DAFTAR ISTILAH

Istilah		Keterangan	Halaman pertama
Data mining	:	Proses pengolahan data dari suatu data yang besar untuk mendapatkan informasi penting pada data.	2
Machine Learning	:	Pengembangan mesin yang dapat digunakan untuk bisa belajar dengan sendirinya tanpa arahan dari pengguna.	2
Database	:	Sekumpulan data yang disimpan secara sistematis.	2
Knowledge Discovery in Database	:	Proses untuk menggali dan menganalisis suatu kumpulan data dan menghasilkan informasi yang dapat digunakan.	2
Artificial Intelligence	:	Teknologi baru yang memanfaatkan kecerdasan buatan yang ditambahkan pada suatu sistem.	3
Ensemble	:	Metode kombinasi dari beberapa algoritma atau model.	3
CART	:	*Classification and Regression Tree* , metode eksplorasi data dengan pohon keputusan dari.	12
SMOTE	:	*Synthetic Minority Oversampling Technique* metode untuk menangani kelas tidak seimbang.	13
ROC	:	Grafik untuk mengukur performa klasifikasi dalam menentukan ambang batas dari suatu model.	15
AUC	:	Luas area di bawah kurva ROC.	15

BAB I PENDAHULUAN

I.1 Latar Belakang

Peningkatan populasi perkotaan yang terus bertumbuh lebih padat atau lebih besar yang mencapai 20 juta penduduk di beberapa kota besar membuat kota-kota sangat rentan terhadap bahaya perubahan iklim. Selain meningkatnya populasi perkotaan, perubahan iklim juga disebabkan oleh adanya emisi gas rumah kaca yang menyebabkan peningkatan panas atomsfer. Pemanasan atmosfer menyebabkan meningkatnya degradasi hujan, polusi udara dan juga air (Rojas-Downing dkk., n.d.,).

Indonesia merupakan wilayah yang hampir seluruhnya memiliki iklim tropis. Mengacu pada variabilitas iklim di Indonesia dari BMKG, suhu wilayah Indonesia memiliki rata-rata sekitar 28ºC di wilayah pesisir, 26ºC di daerah pedalaman, dan 23ºC untuk wilayah dataran tinggi. Dengan adanya perubahan iklim sendiri berdampak pada perubahan anomali suhu udara rata-rata di Indonesia. Menurut (Molle & Larasati, 2021) sebagai daerah yang kebanyakan memiliki iklim tropis Indonesia mengalami variasi suhu yang sempit namun variasi curah hujan yang beragam. Perubahan iklim yang terjadi juga akan berpotensi mengakibatkan perubahan curah hujan. Timbulnya variasi hujan di Indonesia juga dipengaruhi oleh berbagai faktor baik lokal maupun global.

Curah hujan adalah jumlah total presipitasi yang terjadi di suatu daerah tertentu, yang diukur dalam skala harian, mingguan, bulanan, dan tahunan, yang dipengaruhi oleh faktor-faktor tambahan dan merupakan salah satu komponen dari iklim. Di Indonesia, curah hujan memiliki tingkat keberagaman yang signifikan dan merupakan hal yang paling penting dalam kehidupan manusia karena terkait dengan cuaca lain seperti kelembaban, tekanan, suhu, arah dan kecepatan angin (Roqyah dkk., 2023). Menurut (Indra Pratama dkk., 2022) Indonesia sendiri memiliki salah satu daerah yang memiliki curah hujan paling tinggi dengan intensitas curah hujan mencapai 4500mm, daerah tersebut merupakan kota Bogor yang juga dikenal dengan sebutan Kota Hujan. Berdasarkan data yang dikumpulkan oleh BNPB, terjadi sebanyak 2.342 kejadian

bencana di Indonesia pada tahun 2016. Sebanyak 92% dari kejadian tersebut mayoritasnya adalah bencana hidrometeorologi seperti banjir, longsor, dan puting beliung. Secara khusus, terdapat 766 kejadian banjir, 612 kejadian longsor, dan 74 kejadian gabungan banjir dan longsor. Curah hujan yang tinggi di wilayah terdampak menjadi penyebab utama dari banjir dan tanah longsor tersebut. Tragedi-tragedi ini memiliki dampak serius yang tidak boleh diabaikan. Dampak bencana alam yang terjadi menyebabkan korban jiwa, masalah kesehatan, kerusakan rumah, dan penghancuran gedung-gedung umum. Sebanyak 3,05 juta jiwa mengungsi dan menderita, 522 orang kehilangan nyawa atau dinyatakan meninggal dunia, 69.287 rumah mengalami kerusakan, dan 2.311 bangunan fasilitas umum hancur (Setiawan, 2021). Oleh karena itu, diperlukan teknik yang dapat dilakukan untuk melakukan klasifikasi terhadap curah hujan yang dapat digunakan sebagai pencegahan bencana karena intensitas curah hujan yang tinggi. Salah satu metode yang dapat digunakan untuk melakukan analisa dan pengolahan data agar dapat bermanfaat bagi pembaca ialah metode *data mining.*

Menurut (Jackson, 2002) *Data mining* merupakan proses untuk mengidentifikasi dan menghasilkan suatu data yang berguna dan berkolasi yang dapat dimengerti. Proses keseluruhan dari *data mining* sendiri disebut dengan *Knowledge Discovery in Database (KDD)*. KDD memiliki beberapa tahapan proses berupa persiapan, pemilihan, pembersihan dan interpretasi dari hasil proses *data mining*. *Data mining* merupakan proses untuk menemukan korelasi, pola dan tren baru yang berguna dengan melakukan penyaringan sejumlah besar data yang disimpan menggunakan teknologi pengenalan pola serta statistik. *Data mining* merupakan bidang interdisipliner yang menyatukan teknik dari *machine learning*, pengenalan pola, statistik, *database* dan visualisasi untuk mengatasi masalah ekstraksi informasi dari *database* yang besar (Larose, 2005). Sedangkan menurut (Jollyta dkk., 2020) *data mining* merupakan proses untuk menggali informasi dan pengetahuan baru dari suatu data yang berjumlah banyak pada database dengan menggunakan *Artificial Intelligence (AI)*, statistik dan matematika. *Data mining* merupakan teknologi yang dapat menjembatani komunikasi antara data dan kebutuhan pemakai data tersebut.

Dalam penerapannya *data mining* memiliki beberapa metode algoritma yang dapat digunakan diantaranya algoritma random forest dan naïve bayes. random forest merupakan sebuah algoritma klasifikasi yang termasuk dalam kategori *ensemble*. Algoritma ini membangun sebuah hutan (*forest*) yang terdiri dari beberapa pohon keputusan (*decision tree*). Dalam melakukan klasifikasi, random forest akan menggunakan teknik *voting* untuk mengambil keputusan dari seluruh pohon keputusan. Dengan menggunakan banyak pohon keputusan, algoritma random forest dapat mengatasi masalah yang muncul saat hanya menggunakan satu pohon keputusan dalam klasifikasi. Hal ini dapat meningkatkan nilai akurasi secara keseluruhan dan membuat hasil klasifikasi lebih optimal (Kusumarini dkk., 2021). Sedangkan algoritma naïve bayes merupakan sebuah metode klasifikasi yang didasarkan pada teknik-teknik probabilitas dan statistika yang diciptakan oleh seorang ilmuwan asal Inggris bernama Thomas Bayes. Tujuan dari algoritma ini adalah untuk memprediksi hasil yang akan terjadi di masa depan berdasarkan hasil yang terjadi di masa lalu, sehingga dinamai Teorema Bayes. Teori tersebut kemudian dipadukan dengan kata "naive", yang mengasumsikan situasi di mana setiap atribut memiliki sifat saling bebas (Rifai dkk., 2019).

Dalam penelitian yang berjudul "Perbandingan Metode Random Forest Dan Naïve Bayes Dalam Prediksi Keberhasilan Klien Telemarketing" oleh (Leonardo dkk., 2020) mengenai prediksi terhadap keputusan klien untuk membantu kinerja telemarketing, mendapatkan hasil perbandingan algoritma naive bayes dan random forest yaitu akurasi naive bayes sebesar 85% dan random forest 90. Dengan hasil tersebut dapat disimpulkan bahwa pada penelitian yang dilakukan oleh Leonardo dkk terhadap prediksi keberhasilan klien telemarketing, algoritma random forest lebih baik dalam melakukan klasifikasi tersebut.

Berdasarkan latar belakang yang sudah dijelaskan di atas, pada penelitian kali ini akan melakukan analisa perbandingan terhadap penggunaan metode algoritma random forest dan naïve bayes dengan memanfaatkan data mengenai iklim harian di Indonesia untuk melakukan klasifikasi terhadap curah hujan. Hasil dari penelitian ini bertujuan untuk membandingkan kedua metode algoritma dalam proses pengolahan data dan dan mencari hasil akurasi terbaik dari masing-masing

algoritma. Selain itu, dari adanya kasus bencana alam akibat curah hujan tinggi dan melihat kondisi intensitas hujan yang tinggi di wilayah Indonesia, diperlukan adanya pengetahuan yang mendukung untuk mencegah bahaya yang ditimbulkan. Data curah hujan yang telah diproses dengan cermat diharapkan dapat menjadi informasi yang memiliki manfaat yang signifikan. Informasi ini akan sangat berarti bagi berbagai pihak, termasuk masyarakat umum yang sering kali mengabaikan informasi penting tentang potensi bahaya bencana alam. Selain itu, terutama bagi organisasi atau lembaga yang terlibat langsung dalam upaya pencegahan bencana alam di Indonesia, seperti Badan Nasional Penanggulangan Bencana, informasi tersebut dapat digunakan sebagai dasar untuk merencanakan tindakan pencegahan atau mitigasi bencana guna mengurangi risiko dan dampak yang mungkin terjadi. Informasi ini dapat menjadi acuan yang berharga dalam mengetahui pola curah hujan dan mengambil langkah-langkah yang diperlukan untuk menghindari bencana akibat curah hujan tinggi di daerah-daerah rawan bencana di Indonesia. Selain melakukan persiapan terhadap bencana yang ditimbulkan, informasi terkait curah hujan juga dapat digunakan untuk berbagai keperluan seperti pada bidang pertanian, transportasi dan industri.

I.2 Perumusan Masalah

Berdasarkan latar belakang di atas, maka permasalahan yang akan dibahas dalam penelitian ini adalah sebagai berikut :

1. Bagaimana hasil analisa klasifikasi iklim Indonesia menggunakan algoritma random forest?
2. Bagaimana hasil analisa klasifikasi iklim Indonesia menggunakan algoritma naïve bayes?
3. Bagaimana hasil perbandingan analisa klasifikasi iklim Indonesia menggunakan algoritma random forest dan naïve bayes?

I.3 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mendapatkan hasil analisa dari klasifikasi iklim di Indoneisa menggunakan algoritma random forest.

2. Mendapatkan hasil analisa dari klasifikasi iklim di Indoneisa menggunakan algoritma naïve bayes.
3. Mengetahui perbandingan analisa dari klasifikasi iklim di Indoneisa dengan menggunakan algoritma random forest dan naïve bayes.

I.4 Batasan Penelitian

Adapun batasan dalam melakukan penelitian ini adalah sebagai berikut:

1. Penelitian berfokus pada data yang digunakan yaitu *Climate Data Daily IDN* yang didapatkan dari Kaggle dengan rentang waktu tahun 2020 dengan atribut Tavg, RH_avg, ss, ddd_x dan kategori curah hujan (label).
2. Penelitian ini bertujuan untuk melakukan klasifikasi curah hujan berdasarkan data iklim di Indonesia dengan menjadi kategori ringan, sedang, lebat dan sangat lebat.
3. Proses perbandingan klasifikasi menggunakan algoritma random forest dan naïve bayes menggunakan Google Colab.
4. Proses penelitian yang dilakukan hanya sampai tahapan perbandingan analisa algoritma yang didapatkan dalam klasifikasi berdasarkan analisa performa.

I.5 Manfaat Penelitian

Melalui penelitian yang dilakukan, diharapkan hasil yang didapatkan dapat memberi manfaat teoritis dengan mendapatkan hasil analisa klasifikasi iklim Indonesia bagi mahasiswa dan perbandingan mengenai penerapan algoritma yang lebih baik. Selain itu, bermanfaat bagi peneliti untuk mengimplementasikan serta mengetahui *process mining* pada klasifikasi iklim Indonesia dengan memanfaatkan algoritma yang digunakan yaitu algoritma random forest dan naïve bayes hingga mendapatkan hasil perbandingan dari kedua algoritma tersebut.

I.6 Sistematika Penulisan

Sistematika penulisan yang akan digunakan dalam penulisan penelitian, diuraikan dalam pembahasan mengenai setiap bab sebagai berikut,

Bab I Pendahuluan

Pada BAB I Pendahuluan, berisi mengenai latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan permasalahan dan sistematikan penulisan.

Bab II Tinjauan Pustaka

Pada Bab II Tinjauan Pustaka, berisi mengenai studi literatur, teori yang digunakan dan referensi dari penelitian terdahulu yang berkaitan dengan permasalahan penelitian.

Bab III Metodologi Penelitian

Pada Bab III Metodologi Penelitian, berisi tentang metode yang akan digunakan peneliti untuk melakukan penelitian, pada penelitian kali ini yaitu klasifikasi dengan menggunakan algoritma naïve bayes dan random forest. Selain itu juga berisi tentang sistematika penyelesaian masalah dan metode evaluasi yang akan digunakan.

Bab IV Analisis dan Perancangan

Pada Bab IV Analisis dan Perancangan, berisi tentang analisis yang dilakukan dimulai dari proses pengumpulan data, eksplorasi data, pembersihan data, *labeling* data, *balancing* data, penerapan algoritma untuk klasifikasi pada dataset iklim di Indonesia dan evaluasi yang akan digunakan.

Bab V Evaluasi

Pada Bab V Evaluasi, berisi tentang hasil akurasi dan perbandingan algoritma naïve bayes dan random forest untuk klasifikasi curah hujan berdasarkan dataset iklim di Indonesia.

Bab VI Penutup

Pada Bab VI, berisi tentang hasil evaluasi yang telah dilakukan, lalu dirangkum menjadi kesimpulan dan saran.

BAB II TINJAUAN PUSTAKA

II.1 Iklim di Indonesia

Indonesia merupakan negara yang dilewati oleh garis khatulistiwa, hal itu menyebabkan hampir di seluruh wilayah Indonesia memiliki iklim tropis. Selain iklim tropis yang kebanyakan dimiliki oleh wilayah Indonesia, iklim di Indonesia sendiri memiliki iklim muson dan iklim laut. Dengan sebagian besar wilayah yang memiliki iklim tropis menjadikan Indonesia memiliki suhu yang cenderung hangat bahkan cenderung panas dan juga lembab. Iklim tropis yang dimiliki sebagian besar wilayah Indonesia juga menjadikan Indonesia memiliki 2 musim, yaitu musim panas dan musim hujan.

II.2 Curah Hujan

Pola hujan di Indonesia menurut (RAHAYU dkk., 2018) dapat dibagi menjadi tiga wilayah yaitu wilayah A, wilayah B, dan wilayah C. Wilayah tersebut digunakan untuk mengkategorikan pola curah hujan, masing-masing dengan karakteristik mereka sendiri. Wilayah A memiliki puncak musim hujan dan musim kemarau. Karena migrasi Zona Konvergensi Intertropis (ZKI) ke selatan dan utara, wilayah B mengalami dua puncak musim hujan pada bulan Oktober hingga November dan Maret hingga Mei. Sedangkan di wilayah C puncak hujan berlangsung dari Juni hingga Juli dan puncak musim kemarau berlangsung dari November hingga Februari. Kriteria kategori curah hujan menurut BMKG dapat dilihat pada tabel II-1.

Tabel II-1 Kategori curah hujan

Kategori	Rentang Curah Hujan
Sangat Ringan	0
Ringan	0.5 – 20 mm/hari
Sedang	21 – 50 mm/hari
Lebat	51 – 100 mm/hari
Sangat Lebat	100-150 mm/hari

Intensitas terhadap cura hujan, dapat dipengaruhi oleh beberapa variabel. Menurut (Pradipta, 2020) variabel yang dapat mempengaruhi curah hujan yaitu.

a. Suhu rata-rata (C)

   Berdasarkan pernyataan BMKG, suhu udara merupakan indikator dari rata-rata energi kinetik pergerakan molekul. Dalam proses pengukuran suhu, BMKG umumnya menggunakan termometer cair dalam kaca sebagai peralatan konvensional serta termometer PT-100 yang digunakan sebagai peralatan digital pengukur suhu. Secara numerik, suhu rata-rata dinyatakan dalam satuan (*Celsius*) yang digunakan sebagai data untuk variabel.

b. Kelambaban rata-rata (%)

   Kelembaban merupakan jumlah dari kandungan uap air yang terdapat pada udara, hal tersebut berdasarkan pernyataan BMKG. Kehadiran uap air ini bergantung pada suhu udara. Ketika suhu udara semakin tinggi, jumlah uap air yang ada juga meningkat. Alat yang digunakan untuk mengukur kelembaban udara disebut higrometer.

c. Lama penyinaran (jam)

   Istilah "lama penyinaran" menggambarkan durasi atau waktu ketika matahari memberikan penyinaran. Lama penyinaran didefinisikan memiliki intensitas matahari yang lebih besar dari 120 W / m2 dan dapat ditemukan dalam beberapa elemen klimatologis, hal tersebut berdasarkan oleh keterangan BMKG. *Campbell Stokes Recorder* adalah alat utama yang diadopsi oleh BMKG untuk mengukur penyinaran matahari.

d. Arah angin saat kecepatan maksimum (deg)

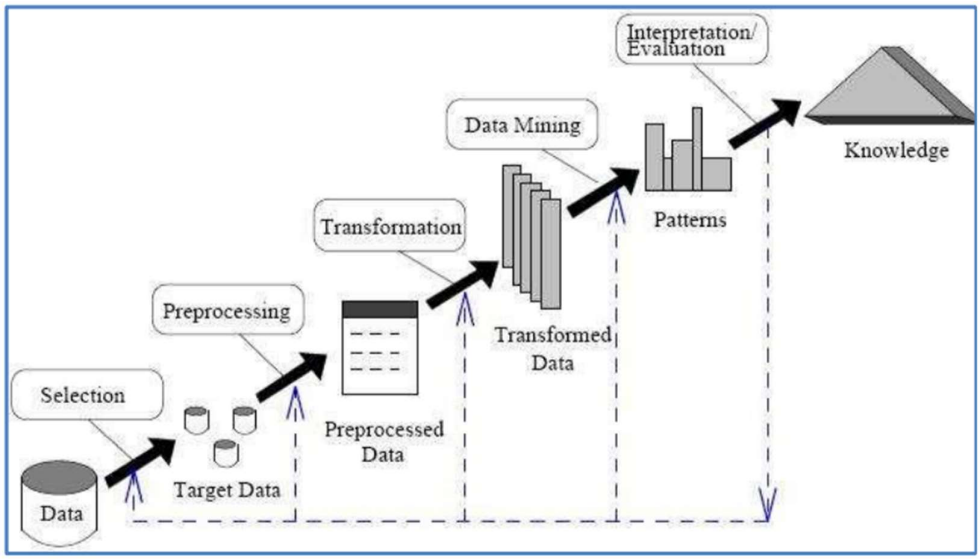
   Gerakan angin dapat disimpulkan dari arah angin. Arah angin dapat ditentukan oleh sudut yang dibuatnya di dalam kompas berdasarkan dari mana arah angin tersebut berhembus.

II.3 *Data mining*

Data mining adalah suatu proses yang melibatkan penerapan teknik-teknik matematika, statistik, kecerdasan tiruan, dan *machine-learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Tujuan utama dari disiplin ilmu *data*

mining adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang tersedia. Selain itu, *data mining* juga dapat dianggap sebagai suatu proses analisis data yang melibatkan eksplorasi dan penelaahan data untuk menemukan hubungan dan pola yang tidak terduga serta menyajikan informasi secara ringkas dan bermanfaat bagi pemilik data. Terdapat beberapa teknik *data mining* yang dapat digunakan, seperti deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi, yang digunakan berdasarkan jenis tugas yang ingin diselesaikan.

Istilah *data mining* dan *knowledge discovery in databases* (KDD) juga sering digunakan secara bergantian untuk merujuk pada proses penggalian informasi yang tersembunyi dalam suatu basis data yang besar. *Knowledge Discovery in Databases* (KDD) merupakan sekumpulan proses yang memiliki tujuan untuk mendapatkan pengetahuan yang bermanfaat dari data. Proses KDD terdiri dari beberapa tahap, yaitu pembersihan data, integrasi data, seleksi data, transformasi data, *data mining*, evaluasi , dan presentasi pengetahuan (Karsito & Sari, 2019).



Gambar II.1 Proses KDD (Mardi, 2017)

Secara garis besar, proses KDD memiliki beberapa tahapan, yaitu (Mardi, 2017) :

1. *Data Selection*

   Sebelum memulai tahap ekstraksi informasi dalam *Knowledge Discovery in Databases* (*KDD*), seleksi atau filter data dari kumpulan data operasional harus dilakukan. Data yang telah dipilih untuk digunakan dalam proses *data mining* akan disimpan dalam file terpisah dari basis data operasional.

2. *Pre-Processing*

   Sebelum memulai proses *mining*, data yang akan digunakan harus menjalani tahap pembersihan (*cleansing*). Pemeriksaan data mencakup penghapusan data yang duplikat, pemeriksaan konsistensi data, dan perbaikan kesalahan pada data, seperti kesalahan cetak. Selain itu, dilakukan juga proses peningkatan data (*data augmentation*) yang bertujuan untuk menambahkan data atau informasi tambahan yang relevan dan diperlukan dalam *Knowledge Discovery in Databases (KDD),* seperti data atau informasi eksternal.

3. *Transformation*

   Koding merupakan suatu proses transformasi terhadap data untuk mengubah data yang telah dikumpulkan sebelumnya menjadi format yang sesuai untuk *data mining*.

4. *Data mining*

   Data mining merupakan proses yang bertujuan untuk menemukan pola atau informasi penting dalam data yang tidak terstruktur dengan menggunakan pendekatan atau teknik yang paling sesuai. Terdapat berbagai macam metode, teknik, dan algoritma yang digunakan dalam *data mining*. Penggunaan metode atau algoritma yang dapat dipercaya sangat membantu dalam mencapai tujuan dan keseluruhan proses.

5. *Interpretation / Evaluation*

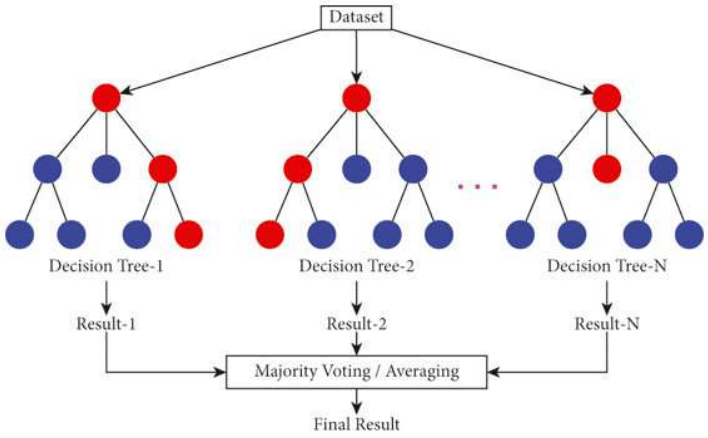
   Tahap *interpretation* merupakan tahapan yang mencakup pemeriksaan terhadap informasi yang ditemukan apakah sesuai dengan fakta atau hipotesis yang ada sebelumnya atau belum. Hasil dari proses *data mining* harus ditampilkan kedalam bentuk yang mudah dimengerti agar memudahkan dalam memahami hasil *data mining.*

II.4 Metode Klasifikasi

Klasifikasi merupakan salah satu metode yang dapat digunakan dalam penerapan *data mining*. Metode klasifikasi adalah salah satu metode analisis data yang menghasilkan model untuk menggambarkan kelas-kelas data yang signifikan. Model tersebut, yang disebut sebagai klasifier, dapat digunakan untuk memprediksi keanggotaan suatu data ke dalam kelas yang telah ditentukan (Han dkk., 2011). Metode klasifikasi biasanya termasuk dalam *supervised learning* yang bertujuan untuk menemukan pola atau hubungan antara atribut masukan dan atribut target. Tujuan klasifikasi adalah untuk dapat memprediksi atau mengklasifikasikan sampel baru ke dalam kelas yang sesuai berdasarkan pola yang telah ditemukan dari data latihan atau data yang sudah terklasifikasi sebelumnya. Hal ini dilakukan dengan tujuan untuk meningkatkan keakuratan atau keandalan hasil dari data yang dihasilkan (Hendrian, 2018).

II.5 Random Forest

Random forest merupakan algoritma yang dikembangkan oleh Leo Breimean. Random forest adalah sekelompok pohon regresi atau klasifikasi yang tidak dipangkas dan dibuat dengan cara memilih sampel acak dari data. Prediksi dibuat dengan menggabungkan hasil prediksi dari seluruh kelompok pohon regresi atau klasifikasi. Random forest memiliki keunggulan seperti mampu mendeteksi kesalahan yang relatif besar, kinerja klasifikasi yang baik, mampu menangani data dengan jumlah sample yang sedikit, dan metode yang efektif untuk mengestimasi data yang hilang (Ali dkk., 2012).



Gambar II.2 Random forest (Khan dkk., n.d., 2021)

Metode random forest merupakan pengembanggan metode *Classification and Regression Tree (CART)* dengan menggabungkan teknik *bootstrapping* atau "*bagging*" dan pemilihan fitur acak. Random forest adalah metode yang digunakan untuk mengklasifikasikan data dengan membuat beberapa pohon klasifikasi. Data dari setiap pohon kemudian digabungkan, dan data yang muncul paling sering diakumulasikan dan dipilih menjadi hasil klasifikasi. Metode ini terdiri dari simpul induk, simpul internal, dan simpul daun. Simpul induk adalah simpul yang terdekat dengan sumber dan biasanya disebut sebagai sumber dari pohon keputusan. Simpul internal adalah simpul dengan satu cabang masukan dan maksimal dua cabang keluaran. Dan simpul daun adalah simpul terakhir yang tidak memiliki keluaran dan hanya memiliki satu cabang masukan (Sandag, 2020).

II.6 Naïve Bayes

Naïve bayes adalah sebuah metode klasifikasi probabilitas yang menghitung total probabilitas dengan meningkatkan frekuensi dan mempertimbangkan data yang tersedia. Algoritma yang menggunakan konsep Bayes menunjukkan bahwa semua atribut independen atau tidak terkait satu sama lain dan diberikan oleh variabel kelas. Naïve bayes didasarkan pada asumsi kuat bahwa nilai atribut dalam kondisi tertentu saling bebas diberikan nilai keluaran. Atau, ketika diberikan nilai keluaran, probabilitas yang dihitung secara bersama-sama adalah produk dari probabilitas individu. Penggunaan naïve bayes dapat membantu karena metode ini

hanya membutuhkan sedikit data pelatihan untuk memperkirakan parameter yang diperlukan untuk klasifikasi secara akurat. Naïve bayes secara konsisten berperforma lebih baik dari yang diharapkan dalam situasi dunia nyata yang lebih kompleks dari yang diharapkan (Triawan & Melinda, 2020).

Dalam penerapan teori Bayes memiliki bentuk umum yang dapat diliha pada persaman II.1 :

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \quad \text{.....(II.1)}$$

Dimana :

X :Data dengan class yang belum diketahui

H :Hipotesis data merupakan suatu class spesifik atau khusus

$P(H|X)$:Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)

$P(H)$:Probabilitas hipotesis H (prior probabilitas)

$P(X|H)$:Probabilitas X berdasarkan kondisi pada hipotesis H

$P(X)$:Probabilitas X

II.7 SMOTE

Pendekatan SMOTE *(Synthetic Minority Oversampling Technique)* menggunakan konsep oversampling, yang melibatkan penambahan data dari kelas minor untuk menyeimbangkan angka dengan data dari kelas utama. SMOTE akan menggunakan teknik ketetanggaan untuk membangkitkan data dari kelas minor. Pendekatan dengan metode SMOTE menggunakan persamaan II.2 berikut:

$$X_{syn} = X_i + (X_{knn} - X_i)\, Y \quad \text{.....(II.2)}$$

Pengamatan baru generasi ini dilambangkan dengan huruf X_{syn}, sedangkan pengamatan ke-i dilambangkan dengan X_i. X_{knn} merupakan jarak terdekat X dari X_i. dan gamma adalah angka peluang dengan bilangan acak antara 0 dan 1. Nilai mayoritas dari K-tetangga terdekat kemudian akan digunakan untuk mengisi

nominal. Saat menghitung jarak dalam SMOTE, variabel kategoris yang memiliki nilai berbeda pada pengamatan ke-i dan ke-j diganti dengan deviasi kuadrat median standar dari variabel kontinu kelas minoritas (Syukron dkk., n.d).

II.8 *Confusion Matrix*

Confusion matrix digunakan dalam *data mining* untuk menghitung akurasi prediksi pada suatu model. *Confusion matrix* berisi informasi tentang jumlah prediksi yang benar atau salah dan digunakan untuk menghitung nilai akurasi, presisi, *recall* dan *F1-Score.* Presisi, juga dikenal sebagai confidence, mengacu pada proporsi kasus yang diprediksi positif yang benar pada data aktual. Recall, juga dikenal sebagai sensitivity, mengacu pada proporsi kasus aktual yang diprediksi positif dengan benar. Model pada *confusion matrix* dapat dilihat pada tabel II-2 (Sandag, 2020).

Tabel II-2 Confusion matrix

		Kelas Prediksi	
		+	-
Kelas Aktual	+	True Positives (TP)	False Negatives (FN)
	-	False Positives (FP)	True Negatives (TN)

Proses perhitungan dari tingkat akurasi, presisi, recall, dan F1-Score yang digunakan sebagai indikator untuk menentukan klasifikasi terbaik dapat dilakukan dengan rumus berikut (Christina Tanujaya dkk., 2020).

Akurasi, merupakan perbandingan dari objek yang diidentifikasi benar dengan jumlah semua subjek.

$$\text{Akurasi} = \frac{TP+FN}{TP+FP+FN+TN} \quad \text{(II.3)}$$

Presisi, merupakan hasil dari perbandingan objek yang secara benar diidentifikasi dengan hasil positif benar dengan jumlah hasil positif benar dan positif palsu.

$$\text{Presisi} = \frac{TP}{TP+FP} \quad \text{(II.4)}$$