BAB 2
LANDASAN TEORI

2.1 Tanaman Obat dan Senyawa Bioaktif

Tanaman obat merupakan sumber utama senyawa bioaktif yang memiliki potensi terapeutik, termasuk sebagai agen antikanker. Senyawa-senyawa ini dapat dikelompokkan ke dalam beberapa kelas utama, seperti flavonoid, alkaloid, terpenoid, dan senyawa fenolik, yang telah terbukti memiliki aktivitas biologis yang signifikan. Misalnya, flavonoid dikenal sebagai antioksidan kuat dan dapat memodulasi jalur sinyal seluler yang terlibat dalam pertumbuhan kanker. Alkaloid, di sisi lain, seringkali menunjukkan aktivitas sitotoksik dengan menginterferensi replikasi DNA atau mitosis sel kanker. Terpenoid juga telah diidentifikasi memiliki kemampuan untuk menginduksi apoptosis atau menghambat proliferasi sel kanker [3].

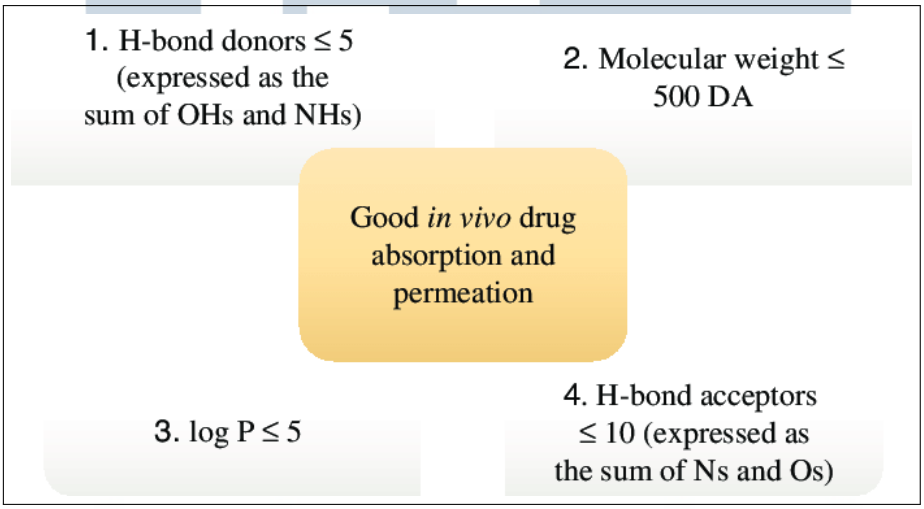
Dalam konteks penemuan obat, identifikasi senyawa bioaktif dari tanaman melalui pendekatan komputasi merupakan tahap awal yang krusial dalam drug discovery pipeline. Sistem klasifikasi seperti yang dikembangkan dalam penelitian ini berfungsi sebagai alat skrining awal untuk mempercepat penemuan kandidat senyawa potensial.

Penentuan potensi suatu senyawa sebagai agen terapi seringkali didasarkan pada sifat fisikokimia molekuler. Dalam penelitian ini, empat parameter utama digunakan sebagai fitur: Hydrogen Bond Donor (HBD), Hydrogen Bond Acceptor (HBA), massa molekul, dan logP. Parameter-parameter ini sangat relevan karena secara kolektif berkorelasi dengan ”Lipinski’s Rule of Five”, suatu pedoman empiris untuk menilai drug-likeness dan oral bioavailability suatu senyawa [12].

- Hydrogen Bond Donor (HBD): Mengacu pada jumlah gugus -OH atau -NH yang dapat mendonasikan ikatan hidrogen. HBD memengaruhi kemampuan senyawa untuk berinteraksi dengan target biologis dan kelarutan dalam air.
- Hydrogen Bond Acceptor (HBA): Mengacu pada jumlah atom oksigen atau nitrogen yang dapat menerima ikatan hidrogen. HBA juga krusial untuk interaksi molekuler dan kelarutan.
- Massa Molekul (Molecular Weight): Berat molekul senyawa. Senyawa dengan massa molekul yang terlalu besar (umumnya diatas 500 Dalton)

cenderung memiliki permeabilitas membran yang buruk, menyulitkan senyawa untuk mencapai target intraseluler.

- logP (Partition Coefficient): Mengukur kelarutan relatif suatu senyawa dalam fase minyak (oktanol) dibandingkan dengan fase air. Nilai logP yang seimbang (tidak terlalu tinggi atau terlalu rendah) penting untuk penyerapan, distribusi, metabolisme, dan ekskresi (ADME) obat dalam tubuh. Senyawa dengan logP terlalu tinggi sulit larut dalam air, sedangkan logP terlalu rendah sulit menembus membran lipid.



Gambar 2.1. Lipinkis rule

Adapun parameter lainnya yang digunakan sebagai fitur yaitu, molecular fingerprint.

- Molecular Fingerprint: Merupakan representasi digital dari struktur kimia suatu senyawa, umumnya dalam bentuk vektor biner (serangkaian angka 0 dan 1). Setiap posisi (bit) dalam vektor ini menandakan ada (nilai 1) atau tidaknya (nilai 0) suatu sub-struktur atau fragmen kimia tertentu. Dibandingkan dengan deskriptor fisikokimia yang hanya memberikan beberapa nilai ringkasan, fingerprint mampu menangkap ribuan detail struktural. Hal ini memberikan representasi yang jauh lebih kaya dan informatif bagi model machine learning untuk dapat mengenali pola-pola kompleks yang membedakan senyawa aktif dan non-aktif.

2.2 Kanker dan Mekanisme Penyembuhannya

Kanker paru-paru adalah penyakit yang ditandai oleh pertumbuhan sel-sel yang tidak terkendali di jaringan paru-paru, yang dapat menyebar (metastasis) ke bagian tubuh lain. Terdapat dua jenis utama kanker paru-paru: Non-Small Cell Lung Cancer (NSCLC) yang lebih umum, dan Small Cell Lung Cancer (SCLC) yang lebih agresif. Mekanisme penyembuhan kanker melibatkan berbagai pendekatan, termasuk kemoterapi, radioterapi, dan terapi target molekuler. Terapi target molekuler, seperti penghambatan reseptor EGFR (Epidermal Growth Factor Receptor) telah menjadi fokus dalam pengembangan obat antikanker modern. Senyawa alami yang dapat berinteraksi dengan target ini memiliki potensi sebagai obat terapi baru. Senyawa alami yang berpotensi menjadi senyawa antikanker bekerja melalui berbagai mekanisme seluler, antara lain:

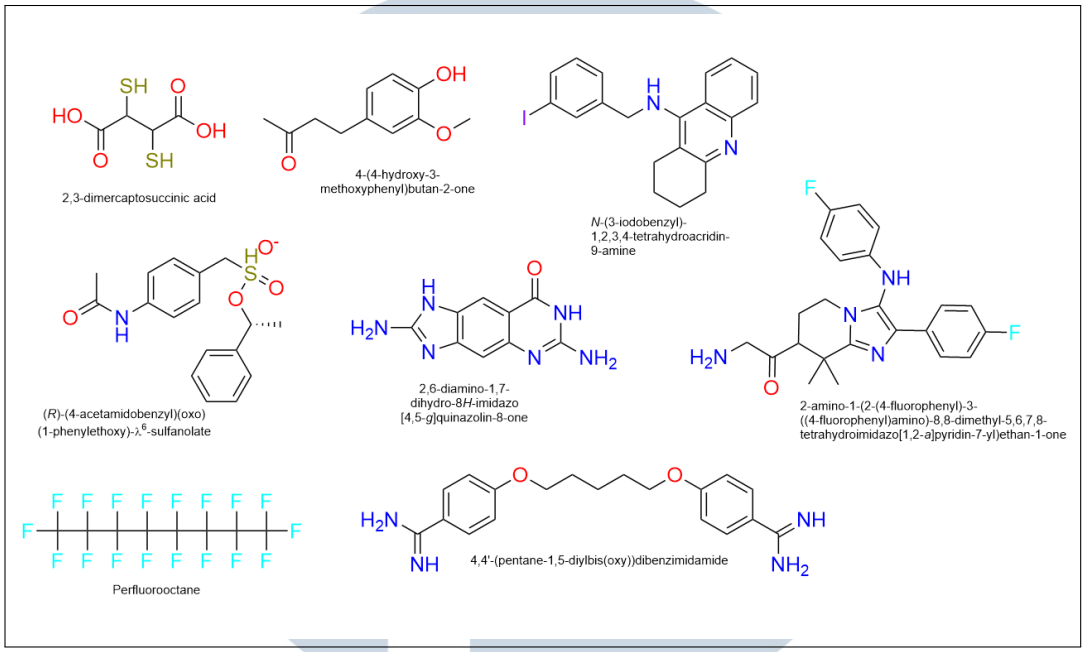
1. Menghambat angiogenesis (pembentukan pembuluh darah baru untuk sel kanker): Angiogenesis adalah proses vital bagi pertumbuhan tumor karena menyediakan nutrisi dan oksigen. Senyawa antikanker dapat memblokir proses ini, secara efektif ”membuat kelaparan” sel kanker.

2. Memicu apoptosis pada sel kanker: Apoptosis adalah kematian sel terprogram yang sehat. Sel kanker seringkali menghindari apoptosis. Senyawa antikanker dapat menginduksi kembali jalur apoptosis, menyebabkan kematian selektif pada sel kanker.

3. Mengganggu siklus pembelahan sel kanker: Sel kanker tumbuh dan membelah dengan cepat. Senyawa antikanker dapat menghentikan atau memperlambat siklus pembelahan sel pada fase tertentu, sehingga mencegah proliferasi tumor.

Senyawa yang memiliki nilai HBA dan HBD tertentu diketahui dapat berinteraksi secara spesifik dengan target protein kanker, membentuk ikatan hidrogen yang stabil. Selain itu, massa molekul dan logP juga memengaruhi kelarutan dan kemampuan senyawa untuk menembus membran sel dan mencapai target intraseluler. Namun, untuk menangkap kompleksitas struktur kimia secara lebih menyeluruh, molecular fingerprint digunakan sebagai representasi fitur yang lebih kaya. Berbeda dengan deskriptor tunggal, fingerprint mengubah struktur molekul menjadi vektor biner (serangkaian angka 0 dan 1), di mana setiap bit menandakan ada atau tidaknya sub-struktur kimia tertentu. Representasi detail ini

memungkinkan model *machine learning* untuk mengenali pola-pola struktural yang lebih kompleks dan spesifik yang berkorelasi dengan aktivitas antikanker.



Gambar 2.2. Gambar molekul dalam senyawa tanaman

2.2.1 Peran EGFR dalam Kanker Paru-Paru

Salah satu target molekuler yang paling penting dalam pengembangan terapi kanker paru-paru adalah EGFR (Epidermal Growth Factor Receptor). EGFR adalah protein reseptor tirosin kinase yang terdapat pada permukaan sel dan memainkan peran krusial dalam regulasi jalur sinyal seluler yang mengontrol pertumbuhan, proliferasi, dan kelangsungan hidup sel.

Pemilihan EGFR sebagai target molekuler dalam penelitian ini didasarkan pada perannya yang sudah tervalidasi secara klinis sebagai pendorong utama dalam banyak kasus kanker paru-paru. Berbagai tinjauan sistematis internasional telah memetakan tingginya frekuensi mutasi EGFR secara global, terutama pada populasi Asia. Lebih lanjut, Organisasi Kesehatan Dunia (WHO) dalam klasifikasi tumor paru terbarunya telah menjadikan pengujian mutasi EGFR sebagai standar dalam diagnosis. Hal ini juga sejalan dengan Pedoman Nasional Pelayanan Kedokteran (PNPK) Tata Laksana Kanker Paru dari Kementerian Kesehatan RI tahun 2023, yang merekomendasikan pemeriksaan mutasi EGFR untuk menentukan strategi terapi target bagi pasien di Indonesia [13, 14].

Dalam kondisi normal, aktivitas EGFR diatur secara ketat. Namun, pada sebagian besar kasus kanker paru-paru jenis Non-Small Cell Lung Cancer (NSCLC), terjadi mutasi pada gen yang mengkode EGFR. Mutasi ini menyebabkan protein EGFR menjadi hiperaktif secara terus-menerus, bahkan tanpa adanya sinyal pertumbuhan dari luar. Aktivitas yang tidak terkendali inilah yang menjadi salah satu pendorong utama di balik proliferasi sel kanker yang agresif, metastasis, dan resistansi terhadap kemoterapi.

Karena perannya yang sentral dalam mendorong pertumbuhan tumor, penghambatan aktivitas EGFR telah menjadi salah satu strategi terapi target (targeted therapy) yang paling berhasil dalam pengobatan kanker paru-paru. Senyawa yang dapat bertindak sebagai inhibitor EGFR berpotensi untuk secara efektif memblokir jalur sinyal abnormal ini, sehingga dapat secara selektif menghambat pertumbuhan sel kanker dan memicu apoptosis (kematian sel terprogram). Dengan demikian, menargetkan EGFR merupakan pendekatan yang sangat valid dan relevan secara klinis dalam upaya penelitian untuk menemukan kandidat obat baru untuk kanker paru-paru, yang menjadi dasar dari pengumpulan data dan pemodelan dalam penelitian ini [15].

2.3 Machine Learning dan Data Mining

Machine Learning (ML) dan Data Mining telah menjadi alat yang sangat penting dalam bidang penemuan obat modern, khususnya dalam proses identifikasi dan klasifikasi senyawa bioaktif dari sumber alami. Dengan kemampuan untuk mengolah dan menganalisis data dalam jumlah besar, algoritma ML memungkinkan peneliti untuk menemukan pola yang kompleks dalam data kimia dan biologis, seperti struktur molekul, sifat fisikokimia, dan aktivitas biologis terhadap target tertentu.

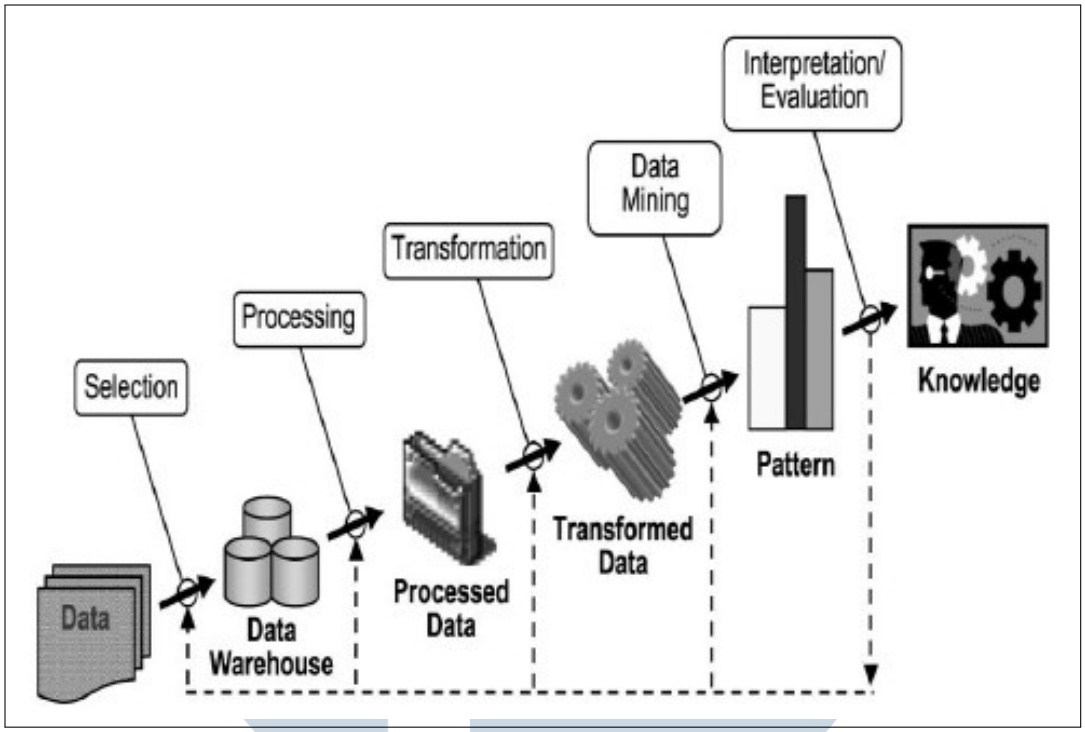
Proses data mining umumnya melibatkan beberapa tahapan:

1. Pengumpulan Data: Mengumpulkan data mentah dari berbagai sumber.
2. Pra-pemrosesan Data: Membersihkan, mengintegrasikan, dan mengubah data ke format yang sesuai untuk analisis.
3. Transformasi Data: Menerapkan teknik untuk mengubah data menjadi format yang cocok untuk algoritma ML, seperti normalisasi atau pemilihan fitur.

4. Pemodelan: Menggunakan algoritma ML untuk membangun model berdasarkan data yang telah diproses.
5. Evaluasi Model: Menilai kinerja model menggunakan metrik yang relevan.
6. Penyebaran: Mengimplementasikan model yang telah dievaluasi ke dalam aplikasi praktis.

Studi menunjukkan bahwa algoritma seperti Random Forest mampu memberikan prediksi yang akurat terhadap aktivitas antikanker dari senyawa tanaman. Dalam penelitian yang dilakukan, Random Forest menunjukkan performa terbaik dalam hal akurasi dan stabilitas prediksi dibandingkan algoritma lain dalam mengidentifikasi aktivitas antikanker senyawa tanaman. Hal ini menegaskan bahwa metode berbasis pembelajaran mesin sangat potensial dalam mengakselerasi proses drug discovery.[5]

Selain itu, penelitian mendukung efektivitas pendekatan machine learning dalam pencarian obat, di mana mereka menerapkan virtual screening berbasis ML dan pemodelan molekuler untuk mengidentifikasi senyawa alami yang berpotensi menghambat kanker paru-paru non-sel kecil (NSCLC). Studi tersebut memperlihatkan bahwa kombinasi data mining dan pemodelan komputasi mampu mempercepat proses identifikasi inhibitor alami dengan akurasi tinggi. Dengan demikian, penerapan machine learning dan data mining memberikan kontribusi signifikan dalam pengembangan obat berbasis senyawa alami, khususnya dalam konteks terapi kanker paru-paru. [4]



Gambar 2.3. Tahapan data mining

2.4 Algoritma Random Forest

Random Forest adalah algoritma ensemble learning yang kuat, menggabungkan beberapa pohon keputusan (decision trees) untuk meningkatkan akurasi dan stabilitas prediksi. Algoritma ini pertama kali diperkenalkan oleh Leo Breiman pada tahun 2001[16]. Random Forest bekerja berdasarkan prinsip bagging (Bootstrap Aggregating) dan metode subruang acak (random subspace method).[10]

2.4.1 Mekanisme Pemisahan Node pada Pohon Keputusan

Unit fundamental yang membangun algoritma Random Forest adalah Pohon Keputusan (Decision Tree). Mekanisme kerja inti dari sebuah pohon keputusan adalah proses pemisahan node (node splitting) secara rekursif. Prinsip dasar dari setiap pemisahan pada sebuah node adalah untuk mempartisi data menjadi dua himpunan bagian (subset) yang lebih homogen atau murni (pure) secara komposisi kelas dibandingkan dengan himpunan data awal.

Untuk mengukur tingkat kemurnian atau homogenitas sebuah himpunan data, metrik matematis seperti Gini Impurity atau Information Gain (berbasis Entropi) digunakan. Algoritma akan mencari kriteria pemisahan yang dapat memaksimalkan penurunan tingkat impuritas (impurity) setelah data dipisah.

Pada kasus fitur numerik, seperti HBD, HBA, Massa Molekul, LogP dan molecular fingerprint yang digunakan dalam penelitian ini, algoritma akan secara iteratif mencari kombinasi fitur dan nilai ambang batas (threshold) yang optimal. Proses ini melibatkan pengujian berbagai nilai threshold pada suatu fitur. Untuk setiap threshold yang diuji, data akan dibagi menjadi dua cabang: satu cabang untuk data yang nilainya kurang dari atau sama dengan threshold (misalnya, LogP lebih kecil dari sama dengan 4.5), dan cabang lainnya untuk data yang nilainya lebih besar dari threshold (LogP lebih besar dari 4.5).

Kombinasi fitur dan threshold yang menghasilkan penurunan Gini Impurity terbesar (atau Information Gain tertinggi) akan dipilih sebagai aturan keputusan (splitting rule) pada node tersebut. Proses ini berlanjut di setiap node turunan hingga kriteria pemberhentian terpenuhi, seperti tercapainya kedalaman maksimum pohon atau jumlah sampel minimum pada node.

2.4.2 Cara Kerja Algoritma Random Forest

Berikut adalah cara kerja algoritma Random Forest:

1. Pembuatan Sampel Bootstrap: Dari dataset pelatihan asli, Random Forest secara acak mengambil beberapa bootstrap sample (sampel dengan penggantian). Setiap sampel ini digunakan untuk melatih satu pohon keputusan. Karena pengambilan sampel dengan penggantian, beberapa data mungkin muncul berkali-kali, sementara yang lain mungkin tidak muncul sama sekali di satu sampel.
2. Pemilihan Fitur Acak: Saat membangun setiap pohon keputusan, pada setiap node (titik pemisahan), bukan semua fitur yang dipertimbangkan untuk pemisahan terbaik. Random Forest hanya memilih subset fitur acak. Ini mengurangi korelasi antar pohon dan meningkatkan keragaman ensemble.
3. Pembangunan Pohon Keputusan: Setiap pohon dibangun secara maksimal tanpa pruning.

4. Prediksi (Voting): Untuk masalah klasifikasi, setiap pohon dalam hutan membuat prediksinya sendiri. Prediksi akhir ditentukan oleh suara mayoritas dari semua pohon. Untuk masalah regresi, prediksi akhir adalah rata-rata dari prediksi semua pohon.

2.4.3 Kelebihan dari Algoritma Random Forest

Berikut adalah kelebihan dari algoritma Random Forest:

1. Mampu Menangani Data Numerik dan Kategorikal: Random Forest secara inheren dapat bekerja dengan berbagai tipe data tanpa perlu pra-pemrosesan yang rumit untuk konversi tipe data.
2. Tahan Terhadap Overfitting: Karena menggunakan ensemble dari banyak pohon yang dibangun dari subset data dan fitur acak, Random Forest cenderung tidak overfit terhadap data pelatihan, sehingga memiliki kemampuan generalisasi yang baik.
3. Memberikan Informasi tentang Pentingnya Fitur (Feature Importance): Random Forest dapat menghitung seberapa besar kontribusi setiap fitur terhadap akurasi prediksi. Ini sangat berguna untuk memahami fitur mana yang paling relevan dalam dataset.
4. Mampu Menangani Data dengan Dimensi Tinggi: Efektif untuk dataset dengan banyak fitur.

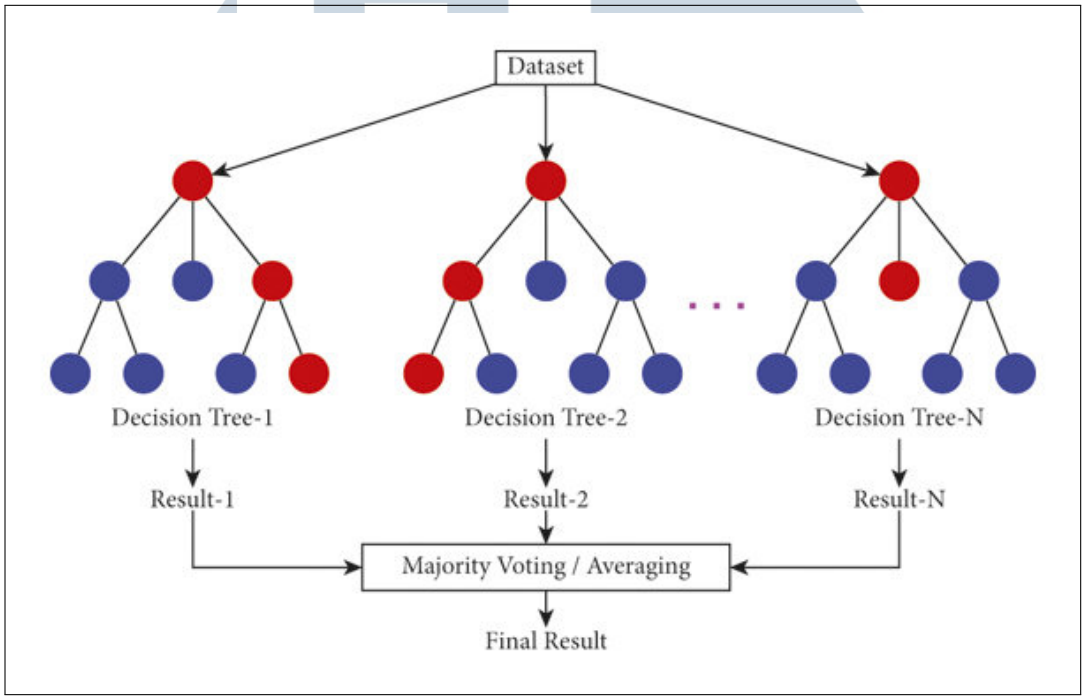
2.4.4 Kekurangan Algoritma Random Forest

Berikut adalah kekurangan dari algoritma Random Forest:

1. Interpretasi yang Sulit: Meskipun memberikan feature importance, model Random Forest secara keseluruhan lebih sulit diinterpretasikan (seperti ”kotak hitam”) dibandingkan dengan satu pohon keputusan sederhana.
2. Komputasi yang Lebih Intensif: Membangun dan melatih banyak pohon membutuhkan waktu dan sumber daya komputasi yang lebih besar dibandingkan algoritma tunggal.

Dalam konteks penelitian ini, Random Forest digunakan untuk mengklasifikasikan senyawa dalam tanaman menjadi dua kategori: berpotensi

sebagai obat kanker paru-paru atau tidak berpotensi, berdasarkan fitur kimia seperti HBA, HBD, massa molekul, LogP dan *molecular fingerprint*. *Random Forest* dipilih karena kemampuannya dalam menangani data dengan dimensi tinggi dan mengurangi *overfitting*, serta kinerjanya yang stabil dan akurat dalam klasifikasi tanaman obat.



Gambar 2.4. Diagram random forest

2.5 Evaluasi Model Klasifikasi

Evaluasi model klasifikasi sangat penting untuk memastikan keandalan, akurasi, dan kemampuan generalisasi model dalam memprediksi kelas baru. Metode evaluasi umum meliputi Akurasi, Presisi, Recall, dan F1-Score. Sebelum membahas metrik-metrik tersebut, penting untuk memahami Confusion Matrix, yang menjadi dasar perhitungannya.

Confusion Matrix:

- Confusion Matrix adalah tabel yang menunjukkan kinerja algoritma klasifikasi. Tabel ini membandingkan prediksi model dengan nilai aktual/sebenarnya.
- True Positive (TP): Jumlah kasus positif yang diprediksi dengan benar

sebagai positif. (Dalam konteks Anda: Senyawa potensial yang benar diprediksi potensial).

- True Negative (TN): Jumlah kasus negatif yang diprediksi dengan benar sebagai negatif. (Dalam konteks Anda: Senyawa tidak potensial yang benar diprediksi tidak potensial).
- False Positive (FP): Jumlah kasus negatif yang salah diprediksi sebagai positif (Kesalahan Tipe I). (Dalam konteks Anda: Senyawa tidak potensial yang salah diprediksi potensial).
- False Negative (FN): Jumlah kasus positif yang salah diprediksi sebagai negatif (Kesalahan Tipe II). (Dalam konteks Anda: Senyawa potensial yang salah diprediksi tidak potensial).

Tabel 2.1. Tabel confusion matrix

Prediksi / Aktual	Positif (Diprediksi Positif)	Negatif (Diprediksi Negatif)
Positif (Aktual Positif)	True Positif (TP)	False Negative (FN)
Negatif (Aktual Negatif)	False Positive (FP)	True Negative (TN)

Metrik Evaluasi:

1. Akurasi (Accuracy): Merupakan proporsi prediksi yang benar dari total jumlah prediksi. Akurasi mengukur seberapa sering model membuat prediksi yang benar secara keseluruhan.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Akurasi cocok digunakan ketika jumlah sampel setiap kelas relatif seimbang.

2. Presisi (Precision): Mengukur seberapa banyak dari kasus yang diprediksi positif, benar-benar positif. Presisi penting ketika biaya dari false positive sangat tinggi.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

Dalam konteks Anda, presisi tinggi untuk kelas ”potensial” berarti sebagian besar senyawa yang diprediksi sebagai antikanker memang benar-benar memiliki potensi tersebut, mengurangi false alarm.

3. Recall (Sensitivity): Mengukur seberapa banyak dari kasus positif yang sebenarnya berhasil dideteksi oleh model. Recall penting ketika biaya dari false negative sangat tinggi.

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

Dalam konteks Anda, recall tinggi untuk kelas ”potensial” berarti model berhasil menemukan sebagian besar senyawa yang sebenarnya berpotensi antikanker, mengurangi missed opportunities.

4. F1-Score: Merupakan rata-rata harmonik dari Presisi dan Recall. F1-Score memberikan keseimbangan antara Presisi dan Recall, dan sangat berguna terutama pada dataset yang tidak seimbang *(imbalanced dataset)*.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

5. ROC Curve dan AUC (Receiver Operating Characteristic Curve and Area Under the Curve):

   - ROC Curve adalah plot yang menggambarkan kinerja model klasifikasi pada semua ambang batas klasifikasi. Kurva ini memplot True Positive Rate (Recall) terhadap False Positive Rate (FP / (FP + TN)) pada berbagai threshold.
   - AUC adalah area di bawah kurva ROC. Nilai AUC berkisar antara 0.5 (kinerja acak) hingga 1.0 (kinerja sempurna). AUC adalah metrik yang robust untuk dataset tidak seimbang dan memberikan gambaran kinerja model secara keseluruhan.

Studi menyatakan Random Forest menunjukkan akurasi tertinggi dalam memprediksi aktivitas antikanker senyawa tanaman. [5] Evaluasi yang tepat memungkinkan peneliti untuk memilih model terbaik dan mengidentifikasi area yang memerlukan perbaikan.