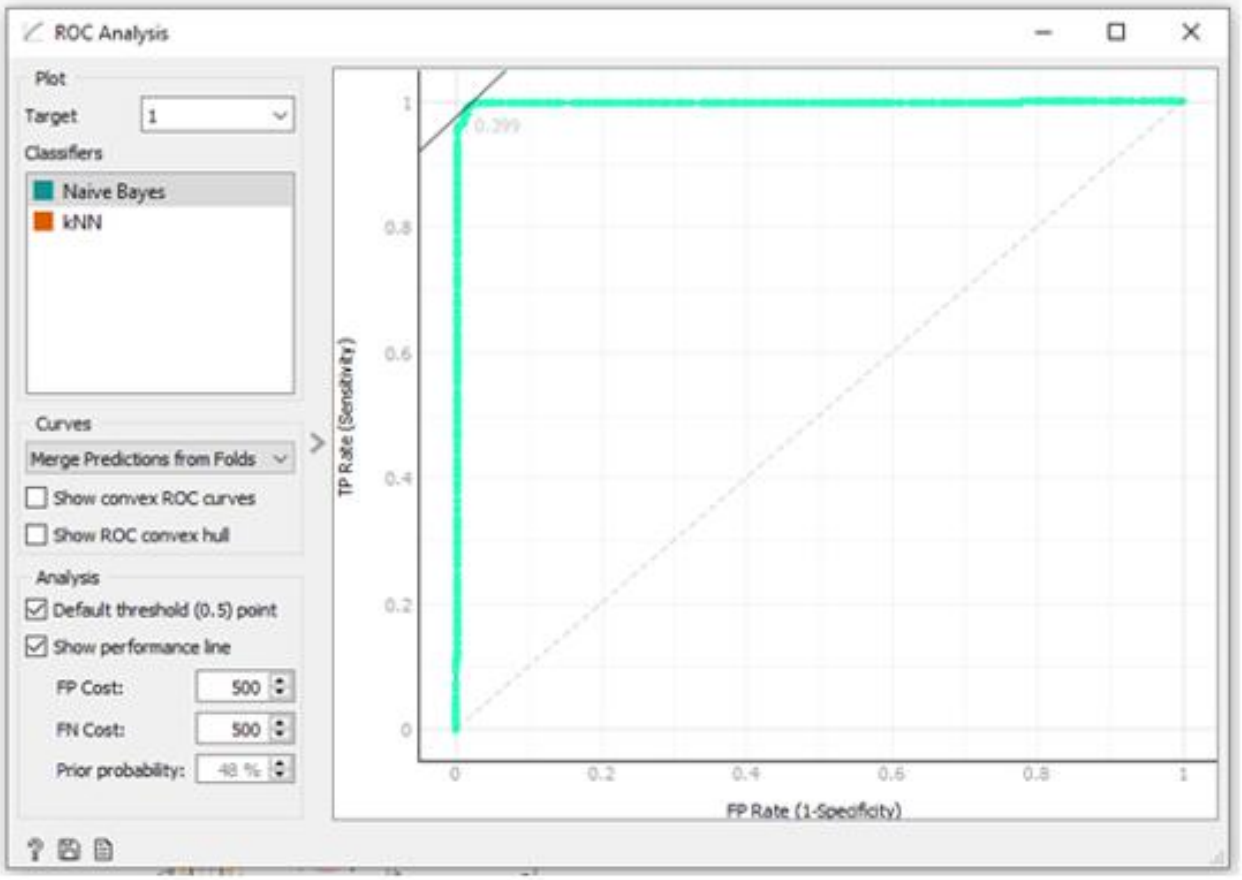
Precision (Hasil = Sehat)$=\frac{667}{667+10} = \frac{667}{677} = 0.985 = 98.5\%$

Recall (Hasil = Sakit) $=\frac{615}{615+10} = \frac{615}{625} = 0.984 = 98.4\%$

Recall (Hasil = Sehat)$=\frac{667}{667+6} = \frac{667}{673} = 0.991 = 99.1\%$

Akurasi (CA) $=\frac{667+615}{667+615+6+10} = \frac{1282}{1298} = 0.988 = 98.8\%$

Hasil perhitungan menunjukkan bahwa Algoritma *Naïve Bayes Classifier* memiliki hasil akurasi sebesar 98.8% dan sesuai dengan pengujian akurasi yang dilakukan pada *tools Orange Data Mining* sebesar 98.8%.



Gambar 4: *Hasil Kurva ROC Naïve Bayes Classifier*

Kurva ROC *Naïve Bayes Classifier*

Hasil dari kurva ROC Algoritma *Naïve Bayes Classifier* disajikan pada Gambar 4.

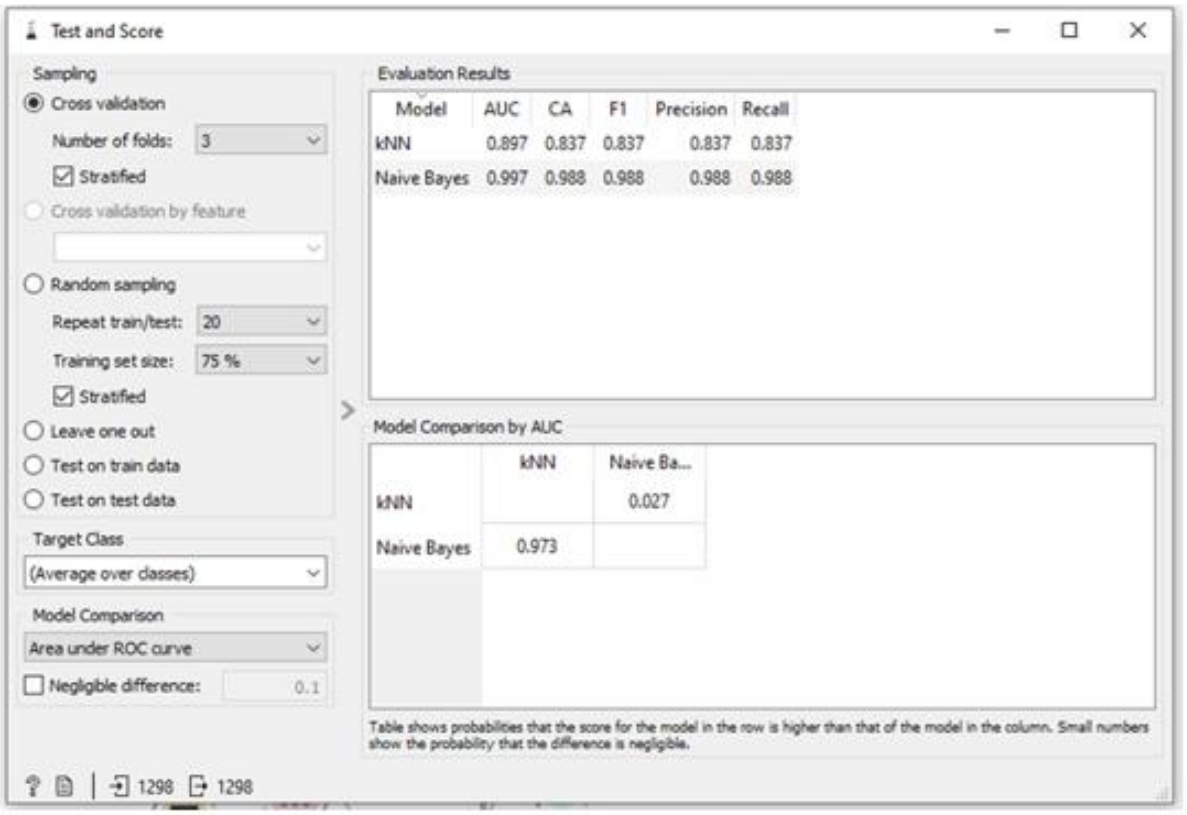
Gambar 4 menunjukkan Kurva *ROC* Algoritma *Naïve Bayes Classifier* dengan nilai *Area Under Curve (AUC)* sebesar 0.997 yang artinya apabila skor terdiagnosa penyakit kanker paru digunakan untuk mendiagnosa ada tidaknya terdiagnosa penyakit kanker paru pada 100 pasien subyek maka kesimpulan yang tepat akan diperoleh pada 99 pasien subyek. Nilai AUC tersebut berdasarkan pada Tabel 3 termasuk kedalam kelas *Excellent Classification.*

Hasil Akurasi Pengujian Algoritma k-Nearest Neighbor (k-NN) dengan Perbandingan 75% sebagai *Data Training* : 25% sebagai *Data Testing*

Setelah dilakukan pengujian dan pemodelan dengan menggunakan *tools Orange Data Mining* dengan Algoritma *k-Nearest Neighbor (k-NN)* menampilkan hasil pada Gambar 5 dan 6.

Dari total 1298 data yang dilakukan pengujian, terdapat 677 pasien tidak terdiagnosa penyakit kanker paru. Dari data tersebut diprediksi pasien yang tidak terdiagnosa penyakit jantung (sehat) sebanyak 578 data dan hasilnya sesuai, serta 99 data pasien diprediksi tidak terdiagnosa penyakit kanker paru (sehat) namun hasilnya terdiagnosa penyakit

kanker paru (sakit). Sedangkan dari 621 data pasien terdiagnosa penyakit kanker paru. Dari data tersebut 112 pasien diprediksi terdiagnosa penyakit kanker paru (sakit) namun hasilnya tidak terdiagnosa penyakit kanker paru (sehat) dan 509 data terdiagnosa penyakit kanker paru (sakit) dan hasilnya sesuai.



Gambar 5: *Hasil Test&Score k-Nearest Neighbor (k-NN)*

Gambar 6: Hasil *Confussion Matrix k-Nearest Neighbor (k-NN)*

Keterangan persamaan *confussion matrix* sebagai berikut :

TP (*True Positive*) : 509

FP (*False Positive*) : 112

TN (*True Negative*) : 578

FN (*False Negative*) : 99

Precision (Hasil = Sakit) = $\frac{509}{509+112} = \frac{509}{621} = 0.082 =$ 8.2%

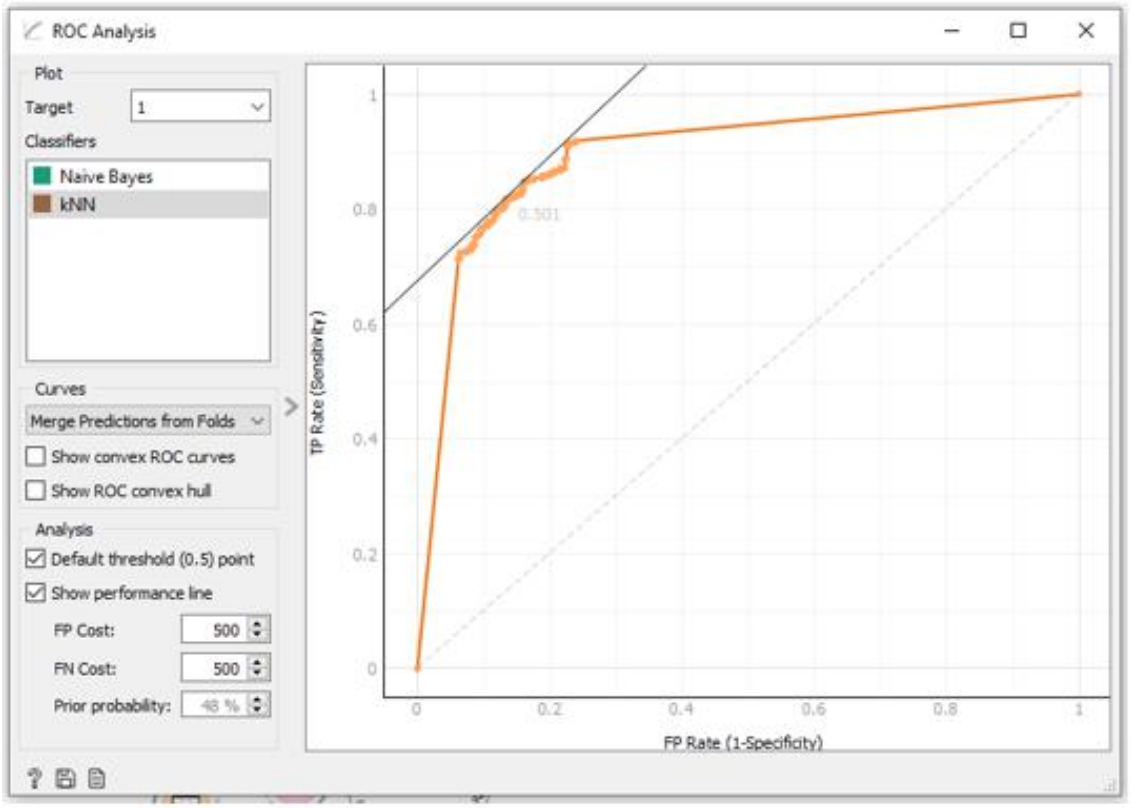
Precision (Hasil = Sehat) = $\frac{578}{578+99} = \frac{578}{667} = 0.854 =$ 85.4%

Recall (Hasil = Sakit) = $\frac{509}{509+99} = \frac{509}{608} = 0.837 = 83.7\%$

Recall (Hasil = Sehat) = $\frac{578}{578+112} = \frac{578}{690} = 0.838 =$ 83.8%

Akurasi (CA) = $\frac{578+509}{578+509+112+99} = \frac{1087}{1298} = 0.837 =$ 83.7%

Hasil perhitungan menunjukkan bahwa Algoritma *k-Nearest Neighbor* memiliki hasil akurasi sebesar 83.7% dan sesuai dengan pengujian akurasi yang dilakukan pada *tools Orange Data Mining* sebesar 83.7%.



Gambar 7: Hasil *Kurva ROC k-Nearest Neighbor (k-NN)*

Kurva ROC k-Nearest Neighbor (k-NN)

Gambar 7 adalah hasil dari kurva ROC Algoritma *k-Nearest Neighbor (k-NN).* Gambar 7 tersebut menunjukkan Kurva *ROC* Algoritma *k-Nearest Neighbor (k-NN)* dengan nilai *Area Under Curve (AUC)* sebesar 0.897 yang artinya apabila skor terdiagnosa penyakit kanker paru digunakan untuk mendiagnosa ada tidaknya terdiagnosa penyakit kanker paru pada 100 pasien subyek maka kesimpulan yang tepat akan diperoleh pada 89 pasien subyek. Nilai AUC tersebut berdasarkan pada Tabel 3 termasuk kedalam kelas *Good Classification.*

Tabel 4: Analisa Komparasi

Algoritma	Akurasi	AUC
Naïve Bayes	98.8%	0.997
k-Nearest Neighbor (k-NN)	83.7%	0.897

Analisa Hasil Komparasi

Setelah melakukan pengujian terhadap dua algoritma tersebut menggunakan *confussion matrix* dan AUC, maka dapat dibuat perbandingan terhadap dua model tersebut dalam memprediksi penyakit kanker paru terhadap pasien disajikan pada Tabel 4.

Dari Tabel 4 dapat dilihat dari segi akurasi bahwa algoritma *Naïve Bayes Classifier* lebih baik dari pada algoritma *k-Nearest Neighbor (k-NN).* Algoritma *Naïve Bayes Classifier* memiliki nilai akurasi 98.8 % sedangkan Algoritma *k-Nearest Neighbor (k-NN)* memiliki nilai akurasi sebesar 83.7%. Selisih akurasi dari kedua model tersebut sebesar 15.1%. Dengan demikian algoritma *Naïve Bayes Classifier* lebih baik dibandingkan dengan algoritma *k-Nearest Neighbor (k-NN).* Kemudian untuk nilai *AUC* yang ditarik dari kurva *ROC* me-

nunjukkan bahwa Algoritma *Naïve Bayes Classifier* memiliki nilai sebesar 0.997, sedangkan model algoritma *k-Nearest Neighbor (k-NN)* memiliki nilai sebesar 0.897. Kedua model memiliki selisih nilai *AUC* sebesar 0.1.

Penutup

Berdasarkan hasil penelitian dan pengujian terhadap *dataset*, maka dapat ditarik kesimpulan bahwa model algoritma *Naïve Bayes Classifier* memiliki akurasi lebih baik dibandingkan dengan model algoritma *k-Nearest Neighbor* (k-NN). Penggunaan *tools* membantu dalam menghitung nilai akurasi maupun AUC dari dua model yang dibandingkan tersebut. Algoritma *Naïve Bayes Classifier* berdasarkan perhitungan dengan tools memiliki akurasi sebesar 98.8% sedangkan algoritma *k-Nearest Neighbor* (k-NN) memiliki akurasi 83.7%. Perhitungan tersebut telah diuji berdasarkan nilai recall maupun presisi baik dari pasien yang berdiagnosa sakit kanker paru maupun yang tidak terdiagnosa penyakit kanker paru. Sedangkan untuk nilai AUC ditarik dari konversi ROC berdasarkan perhitungan dengan *tools* didapatkan hasil bahwa algoritma *Naïve Bayes Classifier* memiliki nilai AUC lebih besar dari *k-Nearest Neighbor* dengan 0.897 berbanding 0.997. Maka dapat ditarik kesimpulan untuk penelitian dengan dataset diagnosa penyakit kanker paru ini lebih baik menggunakan algoritma *Naïve Bayes Classifier* dibandingkan menggunakan algoritma *k-Nearest Neighbor* (k-NN).

Beberapa saran dari peneliti diharapkan dapat membuat penelitian selanjutnya menjadi lebih baik, yaitu penggunaan metode klasifikasi lain seperti *Random Forest, Decession Tree*, SVM dan lain sebagainya dapat dilakukan untuk dibandingkan dengan *Naïve Bayes Classifier* agar dapat melihat model mana yang lebih akurat dalam memprediksi pasien yang terdiagnosa penyakit kanker paru. Kemudian penggunaan data primer dari rumah sakit di Indonesia sehingga dapat diketahui prediksi aktual mengenai kondisi pasien terdiagnosa penyakit kanker paru di Indonesia.

Daftar Pustaka

[1] Priyanka Rajpoot and Mahesh Parmar, "Surveyon Data Mining Classification Techniques for Prediction of Lung Cancer", International Journal of Latest in Engineering and Technology,Vol. 15, No. 2, pp. 61–66, 2019.

[2] E. Yatish Venkata Chandra, K. Ravi Teja, M. Hari Chandra Siva Prasad and Mohammed.Ismail B, "Lung Cancer Prediction Using Data Mining Techniques", International Journal of Recent Technology and Engineering, Vol. 8, No. 4, pp. 12301–12305, 2019..

[3] Ananthnath, Y. Hemalatha G,"Prediction of Lung Cancer Symptoms Using Naïve Bayes and J48 Classification Techniques", International Journal for Scientific Research & Development, Vol. 7, No. 01, pp. 423–427, 2019.

[4] T. Christopher and J Jamera, "Study of Classification Algorithm for Lung Cancer Prediction", International Journal of Innovative Science, Engineering & Technology, Vol. 3 No. 2, pp. 42–49, 2016.

[5] R. Jeena and P. Sarasu, "POPD Disease Diagnosing and Predictions Using Data Mining Algorithms", International Journal of Engineering and Advanced Technology, Vol. 8, No. 2, pp. 258–262, 2019.

[6] M. Bhavani, Sherine Glory, V Pavithra and R Monesh, "Prognosis of Cancer and Proposition of Therapeutics", International Journal of Innovative Technology and Exploring Engineering, Vol. 9, No. 8, pp. 394–397, 2020.

[7] Fitriyani, "Metode Bagging Untuk Imbalance Class Pada Bedah Toraks Menggunakan Naïve Bayes", Jurnal Kajian Ilmiah Universitas Bhayangkara Jakarta Raya, Vol. 18, No. 3, pp. 270–282, 2018.

[8] Pallavi Mirajkar and Andhra Pradesh, "An Integrated Cancer Prediction System Using Data Mining Techniques", International Journal of Scientific Research in Computer Science, Engginering and Information Technology, Vol. 3, No. 1, pp. 1497–1501, 2018.

[9] R. Madana Mohana, R. Delshi Howsalya Devi and Anita Bai, "Lung Cancer Detection Using Nearest Neighbour Classifier", International Journal of Recent Technology and Engineering, Vol. 8, No. 2, pp. 3641–3645, 2019.

[10] Rizki Tri Prasetio and Sari Susanti, "Prediksi Harapan Hidup Pasien Kanker Paru Pasca Operasi Bedah Toraks Menggunakan Boosted K-Nearest Neighbor", Jurnal Responsif, Vol. 1,No. 1, pp. 64–69, 2019.

[11] Soobia Saeed, Afnizanfaizal Abdullah and N Z Jhanjhi, "Analysis of the Lung Cancer Patient ' s for Data Mining Tool", International Journal of Computer Science and Network Security, Vol. 19, No. 7, pp. 90–105, 2019.

[12] Rangga Sanjaya, and Fitriyani, "Prediksi Bedah Toraks Menggunakan Seleksi Fitur", Jurnal Edukasi dan Penelitian Informatika, Vol. 5, No. 3, pp. 316–320, 2019.

[13] E. Sathiyapriya and S. Mary, "A Study on Classification Algorithms and Performance Analysis of Data Mining Using Cancer Data to Predict Lung Cancer Disease", International Jurnal of New Technology and Research, Vol. 3, No. 11, pp. 88–93, 2017.

[14] Yuniar Agung Setyadi, Ibnu Asror, Yanuar Firdaus dan Arie Wibowo, "Prediksi Harapan Hidup Pasca Operasi Toraks Pada Pasien Penderita Kanker Paru-Paru Menggunakan Metode Genetic Algorithm Untuk Feature Selection Dan Naïve Bayes Classifier", e-Proceeding of Engineering,Vol. 7, No. 2, pp. 8349–8360, 2020.

[15] F Leena Vinmalar and A Kumar Kombaiya, "Prediction of Lung Cancer Using Data Mining Techniques", International Journal of Engineering Research & Technology, Vol. 7No. 01, pp. 1–4, 2019.