PERBANDINGAN METODE K-NEAREST NEIGHBOR DAN RANDOM FOREST PADA KLASIFIKASI PENYAKIT KANKER PARU PARU

PROPOSAL TUGAS AKHIR

Disusun Oleh :
Dery Pratama
8020180217

Untuk Persyaratan Penelitian dan Penulisan Tugas Akhir
Sebagai Akhir Proses Studi Strata 1

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS DINAMIKA BANGSA
JAMBI
2022

IDENTITAS PROPOSAL PENELITIAN

Judul Proposal	:	Perbandingan Metode K-Nearest Neighbor dan Random Forest pada Klasifikasi Penyakit Kanker Paru Paru
Program Studi	:	Teknik Informatika
Jenjang Pendidikan	:	Strata 1 (S1)
Peneliti	:	

a. Nama Lengkap	:	Dery Pratama
b. NIM	:	8020180217
c. Jenis Kelamin	:	Laki-Laki
d. Tempat/Tgl. Lahir	:	Jambi, 21 Januari 2001
e. Alamat	:	Jl. Sunan Giri, Lr. Angkasa, No. 93, RT. 32, Kel. Simpang III Sipin, Kec. Kotabaru, Jambi, 36126
f. No. Telepon	:	082170087290
g. E-Mail	:	deryprtm00@gmail.com

1. LATAR BELAKANG

Kanker merupakan pertumbuhan dan penyebaran sel-sel abnormal yang memiliki karakteristik yang khas. Kanker yang sudah menyebar dan tidak dapat terkontrol lagi, biasanya akan menyebabkan kematian. Kanker paru-paru lebih sering menyebabkan pria meninggal dibanding kanker lain, dimana yang sering menjadi penyebab kanker paru-paru adalah merokok. Cara yang digunakan untuk mendeteksi kanker paru-paru ialah melalui pemeriksaan hasil foto rontgen dada [1].

Kanker paru merupakan masalah kesehatan dunia. Dari tahun ke tahun, data statistik di berbagai negara menunjukkan angka kejadian kanker paru cenderung meningkat. Menurut WHO tiap tahun terdapat 1,2 juta penderita kanker paru baru atau 12,3% dari seluruh tumor ganas dan terdapat 1,2 juta atau 17,8% penderita kanker paru yang meninggal dari seluruh tumor ganas. Kanker paru bisa terjadi pada pria maupun wanita. Insiden kanker paru pada pria menduduki urutan kedua setelah kanker prostat, sedangkan pada wanita kanker paru menduduki urutan ketiga setelah kanker payudara dan kanker serviks [2]. Merokok merupakan penyebab utama dari sekitar 90% kasus kanker paru-paru pada pria dan sekitar 70% pada wanita. Semakin banyak rokok yang dihisap, semakin besar resiko untuk menderita kanker paru-paru[3].

Di Indonesia, kanker paru-paru menempati peringkat ke-3 penyakit kanker terbanyak. Paru-paru yang bermasalah akan mempengaruhi fungsi yang ada pada paru-paru-paru, yang mengakibatkan paru-paru tidak bisa berfungsi dengan baik dan akan memberikan efek pada tubuh, kanker paru-paru merupakan salah satu masalah kesehatan yang ada di seluruh dunia dengan hasil yang sangat merugikan bagi penderitanya dan menyebabkan kematian.

Terjadinya kanker ditandai dengan pertumbuhan sel-sel paru yang normal menjadi abnormal atau tidak terbatas dan merusak jaringan-jaringan sel yang normal. Pertumbuhan sel-sel kanker akan menyebabkan jaringan menjadi besar yang biasa disebut tumor ganas [4]. Menurut literature dan penelitian-penelitian terdahulu terdapat beberapa faktor risiko yang menjadi penyebab terjadinya kanker paru. Faktor risiko yang diduga paling berpengaruh terhadap kejadian kanker paru

dan telah banyak penelitian serta bukti statistic yang menunjukkannya adalah merokok. Tiga penyelidikan prospektif yang melibatkan hampir 200.000 pria usia 50-69 tahun, yang diteliti selama 44 bulan menyatakan bahwa angka kematian akibat kanker paru per 100.000 orang diantara mereka yang merokok 10 sampai 20 batang per hari adalah 59,3 dan pada mereka yang merokok 40 batang atau lebih per hari adalah 217,3 [5].

Pengklasifikasian mengenai penyakit kanker paru-paru telah berkembang dari waktu ke waktu.Penelitian kanker paru-paru ini dianggap penting untuk mencari model-model terbaik untuk mengidentifikasi masalah penyakit kanker paru-paru dengan tanda-tanda tertentu yang mendukung penentuan pada klasifikasi penyakit kanker paru paru.dalam hal ini penulis menggunakan algoritma *K-Nearest Neighbor* (KNN) dan *Random Forest* (RF) untuk melihat apa saja faktor-faktor yang menentukan dalam mendeteksi penyakit kanker paru paru, Penulis membandingkan algoritma *Random Forest* (RF) dengan algoritma *K-Nearest Neighbor* (KNN) untuk melihat akurasi klasifikasi yang diberikan

Dalam berberapa penelitian yang ada diatas penulis memutuskan untuk membandingkan algoritma *K-Nearest Neighbor* (KNN) dan Random Forest (RF) untuk melihat akurasi yang dihasilkan dari ke dua model tersebut *K-Nearest Neighbor* dan *Random Forest* sama-sama memiliki akurasi yang tinggi dalam pengklasifikasian penyakit [6].

Berdasarkan uraian diatas, maka diperlukan sistem yang dapat menyelesaikan permasalahan yang ada saat ini. Sehingga penulis mengusulkan judul penelitian **"Perbandingan Metode K-Nearest Neighbor dan Random Forest Pada Klasifikasi Penyakit Kanker Paru Paru"** demi mengatasi berbagai masalah yang telah diuraikan diatas.

2. RUMUSAN MASALAH

Ditinjau bedasarkan latar belakang masalah diatas, maka dapat di tarik menjadi rumusan masalah berikut:

1. Bagaimana cara membuat aplikasi yang bisa mengklasifikasikan metode K-Nearest Neighbor dan Random Forest dari data penyakit kanker paru-paru?

2. Bagaimana hasil perbandingan akurasi dari metode yang telah diterapkan?

3. BATASAN MASALAH

Batasan masalah yang dibahas oleh penulis mencangkup:

1. Ruang lingkup penelitian ini hanya dibatasi oleh sistem aplikasi untuk penerapan metode K-Nearest Neighbor dan Random Forest saja saja.

2. Aplikasi yang digunakan berbasis Web dengan penerapan pada metode K-Nearest Neighbor dan Random Forest,dengan data yang dimiliki sajaTujuan dan manfaat penelitian

4. TUJUAN DAN MANFAAT PENELITIAN

4.1 Tujuan Penelitian

Tujuan yang diinginkan oleh penulis, yakni:

1. Membuat aplikasi pada algoritma K-Nearest Neighbor dan Random Forest untuk klasifikasi data penyakit kanker paru-paru.

2. Untuk mengetahui perbandingan akurasi yang dihasilkan dari metode K-Nearest Neighbor dan Random Forest dan melihat metode mana yang lebih efisien pada klasifikasi data pada penyakit kanker paru-paru.

4.2 Manfaat Penelitian

Manfaat dari penelitian skripsi ini, yakni:

1. Manfaat untuk pengklasifikasian algoritma K-nearest Neighbor dan Random Forest untuk melihat perbandingan akurasi yang dihasilkan dari metode tersebut pada data yang sudah ditentukan.

2. Manfaat untuk penulis, dapat mengimplementasikan ilmu yang telah di dapat pada saat diperkuliahan.

3. Manfaat untuk pembaca, sebagai sumber informasi tambahan untuk penelitian yang sedang dilakukan.

4. Manfaat untuk Program Studi, sebagai acuan perbandingan untuk pengembangan kurikulum dengan pengaplikasian metode yang dilakukan penulis.

5. LANDASAN TEORI

Landasan teori yang penulis gunakan sebagai pedoman untuk mengerjakan penelitian, yakni :

5.1 Kanker Paru

Kanker paru adalah keganasan yang berasal dari luar paru (metastasis tumor paru) maupun yang berasal dari paru sendiri, dimana kelainan dapat disebabkan oleh kumpulan perubahan genetika pada sel epitel saluran nafas, yang dapat mengakibatkan proliferasi sel yang tidak dapat dikendalikan. Kanker paru primer yaitu tumor ganas yang berasal dari epitel bronkus atau karsinoma bronkus [7].

5.2 Definisi Sistem

Azhar Susanto [8] di dalam bukunya, “Bahwa sistem adalah kumpulan atau grup dari sub sistem/bagian/komponen atau apapun baik fisik ataupun non fisik yang saling berhubungan satu sama lain dan dapat bekerja sama untuk mencapai satu tujuan tertentu. Kemudian,dalam bukunya, Sutarman [9] menjelaskan bahwa sistem adalah kumpulan elemen yang saling berinteraksi dalam kesatuan untuk menjalankan suatu proses pencapaian suatu tujuan utama. Sedangkan Jogiyanto [10] dalam bukunya yang berjudul Analisis dan Desain Sistem Informasi bahwa sistem dapat juga didefinisikan dengan pendekatan prosedur dan komponen. Sistem dan prosedur adalah suatu kesatuan yang tidak bisa dipisahkan satu dengan yang lain. Suatu sistem baru dapat terbentuk jika di dalamnya ada beberapa prosedur yang mengikutinya.

Pendekatan dalam mendefinisikan sistem yaitu berdasarkan pendekatan pada prosedurnya dan yang berdasarkan pendekatan komponennya.ada 2 yaitu:

a. Pendekatan sistem pada prosedurnya

Sebuah sistem adalah suatu jaringan dan prosedur yang saling berkaitan satu sama lain, dan bekerja sama dalam melaksanakan suatu pekerjaan atau menyelesaikan suatu masalah

b. Pendekatan sistem pada komponennya

Sebuah sistem adalah sekumpulan dari elemen-elemen yang melakukan interaksi satu sama lain dengan pola teratur sehingga membentuk suatu totalitas untuk menyelesaikan suatu masalah tertentu. Berdasar dari beberapa pendapat ahli yang telah dikemukakan di atas, dapat ditarik sebuah kesimpulan bahwa sistem adalah kumpulan bagian atau beberapa subsistem yang dirancang dan disatukan untuk mencapai suatu tujuan tertentu.

5.3 Artificial Intelligence

Artificial intelligence yang merupakan kecerdasan yang ditunjukkan oleh mesin, berbeda dengan kecerdasan alami manusia dan hewan, kecerdasan buatan tidak melibatkan emosi dan perasaan. Istilah "kecerdasan buatan" sering digunakan untuk mendeskripsikan mesin (atau komputer) yang meniru fungsi "Kognitif" manusia yang diasosiasikan dengan pikiran manusia, seperti belajar dan "pemecahan masalah" [11]. Artificial intelligence menjadi bidang ilmu komputer yang di khususkan untuk membuat perangkat lunak yang cerdas serta mampu melakukan rutin perhitungan yang mirip dengan otak. Banyak sekali metode kecerdasan buatan seperti fuzzy logic, Sistem Pakar (Expert Systems), neural network hingga berhubungan dengan statistik seperti pendekatan Bayesian, Computer Vision, Robot Vision dan Deep Learning.

5.4 Machine Learning

Machine Learning merupakan bidang ilmu komputer yang mempelajari pengenalan pola dan teori pembelajaran komputasi dalam kecerdasan buatan [12]. Machine learning merupakan cabang dari artificial intelligence dengan fokus pada pengembangan sebuah sistem yang mampu belajar sendiri tanpa harus berulang kali

diprogram oleh manusia sehingga sebelum mengeluarkan sebuah hasil dari suatu objek, machine learning membutuhkan data awal sebagai bahan belajar [13]. Machine learning merupakan suatu area dalam artificial intelligence atau kecerdasan buatan yang berhubungan dengan pengembangan teknik-teknik yang diprogramkan dan belajar dari data masa lalu untuk proses pembelajaran agar mendapatkan hasil yang terbaik dari pembelajaran tersebut [14].

5.5 Python

Phyton adalah adalah salah satu bahasa pemrograman tingkat tinggi yang bisa melakukan instruksi multiguna (interpretatif). Phyton biasa digunakan untuk berberapa keperluan seperti mesin pembelajaran. Phyton juga biasa digunakan untuk menangani database,grafik,dan juga game. Phyton dapat dijalankan di berberapa windows seperti Windows, Linux, dan Unix. Tujuan dibuatnya phyton ini adalah untuk pengurangan source program Phyton pertama kali dikembangkan oleh Guiddo van Rossum pada tahun 1990 di belanda [15].

5.6 Django

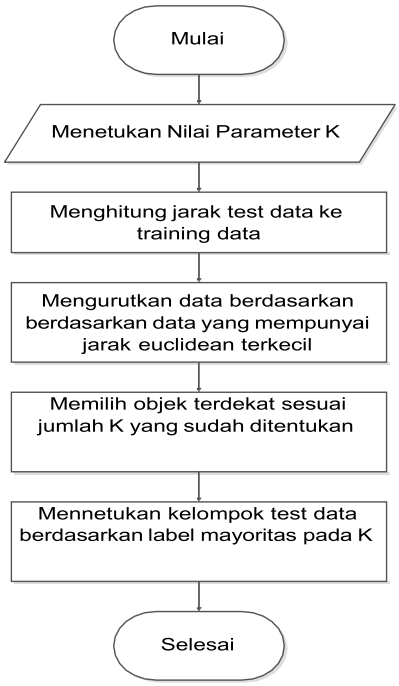
Django adalah sebuah framework full-stack untuk membuat aplikasi web dengan Bahasa pemrograman phyton. Framework ini akan membantu membuat web lebih cepat, daripada menulis kode dari nol. (Django Software Foundation, 2021) Django dibuat pada tahun 2003 oleh Simon Wilison dan Adrian Holovaty. Django digunakan untuk melakukan pengembangkan aplikasi web dengan cepat dan memiliki desain pragmatis yang bersih. Dengan menggunakan Django mebuat pengembangan aplikasi web menjadi lebih mudah, cepat dan lebih sedikit menggunakan kode.

5.7 Metode Klasifikasi

Dalam menganalisa performa dari klasifikasi penulis membandingkan algoritma K-Nearest Neigbor (KNN) dan Random Forest (RF) untuk pengklasifikasian data penyakit Kanker Paru.

5.7.1 K-Nearest Neighbor (KNN)

Algoritma K-Nearest Neighbor adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasi sebelumya. Termasuk dalam supervised learning, dimana hasil query instance yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam (KNN). Adapun tahapan dari algoritma K-Nearest Neighbor.



Gambar 1. 1 Tahapan-tahapan algoritma *K-Nearest Neighbor*

Metode K-Nearest Neighbor ini termasuk dalam metode dengan akurasi yang kuat, dengan menghitung sebuah kasus yang lama dan kasus yang baru dengan pencarian bobot, yang dimana tiap-tiap dimensi untuk mempretasikan fitur yang ada pada data. Prinsip yang ada pada metode K-Nearest Neighbor adalah menggunakan jarak terdekat antara data yang dievaluasi pada tiap atribut-atribut dengan data "K"pada tetangga yang paling dekat dalam data latih [16].

5.7.2 Euclian Distance

Untuk mengurutkan data berdasarkan data yang mempunyai euclidean terkecil dibutuhkannya metode Euclidean distance. Euclidean distance adalah

perhitungan jarak dari dua buah titik dalam euclidean space. Euclidean ini berkaitan dengan Teorama Phytagoras dan biasanya diterapkan pada 1 sampai 3 dimensi. Tapi juga sederhana jika diterapkan pada dimensi yang lebih tinggi [17].

Berikut adalah rumus dari Euclidean distance :

$$dis\ (x_{1i}, x_{2i}) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}$$

Rumus 1.1 Euclidean Distance

5.7.3 Random Forest (RF)

Random Forest adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar klasifikasi yang dilakukan random forest adalah melalui penggabungan pohon (tree) dengan melakukan training dari sampel data yang akan digunakan pemakaian pohon (tree) yang banyak akan dihasilkan akan lebih bagus.penentuan klasifikasi dengan metode random forest diambil berdasarkan hasil voting dan tree yang dibuat. Pemenang dari tree yang terbentuk didapatkan dari vote yang paling banyak. Pohon keputusan dimulai dengan memulai perhitungan nilai entropy sebagai penentu dari tingkat ketidak murnian atribut [18].

Berikut adalah rumus dari menghitung nilai entropy :

$$Entropy\ (Y) = -\sum i\ \ p(c|Y) log_2\ p(c|Y)$$

Rumus 1.2 Entropy

Keterangan :

Y : Himpunan kasus

$(c|Y)$:Proporsi nilai Y terhadap nilai c

Dilanjutkan mencari information gain yang digunakan untuk mengukur efektifitas attribute dalam pengklasifikasian data berikut ini adalah rumus dari information gain.

$$Information\ Gain\ (Y, a) = \text{Entropy(Y)} - \sum Values(a)\ x$$

$$= \frac{|Yv|}{|Ya|}\ Entropy(Yv)$$

Rumus 1.3 Information Gain

Keterangan :

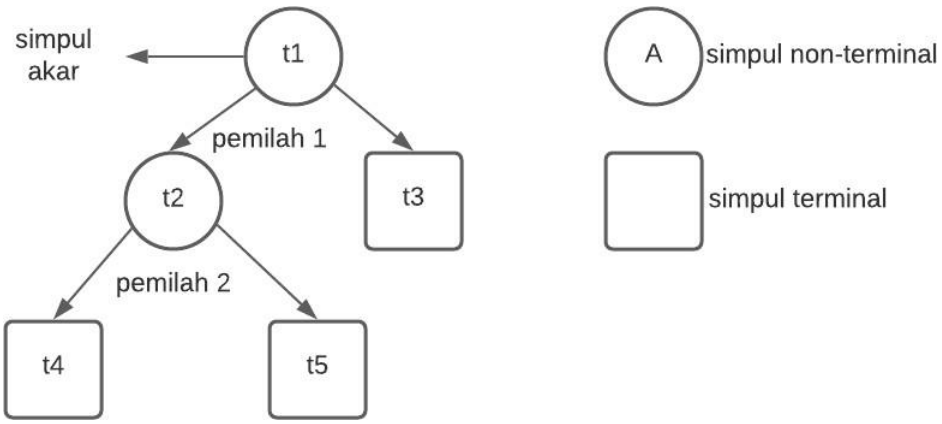
Value (a) : Semua nilai yang mungkin dalam himpunan kasus a

Yv : Semua subkelas dari Y dengan kelas v yang berhubungan dengan kelas a

Ya : Semua nilai yang sesuai dengan a

5.7.4 Classification and Regression Tree (CART)

CART adalah metode eksplorasi data yang didasai dari teknik pohon keputusan. Pohon klasifikasi dihasilkan saat peubah respons berupa data kategorik, sedangkan pohon regresi dihasilkan saat peubah respons berupa data numerik [19]. Pohon terbentuk dari proses pemilahan rekursif biner pada suatu gugus data sehingga nilai peubah respons pada setiap gugus data hasil pemilahan akan lebih homogen [20].



Gambar 1.2 Struktur Pohon pada Metode CART

Berdasarkan ilustrasi pohon diatas yang dapat dilahat pada gambar 2.2 Pohon disimpul oleh simpul t1,t2,t3,t4,t5.Setiap pemilah (split) memilah simpul non-terminal menjadi dua simpul yang saling lepas.

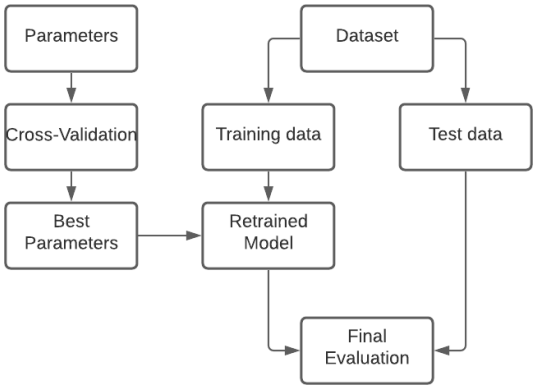
Hasil prediksi respon suatu amatan terdapat pada simpul terminal .

Menurut Breiman et al. (1984), pembangunan pohon klasifikasi CART meliputi tiga hal yaitu :

1. Pemilihan pemilah (split)
2. Penentuan simpul terminal
3. Penandaan label kelas

5.7.5 K-Fold Cross Validation

Metode Cross Validation adalah salah satu metode yang digunakan untuk mengevaluasi algoritma learning dengan membagi data menjadi dua segmen. Yaitu segmen pembelajaran learning (learning) dan segmen latihan (training), segmen latihan digunakan untuk validasi model. Ciri khas dari cross validation yaitu set training dan validasi yang nantinya akan disilangkan (cross over) dalam putaran secara terus menerus sehingga pada tiap titik data mempunyai peluang untuk divalidasi. Dasar dari cross validation adalah K-Fold Validation [21], Cross validation adalah sebuah metode yang banyak digunakan untuk estimasi model dan seleksi variable. Dalam penulisan ini penulis menggunakan prosedur K-Fold Cross Validation, Cara kerja metode K-Fold Cross Validation dengan membagi keseluruhan sampel data latih menjadi k sub-sampel yang berukuran sama. Setiap sub-sampel digunakan untuk pengujian model klasifikasi dan mengulangi proses tersebut sebanyak k kali [22].



Gambar 1.3 Cara Kerja Cross Validation

Dalam penulisan ini penulis menggunakan metode K-Fold Cross Validation, metode ini bekerja dengan membagi keseluruhan sampel data latih menjadi k sub-sampel yang mempunyai ukuran yang sama. Setiap sub-sampel digunakan untuk pengujian dari model klasifikasi dan mengulangi.

5.7.6 Confusion Matrix

Confusion matrix adalah suatu metode yang digunakan untuk menentukan akurasi,recall,precission,dan error rate yang dimana precission untuk mengevaluasi sistem untuk menemukan ranking yang relavan dan precision,dan juga didefinisikan sebagai presentase dokumen yang akan di retrieve dan juga relavan terhadap query.recall bertujuan untuk mengevaluasi kemampuan dari sistem untuk menemukan data yang relevan dari dokumen dan diartikan sebagai presentase dokumen yang relavan pada query.Accuracy adalh perbandingan kasus yang diidentifikasikan benar dengan seluruh jumlah kasus dan eror rate dengan seluruh jumlah kasus yang salah [23].

Tabel 1.1 Confusion Matrix

		PREDIKSI	
		POSITIVE	NEGATIVE
AKTUAL	POSITIVE	TP	FN
	NEGATIVE	FP	TN

Keterangan :

- True Positive adalah jumlah data positif yang diklasifikasikan dengan benar

- True Negative adalah jumlah data negatif yang diklasifikasikan dengan benar

- False Positive adalah jumlah data positif yang diklasifikasikan dengan salah.

- False Negative adalah jumlah data negatif yang diklasifikasikan dengan salah.

Rumus dari pencarian Akurasi :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Rumus 1.4 Accuracy

Rumus untuk mencari Precision :

$$\text{Precision} = \frac{TP}{TP + FP}$$

Rumus 1.5 Precision

Rumus untuk mencari Recall :

$$\text{Recall} = \frac{TP}{TP + FN}$$

Rumus 1.6 Recall

6. METODOLOGI PENELITIAN

6.1 Alat Dan Bahan Penelitian

6.1.1 Alat

Dalam penelitian ini, peneliti menggunakan alat bantu (*tools*) yang digunakan untuk melakukan pengolahan data/bahan penelitian yaitu :

a. Perangkat Keras (*Hardware*)

1. Laptop
   - Monitor : 1920x1080 *pixel*
   - Processor : Intel Core i7
   - RAM : 8 GB
   - Storage : *Harddisk* 1TB, dengan *freespaces* 368GB

b. Perangkat Lunak (*Software*)

1. Windows 10 Home 64 Bit
2. Microsoft Word
3. Microsoft Excel
4. Python
5. Visual Studio Code
6. Web browser (Google Chrome)
7. Lucid chart

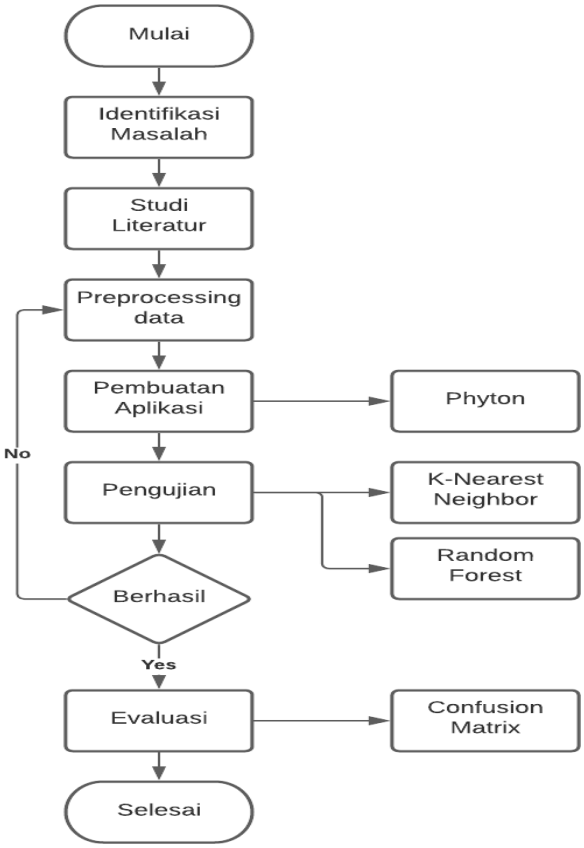
6.1.2 Bahan Penelitian

Bahan yang digunakan untuk penelitian Penyakit Kanker Paru adalah data-data yang diunduh dari https://www.kaggle.com/mysarahmadbhat/lung-cancer pada tanggal 14 September 2022.

6.2 Metode Penelitian

6.2.1 Kerangka Kerja Penelitian

Untuk membantu penelitian ini, diperlukan susunan kerangka kerja (*frame work*) yang jelas urutan dan tahapannya. Kerangka kerja dibawah ini merupakan langkah-langkah yang akan dilakukan dalam penyelesaian masalah yang akan dibahas. Berikut kerangka kerja yang akan peneliti gunakan :



Gambar 1.4 Kerangka Kerja Penelitian

Berdasarkan kerangka kerja penelitian diatas, maka dapat diuraikan pembahasan masing-masing tahapan dalam penelitian sebagai berikut :

1. **Indetifikasi Masalah**

Permasalahan dalam penelitian ini adalah bagaimana cara mengklasifikaskan penyakit kanker paru dengan menggunakan algoritma K-Nearest Neighbor dan Random Forest untuk menampilkan data yang terdeteksi YES dan NO menggunakan aplikasi berbasis web

2. **Studi Literatur**

Pada tahap ini penulis mencari data dari sebuah situs internet dan mencari informasi dengan membaca tiap jurnal atau buku yang berhubungan dengan klasifikasi penyakit kanker paru, metode K-Nearest Neighbor dan Random Forest yang akan menjadi acuan dalam pembuatan tugas akhir ini.

3. Pengumpulan Data

Penelitian ini menggunakan data sekunder, dataset penyakit kanker paru di ambil dari Kaggle komunitas daring ilmuwan data dan pembelajaran mesin dengan mengunduh di situs https://www.kaggle.com/mysarahmadbhat/lung-cancer_data data memiliki 16 fitur dan 310 baris, data memiliki 2 kelas yaitu, YES dan NO.

4. Metode Analisis Data

Metode analisa data yang digunakan penulis adalah metode K-Nearest Neighbor dan Random Forest. Pada metode K-Nearest Neighbor ini akan mengambil jarak terdekat dari data acuan sehingga data yang akan di uji dan diidentifikasi akan diketahui kelasnya dengan melihat jumlah yang paling banyak dari jarak yang paling dekat dan pada Random Forest diambil berdasarkan hasil voting dan tree uang dibuat. Pemenang dari tree yang terbentuk didapatkan dari vote yang paling banyak.

a. Preprocessing Data

Pada tahapan preprocessing ini penulis menggunakan data sekunder yang di dapat dari situs kaggle, data ini memiliki 16 fitur dan 310 baris, data memiliki 2 kelas yaitu, YES dan NO. Dataset mengandung oulier dan noise, sehingga harus dibersihkan pada tahap preprocessing. Tahap preprocessing meliputi estimasi missing value dan menghilangkan noise, seperti oulier, normalisasi, dan pengecekan data yang tidak sembang. Beberapa pengukuran mungkin terlewatkan sehingga menyebabkan nilai yang hilang. Pada data ini harus melewati tahap cleaning data, data yang kosong diisi dengan menggunanakan mean.

b. Pembuatan mesin pembelajaran K-NN

pembuatan mesin pembelajaran yang dibuat menggunakan bahasa pemrograman phyton dari metode K-Nearest Neighbor sebagai model untuk melakukan klasifikasi data

c. Evaluasi K-Nearest Neighbor

Setelah selesai membuat mesin pembelajaran K-NN, lanjutkan ke tahap evaluasi untuk menghitung performa dari metode K-Nearest Neighbor dengan K = 1 pada kelas YES dan NO dibutuhkannya confuson matrix.

d. Pembuatan mesin pembelajaran Random Forest

pembuatan mesin pembelajaran yang dibuat menggunakan bahasa pemrograman phyton dari metode Random Forest sebagai model untuk melakukan klasifikasi data.

e. Evaluasi Random Forest

Pada tahap evaluasi dibutuhkan untuk menghitung performa dari metode Random Forest pada kelas 1 dan 0 dibutuhkannya confusion matrix

5. Perancangan Sistem

Pada tahapan perancangan sistem, penulis menggunakan basis data dari bahasa pemrograman Python. Basis data yang digunakan yaitu untuk menyimpan data latih,menyimpan data pengujian dan menyimpan data hasil dari pengujian yang dilakukan.Untuk pembuatan mesin pembelajaran untuk melakukan proses pengujian menggunakan K-Nearest Neighbor dan aplikasi berbasis web untuk melakukan proses klasifikasi data dibuat dengan menggunakan bahasa pemrograman Python.

6. Penyusunan Laporan

Pada tahap ini peneliti menjelaskan tugas dan kegiatan yang telah dilakukan dengan merangkun hasil penelitian yang telah di lakukan ke dalam laporan tugas akhir dimulai dari indentifikasi masalah hingga sampai pada tahap pengembangan sistem yang telah selesai dirancang.

7. JADWAL PENELITIAN

Berikut jadwal waktu penelitian yang direncanakan peneliti berdasarkan kerangka kerja (***Frame Work***) yang telah disusun, yaitu dilaksanakan pada bulan **Oktober 2022** sampai dengan Februari **2023**. Penelitian dilakukan selama 5 bulan dengan perincian seperti yang tertulis pada tabel berikut ini :

No	Kegiatan	Oktober				November				Desember				Januari				Februari			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Identifikasi Masalah																				
2	Studi Literatur																				
3	Pengumpulan Data																				
4	Perancangan Sistem																				
5	Penyusunan Laporan																				

DAFTAR PUSTAKA

[1] Chen W, Zheng R, Zeng H, dan Zhang S. 2015. Eppidemiology of Lung Cancer in China. Thorac Cancer 6(2):209-15

[2] Lestariningsih Diah. 2010. Evaluasi Penatalaksanaan Mual Muntah Karena Kemoterapi Pada pasien Kanker Paru-paru di Instalasi Rawat Inap RSUD Dr. Moewardi Surakarta Tahun 2009. Fakultas Farmasi.Universitas Muhammadiyah : Surakarta.

[3] Diananda Rama. 2009. Mengenal Seluk Beluk Kanker. Penerbit Kata Hati : Yogyakarta

[4] Benazir, Salsabila. 2013. Faktor Risiko Kejadian Kanke Paru pada Pasien Rawat Inap dan Rawat Jalan di RSUPN Dr. Cipto Mangunkusumo Jakarta Tahun 2011-2012. Skripsi dipublikasikan. Skripsi tidak diterbitkan.Depok: Program Sarjana - Fakultas Kesehatan Masyarakat, Universitas Indonesia.

[5] Price S.A, wilson. (1982). Patofisiologi : Konsep Klinik Proses-proses Penyakit. Jakarta : EGC.

[6] Andiani, L., Sukemi, & Rini, D. P. (2019). Analisis Penyakit Jantung Menggunakan Metode KNN Dan Random Forest. Prosiding Annual Research Seminar 2019, 5(1), 978–979.

[7] Purba, Ardina Filindri. 2015. Pola Klinis Kanker Paru Di RSUP Dr. Kariadi Semarang Periode Juli 2013 – Juli 2014. Fakultas Kedokteran Universitas Diponegoro. Semarang.

[8] Russell, S. J., & Norvig, P. (1996). Artificial intelligence: A modern approach. Artificial Intelligence.

[9] Simon, A., Deo, M. S., Venkatesan, S., & Babu, R. (2020). An Overview of Machine Learning and its Applications. Machine Learning Concepts with Python and the Jupyter Notebook Environment, January, 21–39. https://doi.org/10.1007/978-1-4842-5967-2_2

[10] Ahmad, A. (2017). Mengenal Artificial Intelligence, Machine Learning, & Deep Learning.

[11] Retnosari, D. (2014). Sistem Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa. Jurnal Integrasi Sistem Industri UMJ, 1(2), 13–20.

[12] Dedi Ary Prasetya, I. N. (2012). Deteksi wajah metode viola jones pada opencv menggunakan pemrograman python. Simposium Nasional RAPI XI FT UMS, 18–23.

[13] Admojo, F. T., & Ahsanawati. (2020). Klasifikasi Aroma Alkohol Menggunakan Metode KNN. Indonesian Journal of Data and Science, 1(2), 34–38. https://doi.org/10.33096/ijodas.v1i2.12

[14] Jannah, M., & Humaira, N. (2019). Implementasi Metode Euclidean Distance Untuk Ekstraksi Fitur Jarak Pada Citra Skeleton. Jurnal Ilmiah Informatika Komputer, 24(2), 134–139. https://doi.org/10.35760/ik.2019.v24i2.2368

[15] B. M. Pereira et al., "The role of point-of-care ultrasound in intra-abdominal hypertension management," Anaesthesiol. Intensive Ther., vol. 49, no. 5, pp. 373–381, 2017, doi: 10.5603/AIT.a2017.0074

[16] Raju, K. S., Murty, M. R., Rao, M. V., & Satapathy, S. C. (2018). Support Vector Machine with K-fold Cross Validation Model for Software Fault Prediction. International Journal of Pure and Applied Mathematics, 118(20), 321–334. https://www.researchgate.net/publication/329414359_Support_Vector_Machine_with_K- fold_Cross_Validation_Model_for_Software_Fault_Prediction

[17] Lestari, M. (2014). Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) untuk Mendeteksi Penyakit Jantung. Faktor Exacta, 7(September 2010), 366–371.