**KLASIFIKASI KANKER PARU-PARU MENGGUNAKAN METODE DECISION TREE DAN K-NEAREST NEIGHBOR**

**Muhammad A1 Husein1, Rizka Musyarofatul Hidayah2, Sendy Joan Kevin3**

1,2,3 Prodi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang

12041720040@student.polinema.ac.id, 2 aleshariza@gmail.com, 3  sendyjoan5@gmail.com



**Pendahuluan:** Dalam dunia medis Kanker Paru-paru merupakan salah satu penyakit paling mematikan. Menurut Global cancer statistics(Globocan) 2020 mencatat bahwa kematian akibat dari kanker paru-paru meningkat sebesar 34.3783 kasus.Biasanya penyebab dari kanker adalah adanya pertumbuhan sel yang tidak terkontrol pada jaringan paru-paru. Pendekatan (metode) yang dapat dilakukan untuk melakukan klasifikasi tingkat keseriusan penyakit kanker paru-paru berdasarkan data yang diperoleh adalah dengan Decision Tree method dan K-NN Algorithm (K -Nearest Neighbor). Data yang diperoleh merupakan kumpulan faktor penyebab kanker paru-paru yang umum terjadi mulai dari tingkat obesitas,tingkat seringnya merokok,dan lain sebagainya. Kelas klasifikasi yang ada adalah low,medium,high. Metode Decision Tree mengklasifikasikan dan mengidentifikasi masing masing kelas pada faktor penyebab. Dari hasil tersebut akan dilakukan perhitungan dengan k-Nearest Neighbor yang mana menghasilkan tingkat akurasI 100%.

**Kata kunci :** Kanker paru-paru; Decision tree; K-Nearst Neighbor.

**Abstract**

**Introduction:** In the medical world Lung Cancer is one of the deadliest diseases. According to the 2020 Global Cancer Statistics (Globocan) notes that deaths from lung cancer have increased by 34,3783 cases. Usually the cause of cancer is the presence of unstable cell growth in lung tissue. The approach (method) that can be used to classify the level of seriousness of lung cancer based on the data obtained is the Decision Tree method and the K-NN (K-Nearest Neighbor) Algorithm. The data obtained is a collection of factors that cause lung cancer that commonly occur, starting from the level of obesity, the level of frequent smoking, and so on. Existing classification class is low, medium, high. The Decision Tree method classifies and identifies each class on causal factors. From these results a calculation will be carried out with k-Nearest Neighbor which produces an accuracy rate of 100%.

**Keyword :** Lung cancer; decision tree; K-Nearest Neighbor.



1. **Pendahuluan**

Kanker paru merupakan penyakit yang terbentuk di paru-paru. Kanker ini juga menjadi salah satu kanker yang umum terjadi di Indonesia. Secara global, kanker paru merupakan penyebab pertama kematian akibat kanker pada pria dan penyebab kedua kematian akbiat kanker pada wanita. Meski sering terjadi pada perokok, kanker paru-paru juga bisa terjadi pada orang yang bukan perokok. Pada orang yang bukan perokok, kanker paru-paru terjadi akibat sering terpapar asap rokok dan orang lain (perokok pasif) atau paparan zat kimia di lingkungan kerja.

Pada tahun 2018, terdapat 30.023 kasus kanker paru ganas pada laki-laki dan perempuan di Indonesia. Ini membuat kanker paru menjadi kanker ketiga paling umum di Indonesia. Pada tahun 2018, kanker paru menyebabkan 207.210 kematian di Indonesia.(*Indonesia*, n.d.)

Dengan majunya peranan teknologi maka diperlukan sebuah media IT yang dapat membantu petugas ahli dalam menentukan indikasi kanker paru terhadap pasien yang mana dapat diimplementasikan pada sistem kecerdasan buatan. Nantinya perkembangan teknologi ini dapat diterapkan pada rumah sakit yang memiliki keterbatasan spesialis.

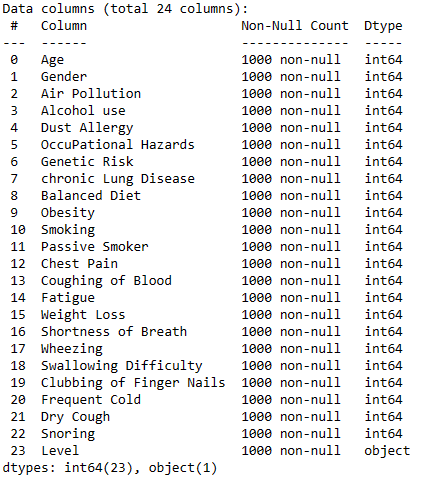
Implementasi kedalam kecerdasan buatan salah satunya adalah menggunakan jenis *supervised learning.* Dalam menggunakan *supervised learning* sendiri terdapat berbagai macam metode yang tersedia. Beberapa diantaranya dan yang akan kami pakai adalah metode *Decision Tree*  dan *K Nearest Neighbor.* Kedua metode ini memiliki tingkatan akurasi yang tinggi dalam melakukan klasifikasi terhadap data yang telah tersedia.

Oleh karena itu, penelitian ini akan mengembangkan teknik komputerisasi klasifikasi pada data pasien penderita kanker paru. Dengan adanya penelitian ini diharapkan sistem klasifikasi yang dikembangkan dapat membantu dokter dan ahli untuk mendeteksi kanker dalam waktu singkat dengan lebih akurat.

1. **Metode Penelitian**

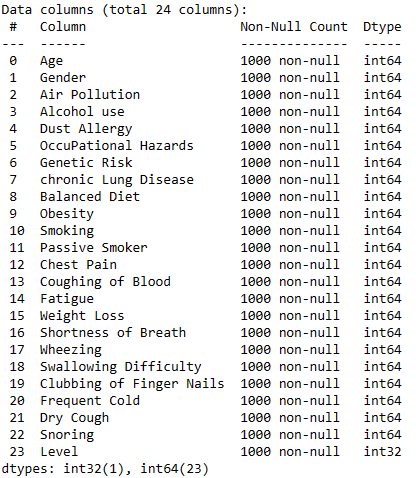
Implementasi klasifikasi kanker paru-paru pada penelitian ini menggunakan dataset sejumlah 303 data dengan level kanker paru-paru rendah, 332 data dengan level kanker paru-paru medium, dan 365 data dengan level kanker paru-paru tinggi. Dataset yang digunakan berupa *csv (comma-separated values file)* diambil dari database kaggle.

Tahap pertama dalam pengolahan citra adalah tahap pemrosesan awal atau sering disebut dengan *cleaning data*. Setiap variable yang ada pada dataset dilakukan proses pemeriksaan informasi mengenai tipe data yang ada pada dataset tersebut. *Cleaning data* dilakukan dengan cara menghapus variable yang tidak digunakan yang mana pada dataset yang digunakan terdapat variable index dan patient id yang harus dihapus dikarenakan tidak memiliki korelasi dengan hasil yang akan digunakan. Sehingga dataset yang akan digunakan hanya akan memiliki variable seperti pada gambar berikut



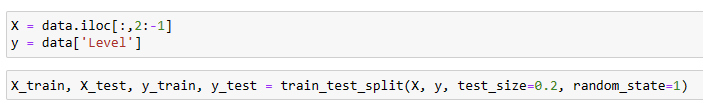
Gambar 1 Variabel yang digunakan

Setelah melakukan penghapusan variable yang tidak digunakan pada dataset, maka langkah selanjutnya adalah melakukan penyamaan tipe data untuk mempermudah dalam melakukan pemrosesan data. Pada dataset tersebut terdapat variable level yang harus dilakukan penyamaan tipe dengan cara melakukan encoding. Langkah encoding sendiri memanfaatkan library yang ada pada *python* yakni library *LabelEncoder* dari *package* sklearn.preprocessing. Sehingga setelah dilakukan cleaning maka informasi mengenai dataset akan tertampil seperti berikut



Gambar 2 Variabel setelah dilakukan cleaning data

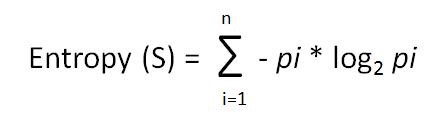
Setelah tahap pemrosesan awal atau *cleaning data* tahap selanjutnya adalah tahap splitting data menjadi data training dan data testing. Yang mana data training nantinya digunakan untuk melatih program agar dapat mengidentifikasi golongan pasien tersebut berdasarkan pada variable yang dimiliki. Sedangkan data testing dilakukan untuk mengevaluasi atau melakukan percobaan terhadap hasil pembelajaran program itu sendiri. Dengan komposisi 8:2 yang wajar digunakan dalam melakukan pembagian data train dan testing maka *pseudocode* yang digunakan pada data sebagai berikut



Gambar 3 Pseudocode Splitting Data

Berdasarkan pada gambar 3 diatas maka sebelum melakukan splitting, data terlebih dahulu didefinisikan sebagai variable X dan Y yang mana variable X akan digunakan sebagai variable yang mempengaruhi hasil dan variable Y sebagai variable hasil dari variable X tersebut.

Setelah data berhasil dilakukan *cleaning* dan *splitting* maka langkah selanjutnya adalah melakukan input kepada metode yang digunakan. Pertama metode yang digunakan adalah *decision tree*. *Decision tree* sendiri nantinya akan menjadikan data yang kita miliki dihitung masing-masing bobotnya sehingga diperoleh sebuah bagan seperti pohon untuk menarik kesimpulan dari data tersebut. Dalam perhitungan *decision tree* diperlukan juga perhitungan *entropy* dan *gain.* Membuat *decision tree* adalah perkara memilih atribut mana yang harus diuji pada setiap simpul pada pohon untuk menentukan ukuran dimana proses ini disebut informasi *gain,* yang berguna untuk menentukan atribut mana yang akan digunakan pada setiap simpul. Informasi *gain* itu sendiri didapatkan dari perhitungan yang menggunakan satuan yang disebut *entropy*. Mendefinisikan kasus pada keputusan biner dan kemudian menentukan kasus umum. Rumus dalam penentuan entropy dapat dilihat sebagai berikut



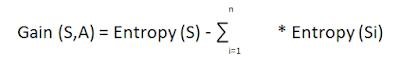
Gambar 4 Rumus Entropy

n : merupakan jumlah nilai yang ada pada atribut target (jumlah kelas klasifikasi)

pi : porsi sampel untuk kelas i

Berikutnya adalah perhitungan informasi gain. Informasi gain merupakan suatu nilai statistic yang digunakan untuk memilih atribut yang akan mengekspansi *tree*  dan menghasilkan *node* baru. Suatu *entropy*  dipergunakan untuk mendefinisikan nilai informasi *gain.*

*Entropy* digunakan sebagai parameter untuk mengukur heterogenitas (keberagaman) dari kumpulan sample data dan jika kumpulan sample data semakin heterogeny, maka nilai *entropynya* semakin besar. *Information gain* dimanfaatkan sebagai efektivitas atribut dalam mengklasifikasikan data dihitung berdasakan *entropy* dengan ketentuan rumus seperti berikut



Gambar 5 Rumus Information Gain

s : himpunan

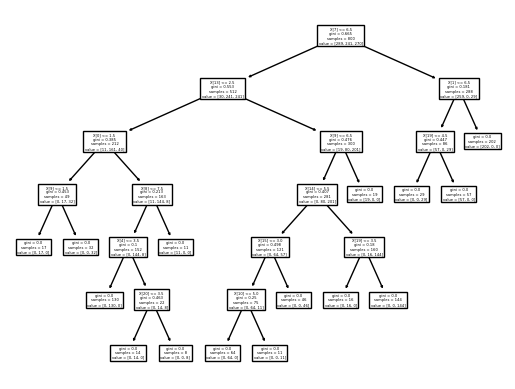
a : atribut

n : jumlah partisi atribut a

|Si| : jumlah kasus pada partisi ke-i

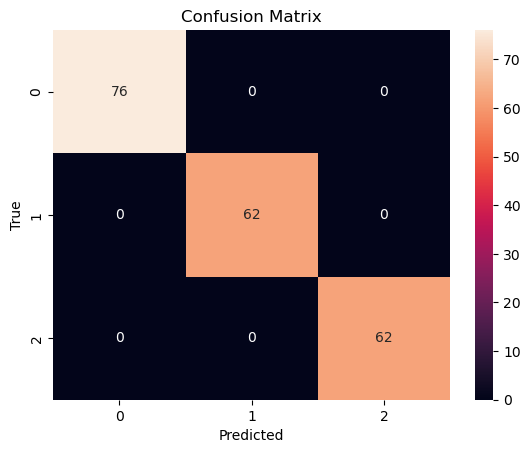
|S| : jumlah kasus dalam S

Setelah melakukan perhitungan berdasarkan dataset sebelumnya makan ditemukan susunan pohon yang ada dalam algoritma *decision tree* pada kasus kanker paru paru ini sebagai berikut



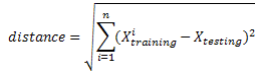
Gambar 6 Hasil Visualisasi Decision Tree

Selain pohon keputusan hasil dari perhitungan *decision tree* tersebut, setelah dilakukan pelatihan dan pengetesan kepada mesin diperoleh sebuah *confusion matrix* yang mana dari *matix* tersebut dapat diketahui berapa besar akurasi yang dihasilkan dari perhitungannya. *Confusion matrix* yang dihasilkan adalah sebagai berikut



Gambar 7 Hasil Confusion Matix Decision Tree

Kedua adalah melakukan langkah yang sama pada data yang telah dilakukan *cleaning* dan *testing* namun menggunakan metode *K Nearest Neighbor*. Algoritma *K Nearest Negihbor (KNN)* adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan keruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasakan klasifikasi data pembelajaran. Sebuah titik pada ruangn ini ditandai kelas c jika kelas c merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat titik tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak *Euclidean* dengan rumus seperti pada persamaan berikut



Gambar 8 Persamaan Euclidean

Xitraining : data training ke-I,

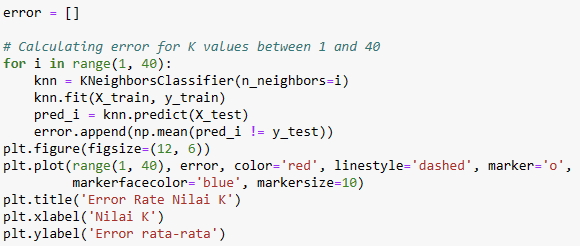
Xtesting : data testing,

i : record (baris) ke i dari tabel,

n : jumlah data mining

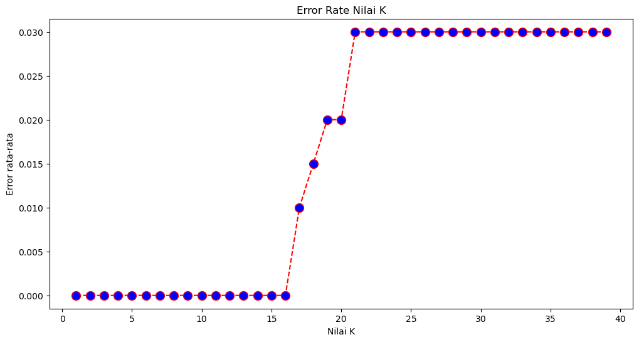
Selama fase pelatihan, algoritma ini hanya menyimpan vektor fitur dan mengklasifikasikan data pelatihan. Selama fase klasifikasi, fitur serupa dihitung untuk data uji (diklasifikasikan tidak terdefinisi). Jarak antara vektor baru ini dan semua vektor data pelatihan dihitung dan elemen K terdekat diperoleh. Skor peringkat baru diprediksi menjadi salah satu yang terbesar dari skor ini. Nilai K terbaik untuk algoritma ini bergantung pada data. Secara umum, nilai K yang tinggi mengurangi pengaruh noise pada classifier, tetapi mengaburkan batas antara masing-masing classifier. Nilai K yang baik dapat dipilih dengan optimasi parameter, misalnya menggunakan cross-validation. Kasus khusus di mana pengklasifikasi diprediksi berdasarkan data pelatihan terdekat (dengan kata lain, K = 1) disebut algoritma tetangga terdekat.

Dalam menentukan besarnya nilai n\_neighbor pada mesin, perlu dilakukan pengechekan Kembali melalui rumus yang telah tersedia sebagai berikut



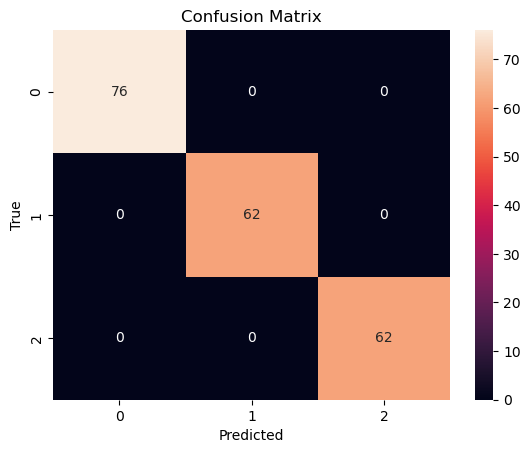
Gambar 9 Pseudocode Hitungan n\_neighbor

Setelah menjalankan code tersebut maka akan muncul sebuah grafik yang menampilkan berapa nilai n yang sesuai dengan nilai error yang mungkin terjadi seperti berikut



Gambar 10 Hasil Perhitungan n\_neighbor

Dalam visual tersebut terlihat bahwasannya metode *K Nearest Neighbor* sesuai menggunakan nilai *n\_neighbor* sebesar 1 sampai 16. Namun tidak sesuai ketika menggunakan nilai 17 sampai 40 karena akan menurunkan akurasi dari metode *K Nearest Neighbor* itu sendiri. Setelah diketahui maka dilanjutkan dengan menggunakan nilai n yang sesuai maka akan mendapatkan *confusion matrix* dengan nilai akurasi sebagai berikut



Gambar 11 Confussion Matrix K Nearest Neighbor

1. **Hasil dan Diskusi**

Hasil klasifikasi dari dataset kanker paru-paru menggunakan metode *decision tree* dan *k nearest neighbor* menghasilkan akurasi yang lumayan baik yakni masing masing senilai 1.0. Performa klasifikasi dari hasil oengujian kepada 20% dataset menghasilkan presentase sebesar 100%. Akurasi klasifikasi semakin baik jika data latih terus ditambah. Hasil dari penelitian ini juga menunjukkan bahwa algoritma *decision tree* dan *k nearest neighbor* handal dalam menangani kesalahan inputan dengan memperbaharui bobot.

1. **Kesimpulan**

Dari penelitian ini dapat disimpulkan bahwa pengembangan teknik klasifikasi menggunakan metode *Decision Tree* dan *K Nearest Neighbor* berhasil mengklasifikasikan pasien berdasarkan pada variable-variable yang digunakan pada dataset. Keluaran dari program klasifikasi yang dikembangkan ini dapat menggolongkan pasien apakah termasuk kedalam kategori level rendah, menengah, hingga tinggi.

1. **Daftar Pustaka**

*Indonesia*. (n.d.).