BAB II
KAJIAN LITERATUR

2.1. Kondisi Bencana Hidrometeorologi Indonesia

Bencana alam menurut BNPB dikelompokkan menjadi bencana hidrometeorologi dan bencana geologi. Bencana hidrometeorologi adalah bencana yang terjadi akibat gangguan pada sistem siklus air yang mengganggu stabilitas iklim serta ketersediaan air di permukaan bumi. Fenomena ini muncul dalam berbagai bentuk, seperti banjir, kekeringan, badai, tanah longsor, hingga cuaca ekstrem berupa angin puyuh serta gelombang panas dan dingin [24].

Sebagai negara tropis, Indonesia memiliki kerentanan tinggi terhadap berbagai bencana hidrometeorologi, mulai dari banjir dan tanah longsor hingga cuaca ekstrem akibat perubahan iklim global. Data dari BNPB tahun 2022 menunjukkan bahwa mayoritas bencana tahunan di tanah air yakni lebih dari 80% didominasi oleh fenomena ini. Namun, meski pola bencana tersebut cenderung berulang dan sebenarnya dapat diprediksi, tingkat kewaspadaan serta kesadaran masyarakat dalam menghadapi risiko tersebut masih tergolong minim [25].

2.1.1. Definisi Banjir

Banjir adalah peristiwa air yang meluap ke wilayah daratan yang kering. Banjir dapat terjadi secara spontan atau perlahan, dan dapat berlangsung dalam waktu singkat atau lama. Peristiwa ini bisa disebabkan oleh beberapa faktor seperti curah hujan yang tinggi, meluapnya sungai, ataupun sistem drainase yang buruk [26]. Menurut Syafitri (2023) banjir merupakan keadaan dimana air yang ada melebihi daya tampung kawasan dan tidak bisa untuk dialirkan maupun diresapkan dengan cepat sehingga memberikan dampak berupa terendamnya suatu kawasan oleh air [27].

Di wilayah Indonesia, fenomena ini sering muncul dalam berbagai ragam kejadian, mulai dari luapan air sungai, genangan di area cekungan akibat curah hujan tinggi, banjir bandang di wilayah hulu, hingga banjir rob akibat pasang air laut di wilayah pesisir. Meskipun pemicu spesifiknya beragam, dalam konteks permodelan statistik, seluruh kejadian tersebut dipandang sebagai satu variabel dependen biner yang dipengaruhi oleh parameter fisik yang seragam, seperti ketinggian lahan, kemiringan lereng, curah hujan, dan jarak terhadap jaringan drainase atau sungai [12], [21].

2.1.2. Karakteristik Geografis dan Kompleksitas Topografi

Topografi merupakan cabang ilmu yang mempelajari wujud permukaan bumi serta objek lain seperti planet, satelit, dan asteroid. Objek geografis sendiri terbagi menjadi bentang budaya dan bentang alam. Bentang budaya mencakup segala hal buatan manusia, contohnya jalan raya, rel kereta api, kawasan permukiman, hingga area pertanian. Sementara itu, bentang alam meliputi segala sesuatu yang terbentuk secara alami tanpa campur tangan manusia, seperti dataran tinggi dan rendah, pegunungan, sungai, danau, kondisi iklim, hingga karakteristik tanah [28].

Indonesia berada di 6^0 LU hingga 11^0 LS dan 92^0 BT hingga 142^0 BT dan merupakan negara kepulauan yang memiliki topografi yang kompleks, seperti yang ditunjukkan pada gambar berikut

Gambar 2.1 Ketinggian Daratan Indonesia [29].

Ketinggian daratan di Indonesia bervariasi antara 0 dan 4.535 meter di atas permukaan laut berdasarkan data dari [29]. Secara geologis, keberadaan daratan yang membentuk pulau-pulau besar di wilayah ini memengaruhi proses transportasi massa dan panas antara Samudra Hindia dan Pasifik Daratan ini juga mendukung proses transfer panas dari laut ke atmosfer, sehingga meningkatkan curah hujan di wilayah ini. Selain itu, interaksi antara daratan dan angin juga memiliki pengaruh penting terhadap variasi intensitas curah hujan di wilayah Indonesia [30].

2.2. Faktor Pendukung Banjir

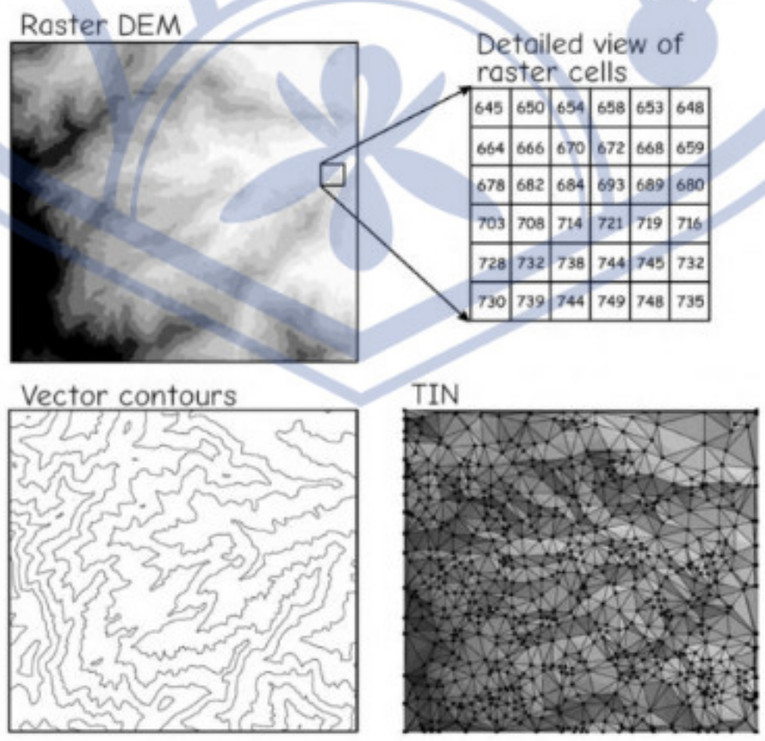
Secara mekanis, fenomena banjir di perkotaan dipicu oleh ketidakmampuan infrastruktur drainase, seperti sungai dan saluran pembuangan, dalam menampung debit air yang melampaui kapasitas tampungnya. Selain faktor perubahan lingkungan tersebut, karakteristik fisik alami seperti intensitas curah hujan yang tinggi dan kondisi topografi berupa dataran rendah yang cekung turut memperbesar probabilitas terjadinya genangan.

Ketidakseimbangan antara input air meteorologis dengan kapasitas sistem pengaliran di hilir menyebabkan air melimpah dan menggenangi wilayah sekitarnya. Oleh karena itu, penelitian ini difokuskan untuk menganalisis variabel-variabel determinan penyebab banjir di area perkotaan guna memberikan dasar yang objektif dalam upaya mitigasi bencana [31].

2.2.1. Elevasi (*Elevation*)

Elevasi dapat didefinisikan sebagai ketinggian suatu tempat terhadap daerah sekitarnya (di atas permukaan laut) [32]. Dalam pemodelan dan pengelolaan data elevasi digital, struktur representasi spasial yang paling umum diaplikasikan meliputi raster grid, Triangulated Irregular Networks (TIN), dan kontur vektor. Data spasial berbasis raster maupun TIN secara akademis sering diistilahkan sebagai Digital Elevation Model (DEM) atau Digital Terrain Model (DTM), yang merupakan instrumen fundamental dalam pelaksanaan analisis medan dan topografi [33].

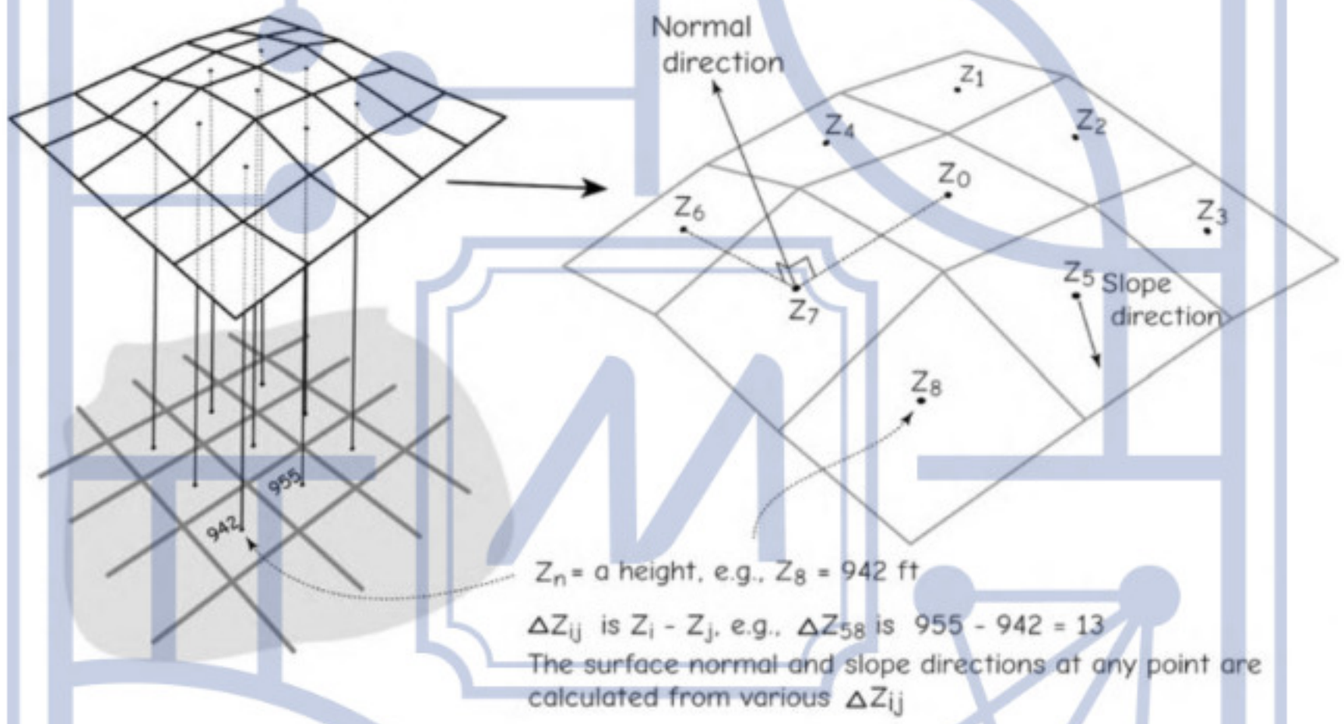
Secara historis, variasi ketinggian permukaan bumi dipresentasikan melalui garis kontur. Secara definisi, kontur merupakan garis yang merepresentasikan titik-titik dengan nilai elevasi yang ekuivalen, yang umumnya didistribusikan secara spasial dengan interval vertikal yang tetap di seluruh area pemetaan. Dalam praktik Sistem Informasi Geografis (GIS) modern, garis kontur kini lebih banyak difungsikan sebagai data masukan (input) dasar atau sebagai visualisasi keluaran (output) yang mudah dipahami. Namun, mengingat bahwa eksekusi analisis spasial tingkat lanjut memiliki tingkat kesulitan komputasi yang jauh lebih tinggi jika menggunakan data berbasis garis kontur, mayoritas penyimpanan dan pemrosesan data elevasi digital saat ini mengadopsi arsitektur model raster atau TIN [33].



Gambar 2.2 Data sering kali dapat direpresentasikan dalam beberapa model data. Data ketinggian digital umumnya direpresentasikan dalam model data raster (DEM), vektor (kontur), dan TIN [33].

2.2.2. Kemiringan Lereng (*Slope*)

Kemiringan lereng (slope) merupakan salah satu variabel topografi yang paling fundamental dan lazim digunakan dalam analisis spasial. parameter ini sangat esensial dalam mendukung berbagai studi hidrologi, konservasi, perencanaan wilayah, serta pengembangan infrastruktur, sekaligus menjadi landasan bagi berbagai fungsi analisis medan lainnya. Penentuan batas Daerah Aliran Sungai (DAS), pemodelan jalur dan arah aliran air (flowpaths), estimasi tingkat erosi, hingga analisis jarak pandang (*viewshed*), seluruhnya memanfaatkan data slope sebagai salah variabel masukan yang krusial. Selain itu, parameter topografi ini juga berkontribusi signifikan dalam pemetaan vegetasi dan sumber daya tanah guna meningkatkan akurasi analisis keruangan [33].



Gambar 2.3 Kemiringan Lereng Dalam Topografi Digital [33].

2.2.3. Curah Hujan

Curah hujan didefinisikan sebagai akumulasi volume air yang jatuh ke permukaan bumi dalam rentang waktu tertentu. Secara teknis, parameter ini merujuk pada jumlah air yang terakumulasi di atas permukaan tanah datar selama periode tertentu yang diukur dalam satuan tinggi (mm) di atas bidang horizontal, dengan asumsi tidak terjadi fenomena penguapan (evaporasi), aliran permukaan (limpasan), maupun penyerapan ke dalam tanah (infiltrasi) [34].

Besaran curah hujan yang dinyatakan dalam bentuk tinggi hujan atau volume per satuan waktu yang terjadi selama air hujan terkonsentrasi disebut sebagai intensitas curah hujan. Dalam standar pengukuran, nilai curah hujan sebesar 1 mm merepresentasikan volume air yang tertampung pada luasan satu meter persegi dengan ketinggian air mencapai 1 mm pada tempat yang datar [35].

2.2.4. Proksimitas Sungai (*Distance to River)*

Dalam pendekatan kuantitatif berbasis Sistem Informasi Geografis (SIG), jarak terhadap sungai dihitung menggunakan analisis *Euclidean Distance* atau *proximity analysis*, yang menghasilkan nilai kontinu dalam satuan meter. Nilai ini kemudian digunakan sebagai variabel independen dalam model statistik maupun machine learning, termasuk Regresi Logistik. Beberapa penelitian menunjukkan bahwa variabel jarak ke sungai memiliki kontribusi signifikan dalam meningkatkan akurasi model prediksi banjir, terutama pada wilayah dataran rendah dan daerah aliran sungai (DAS) [36].

2.3. Euclidean Distance Sungai

Jarak Euklides (*Euclidean Distance*) merupakan ukuran jarak yang paling intuitif karena sesuai dengan persepsi kita sehari-hari mengenai jarak. Jarak Euklides d dari dua titik data (x_1, x_2) didefinisikan sebagai akar kuadrat dari jumlah kuadrat selisihnya [37]

$$d(p,R) = \min_{q \in R} \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} \quad \text{.....(1)}$$

Di mana:

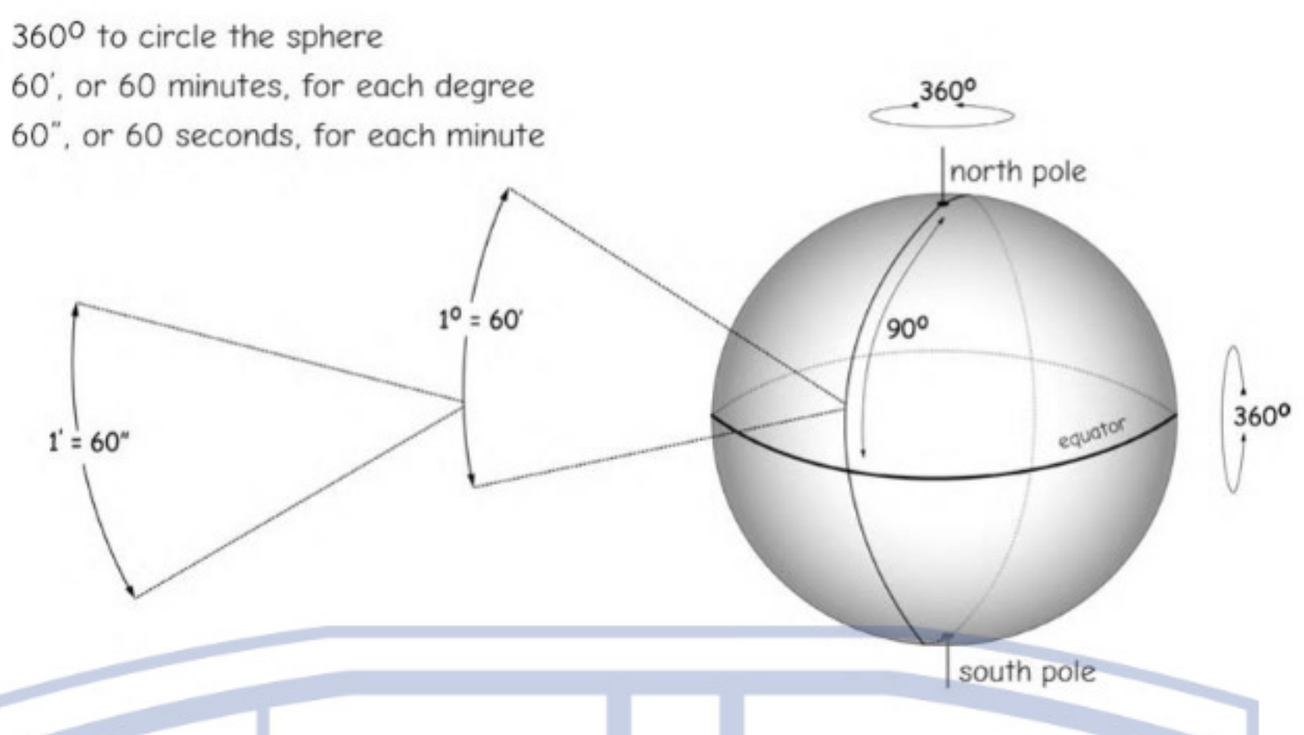
$d(p,R)$ = Jarak terpendek (meter).

$(x_p - y_p)$ = Koordinat lokasi pengamatan dalam proyeksi UTM.

$(x_q - y_q)$ = Titik pada geometri sungai yang memiliki jarak minimal terhadap p.

2.3.1. Transformasi Sistem Koordinat

Transformasi sistem koordinat dari format DMS (*Degrees*, *Minutes*, *Seconds*) ke DD (*Decimal Degrees*) adalah proses mengonversi pembacaan lokasi di permukaan bumi dari format sudut seksagesimal (basis 60) menjadi format angka desimal murni (basis 10). Konversi ini mutlak diperlukan karena perangkat lunak Sistem Informasi Geografis (GIS) dan algoritma pemodelan statistik komputasional tidak dapat membaca data yang mengandung teks atau simbol derajat. Sistem analitik tersebut membutuhkan input berupa nilai numerik yang kontinu (*floating-point*) agar dapat melakukan operasi matematika, menghitung jarak spasial antar titik secara presisi, dan mengeksekusi analisis keruangan [33].



Gambar 2.4 Transformasi Koordinat Bola [33].

Koordinat bola (*spherical coordinates*) umumnya dicatat menggunakan notasi derajat-menit-detik (DMS); contohnya U43° 35' 20" yang berarti 43 derajat, 35 menit, dan 20 detik garis lintang. Dalam format DMS, setiap satu derajat tersusun dari 60 menit busur, dan setiap menit dibagi lagi menjadi 60 detik busur. Sistem ini menghasilkan perhitungan 60 dikali 60, atau 3600 detik untuk setiap derajat lintang maupun bujur [33].

Format DMS dapat dikonversi menjadi DD dengan cara:

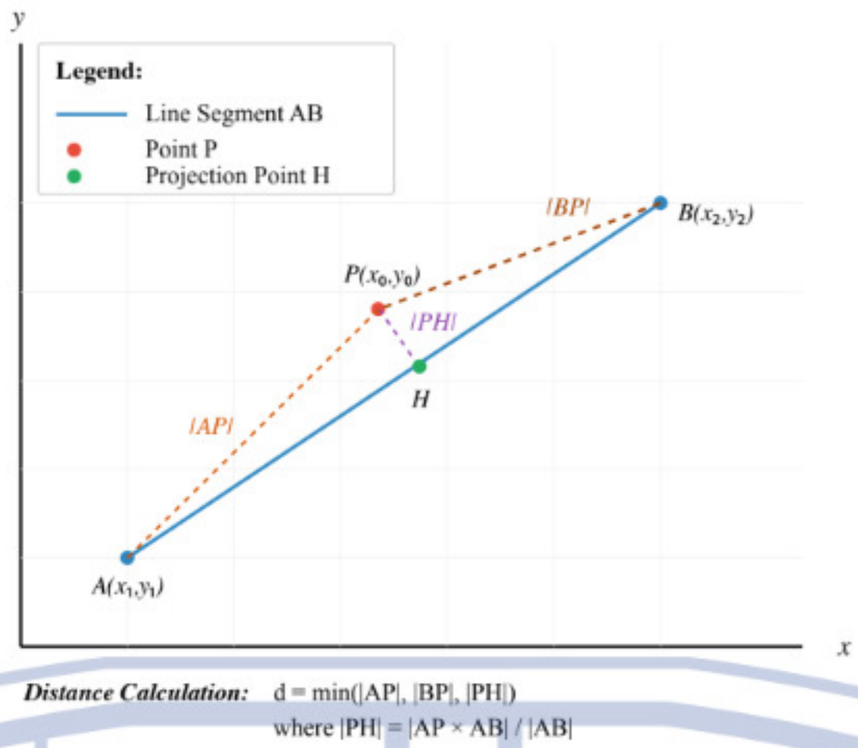
$$DD = DEG + \frac{MIN}{60} + \frac{SEC}{3600} \quad \text{.....(2)}$$

2.4. Penentuan Titik Terdekat pada Segmen

Dalam analisis spasial dan komputasi visual, penentuan posisi suatu titik (misalnya titik fiksasi atau objek observasi) terhadap sebuah batasan sering kali membutuhkan perhitungan jarak terpendek dari titik tersebut ke sebuah segmen garis. Berbeda dengan garis lurus biasa yang memiliki panjang tak terhingga, segmen garis memiliki batas awal dan akhir.

Secara teoritis, penentuan jarak terpendek dari sebuah titik ke segmen garis tidak selalu berupa garis proyeksi yang tegak lurus. Menurut literatur Chen K (2025), jarak terpendek ini dievaluasi berdasarkan tiga kemungkinan kondisi geometris [38]:

1. Proyeksi berada di luar segmen (sebelum titik awal): Jarak terpendek diukur langsung dari titik observasi ke titik ujung pertama dari segmen garis.
2. Proyeksi berada di luar segmen (setelah titik akhir): Jarak terpendek diukur langsung dari titik observasi ke titik ujung kedua dari segmen garis.
3. Proyeksi jatuh tepat di sepanjang segmen: Jarak terpendek adalah jarak proyeksi ortogonal (tegak lurus) dari titik observasi ke garis tersebut, yang umumnya dikomputasi menggunakan prinsip vektor.



Gambar 2.5 Proyeksi Ortogonal [38].

Perhitungan titik yang terdapat pada garis dapat dihitung menggunakan rumus sebagai berikut [38]:

$$|PH| = \frac{|AP \times AB|}{|AB|} \quad \text{.....(3)}$$

2.5. Sistem Informasi Geografis (SIG)

Sistem Informasi Geografis (SIG) merupakan sebuah sistem informasi berbasis komputer yang dirancang khusus untuk mengelola dan menganalisis data dengan atribut spasial atau referensi geografis. Sebagai perangkat lunak teknis, SIG memfasilitasi serangkaian proses yang komprehensif, mulai dari tahap akuisisi data (input), penyimpanan, pemrosesan (manipulasi), hingga penyajian (visualisasi) dan publikasi informasi mengenai karakteristik wilayah tertentu [39].

2.5.1. Konsep Analisis Spasial

Menurut Faisol (2012), analisis spasial adalah proses pengolahan dan evaluasi data yang memiliki informasi lokasi untuk menghasilkan informasi baru melalui teknik analisis tertentu. Analisis ini tidak hanya berfokus pada karakteristik suatu objek, tetapi juga mempertimbangkan aspek lokasi, distribusi, pola, dan hubungan antarobjek di dalam ruang geografis. Dengan demikian, analisis spasial memungkinkan peneliti untuk memahami fenomena berdasarkan dimensi "di mana" suatu kejadian terjadi, selain "apa" dan "mengapa" [40].

Dalam SIG, data spasial umumnya direpresentasikan dalam bentuk *layer* (lapisan) yang menggambarkan berbagai unsur geografis seperti sungai, jalan, penggunaan lahan, kemiringan lereng, curah hujan, maupun batas administrasi. Setiap layer memiliki atribut

yang dapat dianalisis secara individual maupun dikombinasikan dengan layer lain melalui teknik tertentu untuk memperoleh informasi yang lebih komprehensif.

Beberapa konsep dasar dalam analisis spasial meliputi [40]:

1. Konsep Lokasi (*Location*)

   Lokasi menunjukkan posisi suatu objek di permukaan bumi yang dinyatakan dalam sistem koordinat tertentu. Konsep ini menjadi dasar dalam menentukan posisi absolut maupun relatif suatu fenomena geografis.

2. Analisis Proksimitas (*Proximity Analysis*)

   Analisis jarak digunakan untuk mengetahui tingkat kedekatan antarobjek dalam ruang. Dalam kajian kebencanaan, misalnya, jarak permukiman terhadap sungai dapat menjadi salah satu variabel penentu tingkat risiko banjir.

3. Analisis *Buffer*

   Buffer adalah zona dengan radius tertentu di sekitar objek spasial. Dalam studi banjir, buffer sering digunakan untuk mengidentifikasi zona rawan di sekitar sungai atau badan air.

4. Analisis Interpolasi

   Metode ini digunakan untuk memperkirakan nilai pada lokasi yang tidak memiliki data pengamatan langsung berdasarkan titik-titik terdekat, seperti interpolasi curah hujan menggunakan metode Kriging atau IDW (*Inverse Distance Weighting*).

5. Analisis Raster Berbasis Sel (*Grid-based Analysis*)

   Dalam pendekatan raster, wilayah studi dibagi menjadi sel-sel grid dengan resolusi tertentu. Setiap sel menyimpan nilai numerik yang mewakili parameter tertentu (misalnya elevasi atau curah hujan). Model statistik seperti Regresi Logistik kemudian diaplikasikan pada setiap sel untuk menghasilkan peta probabilitas risiko banjir.

Dalam penelitian ini, analisis spasial berfungsi sebagai tahap praproses (*pre-processing*) untuk mengekstraksi dan menormalisasi variabel prediktor yang akan digunakan dalam model Regresi Logistik. Integrasi antara SIG dan metode statistik memungkinkan pengembangan model prediksi berbasis data yang objektif, replikatif, dan mampu diterapkan dalam skala nasional [40].

2.5.2. Open Street Map (OSM)

OpenStreetMap (OSM) merupakan proyek pemetaan dunia kolaboratif yang bersifat sumber terbuka (*open-source*), terbuka untuk penyuntingan, dan dapat diakses secara bebas

oleh publik. Proyek ini diinisiasi pada tahun 2004 di Inggris sebagai solusi atas keterbatasan ketersediaan data geospasial berkualitas tinggi yang bersifat komersial dan tertutup [41].

Data yang termuat dalam OSM mencakup berbagai elemen infrastruktur dan fitur alam yang sangat mendalam, di antaranya [41]:

1. Infrastruktur Transportasi: Jaringan jalan, jalur kereta api, jalur transit, dan rute pendakian.
2. Lingkungan Binaan: Bangunan, alamat, pusat bisnis, toko, dan titik kepentingan (*Point of Interest*).
3. Fitur Alami: Penggunaan lahan (land use), area vegetasi, serta fitur hidrografi seperti jaringan sungai.

Secara operasional, peta ini dibuat dan dipelihara secara mandiri oleh jutaan kontributor terdaftar dari seluruh dunia dengan memanfaatkan berbagai perangkat lunak pemetaan bebas. Tata kelola proyek dijalankan melalui struktur organisasi yang ramping oleh *OpenStreetMap Foundation*, sebuah organisasi mandiri yang didanai melalui donasi serta keanggotaan korporasi [41].

Dalam penelitian ini, OSM berperan sebagai penyedia *dataset* vektor jaringan sungai yang menjadi dasar perhitungan variabel proksimitas (*distance to river*). Sifat datanya yang dinamis dan berbasis kontribusi sukarela (*Volunteered Geographic Information*) memungkinkan akses data spasial yang luas dan mendalam untuk wilayah Indonesia tanpa kendala lisensi komersial [41].

2.5.3. GeoPandas

GeoPandas merupakan proyek sumber terbuka (*open-source*) yang dirancang untuk menambahkan dukungan data geografis pada objek-objek dalam pustaka Pandas. Pustaka ini memperluas fungsionalitas Pandas dengan mengimplementasikan tipe data *GeoSeries* dan *GeoDataFrame*, yang masing-masing merupakan turunan (*subclass*) dari *pandas.Series* dan *pandas.DataFrame*. Melalui GeoPandas, manipulasi data tabular dapat dilakukan bersamaan dengan operasi spasial yang kompleks [42].

Operasi geometri dalam *GeoPandas* dijalankan melalui integrasi dengan pustaka *Shapely*. Hal ini memungkinkan *GeoPandas* untuk melakukan berbagai analisis spasial, seperti perhitungan luas area, penggabungan data berdasarkan lokasi (*spatial join*), hingga perhitungan jarak antar objek geospasial. Objek geometri yang didukung mencakup titik (*Point*), garis (*LineString*), dan poligon (*Polygon*) [42].

Dalam penelitian ini, *GeoPandas* memiliki peran krusial pada tahap pra-pemrosesan data spasial. Pustaka ini digunakan untuk mengonversi data koordinat (lintang dan bujur) dari *dataset* kejadian banjir menjadi objek *GeoDataFrame*. Fitur utama yang dimanfaatkan adalah kemampuan untuk menghitung Jarak Euclidean antara titik kejadian banjir dengan vektor jaringan sungai yang diperoleh dari *OpenStreetMap* [42].

2.6. Regresi Logistik dan Fitur Polinomial

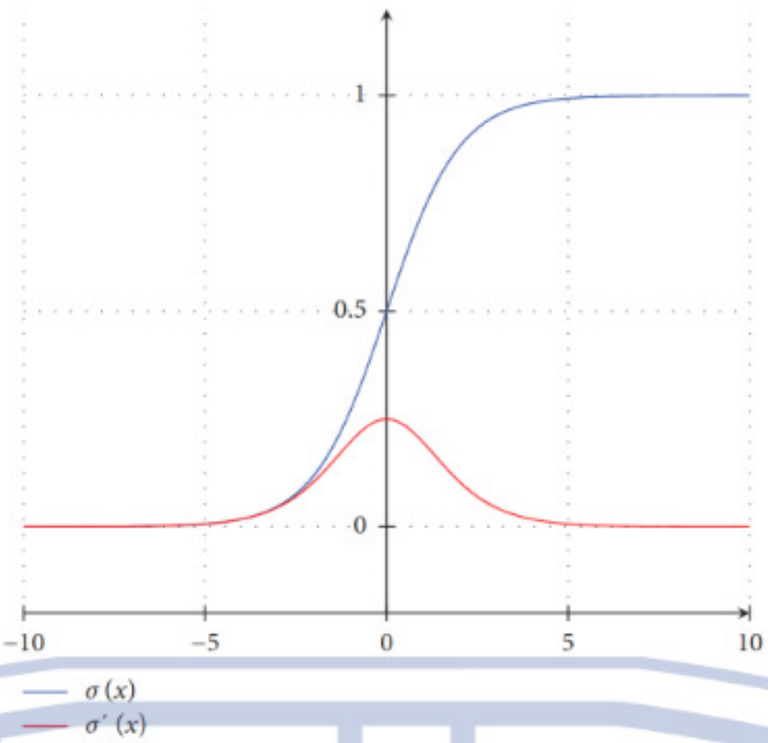
Meskipun konsepnya telah berkembang sejak lama, Regresi Logistik (LR) menjadi semakin populer dalam beberapa dekade terakhir sebagai alat analisis data. Fokus utama model ini adalah pada variabel dependen kategorik, khususnya untuk klasifikasi biner yang menggunakan nilai 0 dan 1 sebagai representasi kondisi. Secara umum, nilai 0 digunakan untuk menunjukkan kegagalan atau kondisi tidak terjadi, sedangkan nilai 1 menunjukkan keberhasilan atau terjadinya suatu fenomena [43].

Melalui model ini, peneliti dapat menghitung peluang variabel respon bernilai 1, sehingga hasil klasifikasi biner dapat diprediksi dan dipahami secara lebih akurat. Selain tipe biner, terdapat pula pengembangan berupa Regresi Logistik ordinal dan multinomial yang mampu mengklasifikasikan data dengan lebih dari dua kelompok. Dalam penerapannya di bidang kecerdasan buatan, model ini sangat diandalkan karena kemampuannya mentransformasikan input numerik menjadi probabilitas melalui fungsi logistik khusus (Sigmoid) [43].

2.6.1. Fungsi Sigmoid Logistik

Fungsi matematika yang disebut fungsi sigmoid menerima masukan berupa berbagai macam angka dan menghasilkan nilai keluaran antara 0 dan 1. Fungsi ini merupakan salah satu jenis fungsi aktivasi yang digunakan oleh banyak jaringan saraf tiruan untuk menentukan cara mereka menghasilkan informasi. Dalam Regresi Logistik, fungsi ini juga digunakan untuk mengubah data menjadi probabilitas antara 0 dan 1 [44]. Secara matematis, fungsi Sigmoid didefinisikan sebagai berikut [45][45]:

$$Logistic\ Sigmoid(x) = \frac{1}{1+e^{-z}} \quad \text{.....(4)}$$



Gambar 2.6 Fungsi Sigmoid [44]

Di mana:

e = Konstanta Euler ~2,718, dasar dari logaritma alami.

z = Prediktor Linear. Ini mewakili jumlah tertimbang dari data.

Fungsi aktivasi ini memampatkan (*squash*) nilai keluaran agar selalu berada dalam rentang [0, 1], seperti yang ditunjukkan pada Gambar 2.4. Namun, keluaran dari fungsi Sigmoid Logistik rentan mengalami kejenuhan (saturasi) ketika menerima nilai input yang terlalu ekstrem (baik terlalu tinggi maupun terlalu rendah), yang berujung pada masalah gradien hilang (*vanishing gradient*) [45].

Fenomena vanishing gradient ini merujuk pada suatu kondisi di mana nilai gradien dari fungsi objektif terhadap suatu parameter menyusut tajam hingga mendekati titik nol. Akibatnya, hampir tidak terjadi pembaruan parameter secara signifikan selama fase pelatihan jaringan yang menggunakan teknik *stochastic gradient descent* (SGD). Dengan kata lain, proses pelatihan praktis terhenti total di bawah skenario gradien hilang tersebut. Selain itu, karakteristik keluarannya yang tidak berpusat pada angka nol (*non-zero-centric*) turut mengakibatkan buruknya proses konvergensi [45].

2.6.2. Regresi Logistik

Regresi Logistik adalah metode statistik yang digunakan untuk mendeskripsikan dan memprediksi hubungan antara variabel dependen yang bersifat biner, yang dikenal dengan variabel hasil, dengan satu atau lebih variabel bebas, yang dikenal sebagai prediktor atau kovariat. Metode ini dikenal karena fleksibilitasnya, interpretasi yang kuat, serta penerapannya yang meluas dalam berbagai bidang. Model Regresi Logistik sering digunakan untuk menganalisis data retrospektif, termasuk studi kasus-kontrol, serta membuat algoritma prediksi [46].

Karena model linear yang digunakan untuk mendeskripsikan hasil kontinu tidak dapat digunakan untuk memodelkan hasil yang hanya memiliki 2 kemungkinan (seperti 0 atau 1, ya atau tidak), yang mana dapat memberikan hasil kecocokan buruk jika tetap dilakukan, model Regresi Logistik menjadi solusi untuk model regresi linear dengan mengubah hasil biner menjadi hasil kontinu. Dalam Regresi Logistik, alih-alih memodelkan hasil biner secara langsung, yang dimodelkan adalah *log-odds*, yang disebut juga fungsi logit. Dalam konteks regresi logistik, persamaan ini dijabarkan sebagai [46]:

$$logit(P_x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad \text{.....(5)}$$

Di mana:

$logit(P_x)$ = Logaritma dari peluang kejadian

β_0 = Nilai potong titik Y saat semua variabel bernilai 0

$\beta_1, \beta_2, \ldots, \beta_n$ = Koefisien (nilai bobot yang diberikan kepada setiap variabel).

$X_1, X_2, \ldots, X_n$ = Variabel Independen

Regresi Logistik menonjol karena kesederhanaannya, kemampuan interpretasinya, dan efektivitasnya dalam masalah klasifikasi biner (area banjir vs. tidak banjir). Penelitian seperti Edamo (2024) telah menunjukkan kegunaan Regresi Logistik dalam pemodelan bahaya lingkungan, menampilkan potensinya dalam analisis kerentanan banjir [47].

2.6.3. Regresi Logistik Dengan Fitur Polinomial

Regresi Polinomial adalah bentuk dari regresi yang mana hubungan antar variabel dependen dan variabel independen dimodelkan menggunakan polinomial dengan derajat tertentu, sehingga menghasilkan fungsi non linear. Secara umum, Regresi Polinomial didefinisikan sebagai berikut [48]:

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \ldots + \beta_p \cdot X^p + \varepsilon \quad \text{.....(6)}$$

Di mana:

Y = Variabel yang diprediksi

β_0 = Nilai potong titik Y saat semua variabel bernilai 0

$\beta_1, \beta_2, \ldots, \beta_n$ = Koefisien (nilai bobot yang diberikan kepada setiap variabel).

$X_1, X_2, \ldots, X_n$ = Variabel Independen

2.7. Regulariasi (L1 *Lasso* & L2 *Ridge*)

Regularisasi merupakan prosedur dalam regresi penyusutan (*shrinkage regression*) yang diterapkan ketika metode *Ordinary Least Squares* (OLS) tidak dapat digunakan akibat

adanya masalah matriks desain yang bersifat singular. Teknik ini melakukan penyesuaian minor pada parameter model agar mampu menghasilkan performa yang optimal, baik pada data pelatihan maupun data pengujian [49].

Secara teknis, fungsi kerugian (*loss function*) yang digunakan untuk mengestimasi parameter model dikenal sebagai *Residual Sum of Squares* (RSS). Parameter model tersebut disesuaikan melalui mekanisme pertukaran antara bias dan varians (*bias-variance tradeoff*) guna meminimalkan nilai fungsi kerugian [49].

Persamaan RSS dinyatakan dalam rumus berikut:

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{.....(7)}$$

Di mana

n = Jumlah observasi

p = Jumlah variabel prediktor

2.7.1. L1 *Lasso*

Lasso merupakan prosedur regresi penyusutan yang secara spesifik menggunakan penalti L1. Metode ini berfungsi ganda, yaitu sebagai prosedur penyusutan koefisien sekaligus sebagai teknik seleksi variabel. Karakteristik utama dari penalti L1 adalah kemampuannya untuk memaksa estimasi koefisien regresi tertentu menjadi tepat bernilai nol, terutama ketika parameter memiliki nilai yang besar.Proses minimalisasi fungsi pada teknik Lasso, yang juga dikenal sebagai normalisasi L1 [49], dinyatakan dalam persamaan berikut:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \quad \text{.....(8)}$$

Di mana:

λ = Konstanta pinalti

2.7.2. L2 *Ridge*

Regresi *Ridge* merupakan teknik regresi yang menerapkan penalti L2 untuk membatasi kompleksitas model. Metode ini sangat berguna sebagai instrumen untuk menangani masalah multikolinearitas atau ketika matriks desain bersifat singular (tidak dapat dibalik). Fungsi objektif dalam regresi *Ridge* memodifikasi RSS dengan memperkenalkan parameter penyusutan (λ) sebagai berikut:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta^2 = RSS + \lambda\sum_{j=1}^{p}\beta^2 \quad \text{.....(9).}$$

Di mana:

λ = Konstanta pinalti

Parameter (λ) berfungsi sebagai parameter penala (*tuning parameter*) atau hiperparameter yang menentukan tingkat penalti terhadap model. Semakin besar nilai koefisien, maka kompleksitas model akan meningkat; oleh karena itu, penambahan komponen penalti ini akan menyusutkan koefisien regresi mendekati nol untuk meminimalkan fungsi tersebut [45][44].

2.8. *Standard Scaler* (Z-Score)

Z-score mewakili bentuk terstandarisasi dari nilai mentah (x), memberikan wawasan tentang posisi relatif nilai tersebut di dalam distribusinya[50]. *Z-Score* normalization dilakukan dengan memusatkan data pada nilai rata-rata (*mean*) 0 dan mengubah skala data sehingga memiliki varians 1. Proses ini memastikan bahwa data telah terstandarisasi, sehingga fitur-fitur yang berbeda dalam penelitian menjadi sebanding satu sama lain. Secara matematis, persamaannya adalah [51]:

$$z = \frac{x-\mu}{\sigma} \quad \text{.....(10)}$$

Di mana:

z = Nilai hasil standarisasi

x = Nilai asli data mentah

μ = Rata-rata dari seluruh data pada satu fitur

σ = Standar deviasi dari fitur tersebut

Persamaan dari standar deviasi:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad \text{.....(11)}$$

Di mana:

N = Jumlah data

μ = Rata-rata dari seluruh data pada satu fitur

x = Nilai asli data mentah

2.9. Validasi Model

Dalam praktiknya, berbagai strategi validasi model diterapkan untuk mencegah terjadinya *overfitting* (pencocokan berlebih). Dalam kondisi ideal, *dataset* yang tersedia dibagi menjadi tiga bagian utama: data latih (*training*), data validasi (*validation*), dan data uji (*testing*). Strategi ini sering dikenal sebagai metode holdout [52].

1. Data Latih (*Training Set*): Digunakan untuk membangun model dengan berbagai bentuk atau tingkat kompleksitas. Namun, estimasi galat (*error*) yang dihitung hanya berdasarkan data latih cenderung terlalu optimis (lebih rendah dari kenyataan) dan tidak dapat dijadikan acuan tunggal dalam pemilihan model.
2. Data Validasi (*Validation Set*): Digunakan sebagai pembanding untuk memilih model terbaik yang memiliki tingkat galat prediksi terkecil.
3. Data Uji (*Testing Set*): Merupakan data independen yang tidak terlibat dalam proses pelatihan maupun pemilihan model. Data ini digunakan untuk memberikan penilaian akhir yang objektif terhadap kemampuan generalisasi model (*test error*).

Namun, pada penelitian dengan ketersediaan data yang terbatas, pembagian data menjadi tiga bagian sering kali tidak memungkinkan. Dalam kondisi tersebut, proses validasi dapat dilakukan dengan metode *K-fold Cross-Validation* (CV). Pada metode ini, data dibagi menjadi K bagian yang sama besar. Secara bergantian, satu bagian disisihkan sebagai data uji, sementara K-1 bagian lainnya digunakan sebagai data latih. Prosedur ini diulangi sebanyak K kali hingga seluruh bagian pernah menjadi data uji. Hasil akhir dari galat prediksi model diperoleh dari nilai rata-rata galat yang dihasilkan pada setiap iterasi [52].

2.10. Hyperparameter Tuning dengan GridSearchCV

Penyetelan hiperparameter (*Hyperparameter Tuning*) adalah proses menemukan kumpulan hiperparameter optimal yang menghasilkan kinerja terbaik untuk model pembelajaran mesin. Proses ini sangat penting karena hiperparameter mengontrol proses pembelajaran dan struktur model seperti tingkat pembelajaran, jumlah neuron dalam jaringan saraf, atau ukuran *kernel* dalam *support vector machine*, yang berdampak langsung pada kinerjanya. Tidak seperti parameter model, yang dipelajari dari data, hiperparameter ditetapkan sebelum pelatihan dan memerlukan pemilihan yang cermat. Penyetelan hiperparameter dapat meningkatkan kinerja dan generalisasi model [53].

Grid Search didefinisikan sebagai metode pencarian sistematis yang mengevaluasi seluruh ruang parameter (*parameter space*) yang telah ditentukan secara *exhaustive*. Strategi ini bertujuan untuk mengidentifikasi kombinasi nilai hiperparameter yang menghasilkan metrik performa paling optimal berdasarkan data validasi [54].

Prosedur operasional algoritma *Grid Search* dilaksanakan melalui tahapan-tahapan terstruktur sebagai berikut:

1. Spesifikasi Ruang Parameter: Menetapkan himpunan nilai diskrit atau rentang kontinu untuk setiap hiperparameter yang akan diuji.

2. Konstruksi Kartesius (*Grid*): Membentuk matriks kombinasi dari seluruh nilai hiperparameter yang tersedia, menciptakan sebuah ruang pencarian multidimensi.
3. Pelatihan dan Evaluasi Iteratif: Untuk setiap koordinat kombinasi dalam grid:
   a. Melakukan inisialisasi dan pelatihan model pada partisi data latih (*training set*).
   b. Mengukur generalisasi model menggunakan teknik
   c. Validasi Silang (misalnya k-*fold Cross-Validation* dengan $k = 5$). Penggunaan $k = 5$ bertujuan untuk mereduksi bias estimasi performa pada data yang belum pernah dilihat sebelumnya.
   d. Mendokumentasikan metrik evaluasi (seperti akurasi, presisi, atau *F1-score*) sebagai representasi kualitas kombinasi tersebut.
4. Seleksi Model Optimal: Mengidentifikasi dan menetapkan kombinasi hiperparameter yang menunjukkan nilai signifikansi tertinggi pada metrik performa yang dipilih.

2.11. Metrik Evaluasi Klasifikasi Biner

Dalam tugas klasifikasi biner, instansi data dikategorikan ke dalam dua label diskrit, yaitu positif dan negatif. Secara umum, label positif merepresentasikan keberadaan suatu kondisi spesifik seperti penyakit, anomali, atau deviasi tertentu, sedangkan label negatif mengindikasikan kondisi dasar (*baseline*) atau ketiadaan kondisi tersebut [55] [56].

Matriks Kebingungan (*Confusion Matrix*) didefinisikan oleh [55], setiap hasil prediksi pada klasifikasi biner diklasifikasikan ke dalam empat penamaan fundamental yang membentuk matriks kontingensi 2×2, atau yang dikenal sebagai *Confusion Matrix*:

1. *True Positive* (TP) : Data positif yang diprediksi secara tepat sebagai positif.
2. *True Negative* (TN) : Data negatif yang diprediksi secara tepat sebagai negatif.
3. *False Positive* (FP) : Data negatif yang secara salah diprediksi sebagai positif.
4. *False Negative* (FN) : Data positif yang secara salah diprediksi sebagai negatif.

2.11.1. Akurasi, Sensitivitas, Spesifisitas, dan Presisi

Berdasarkan parameter pada matriks tersebut, performa klasifikator biner diukur menggunakan metrik-metrik berikut:

1. Akurasi (*Accuracy*): Merupakan rasio prediksi benar (positif dan negatif) terhadap keseluruhan total instansi data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

2. Sensitivitas (*Sensitivity*/*Recall*): Mengukur proporsi instansi positif aktual yang berhasil diidentifikasi dengan benar. Dalam konteks diagnostik medis, metrik ini dikenal sebagai *True Positive Rate* (TPR).

$$Sensitivity = \frac{TP}{TP+FN} \quad (13)$$

3. Spesifisitas (*Specificity*): Mengukur proporsi instansi negatif aktual yang berhasil diidentifikasi dengan benar, atau disebut juga sebagai *True Negative Rate* (TNR).

$$Specificity = \frac{TN}{TN+FP} \quad (14)$$

4. Presisi (*Precision*): Menunjukkan tingkat akurasi pada hasil yang diprediksi positif, yang secara formal disebut sebagai *Positive Predictive Value* (PPV).

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

2.11.2. F1-*Score*

F1-*score* merupakan metrik evaluasi yang menggabungkan presisi (*precision*) dan sensitivitas (*recall*). Metrik ini menjadi sangat krusial pada model yang menempatkan bobot kepentingan yang sama pada kedua nilai tersebut. Secara matematis, F1-*score* adalah rata-rata harmonik dari presisi dan recall yang mempertimbangkan baik kesalahan *false negative* (FN) maupun *false positive* (FP) [57].

Rata-rata harmonik dihitung dengan membagi jumlah nilai dalam suatu seri data dengan jumlah kebalikan (*reciprocal*) dari setiap nilai tersebut. Karakteristik rata-rata harmonik adalah nilainya akan selalu lebih kecil atau sama dengan rata-rata aritmatika maupun rata-rata geometrik. Hal ini disebabkan karena rata-rata harmonik memberikan bobot lebih besar pada nilai yang lebih kecil dalam seri data tersebut [57].

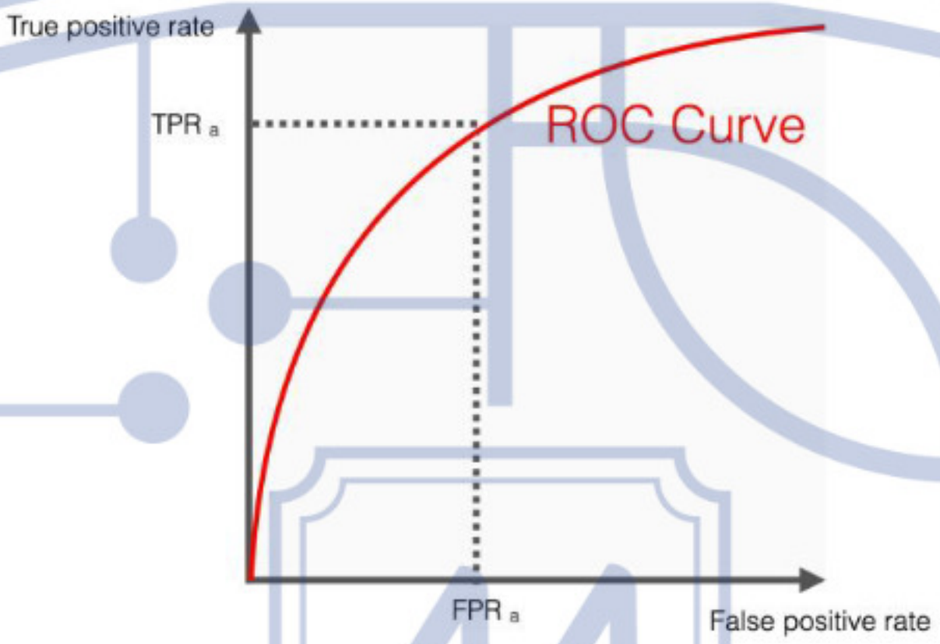
Apabila nilai presisi dan *recall* setara, maka rata-rata harmonik akan menunjukkan nilai rata-ratanya. Namun, jika terdapat perbedaan signifikan antara keduanya, rata-rata harmonik akan mendekati nilai yang lebih kecil. Oleh karena itu, F1-*score* jauh lebih representatif dibandingkan metrik akurasi (*accuracy*) ketika menangani data dengan ketidakseimbangan kelas (*class imbalance*). Diagambarkan dengan persamaan:

$$F1 - score = \frac{2 \times Precision \times Recall}{Percision + Recall}$$

$$F1 - Score = \frac{2TP}{2TP+FP+FN} \quad (16)$$

2.11.3. *Area Under ROC Curve* (AUC)

Implementasi metrik evaluasi yang representatif dalam tugas klasifikasi biner telah menjadi diskursus ilmiah yang intensif dalam beberapa tahun terakhir, khususnya pada domain data dengan distribusi kelas yang tidak seimbang (*imbalanced data*). Dalam konteks ini, *Area Under the Receiver Operating Characteristic Curve* (AUC) menempati posisi sebagai metrik yang paling ekstensif diadopsi dan mendapatkan atensi akademik yang signifikan dalam prosedur evaluasi performa model [56].



Gambar 2.7 AUC-ROC [56].

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP+FN} \quad \text{.....(17)}$$

$$False\ Positive\ Rate\ (FPR) = 1 - TNR \quad \text{.....(18)}$$

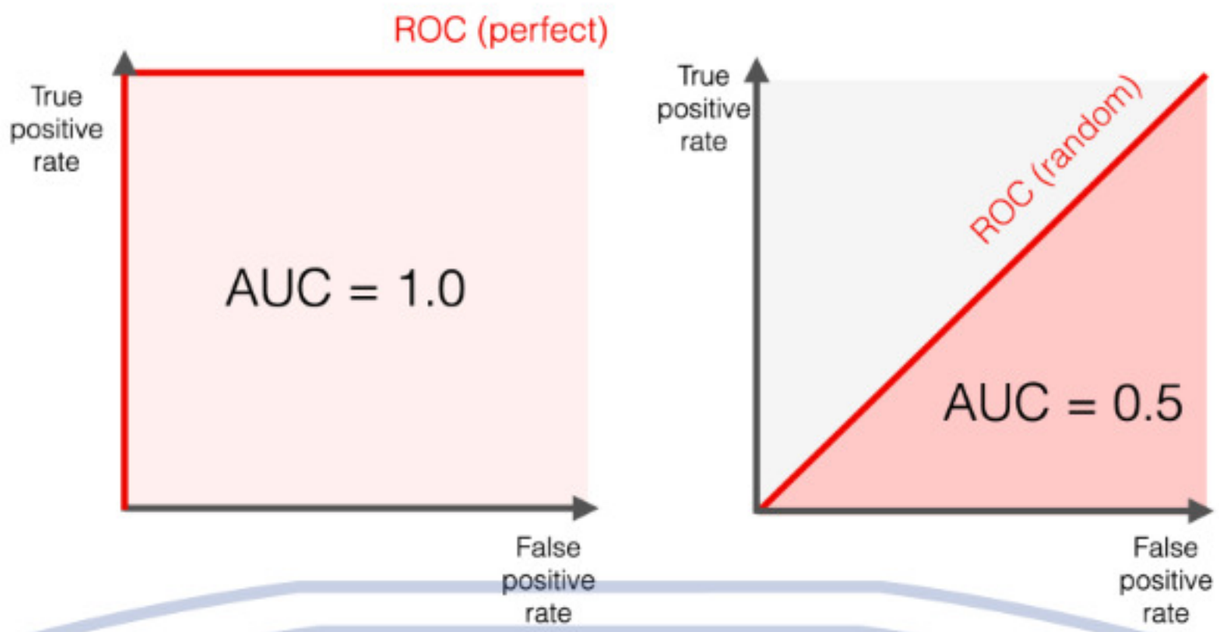
1. Karakteristik Kurva ROC

   Pada klasifikasi biner, kurva ROC merupakan representasi grafis yang mengintegrasikan seluruh ambang batas keputusan (*decision thresholds*) untuk mengevaluasi kemampuan diskriminasi model. Kurva ini memvisualisasikan korelasi antara dua parameter utama:

   a. Sumbu Absis (X): *False Positive Rate* (FPR), yang merepresentasikan proporsi kesalahan klasifikasi pada kelas negatif.
   b. Sumbu Ordinat (Y): *True Positive Rate* (TPR) atau Sensitivitas, yang menunjukkan kemampuan model dalam mengidentifikasi instansi positif secara akurat.

2. Interpretasi Probabilistik AUC

   Secara esensial, nilai AUC merepresentasikan probabilitas bahwa sebuah sampel positif yang diambil secara acak akan mendapatkan peringkat atau skor probabilitas yang lebih tinggi dibandingkan dengan sampel negatif yang diambil secara acak.



Gambar 2.8 Probabilistik AUC [56].

3. Signifikansi Metrik

   Penggunaan AUC dianggap lebih unggul dalam analisis stabilitas model karena sifatnya yang *threshold-independent* (tidak bergantung pada ambang batas tunggal). Model dengan nilai AUC yang mendekati 1 menunjukkan kemampuan separasi kelas yang superior, sementara nilai 0,5 mengindikasikan bahwa model tidak memiliki kemampuan diskriminasi yang lebih baik daripada klasifikasi acak (*random chance*).

2.12. FURPS

Model FURPS dikembangkan oleh Rebort Grady dan Hewlett Packard Co. Model FURPS merupakan akronim dari Functionality, Usability, Reliability, Performance, Supportability. Kerangka kerja tersebut dapat digunakan baik sebagai persyaratan sistem maupun di penilaian kualitas sistem. Dalam model FURPS, setiap indikator memiliki sub indikator yang dapat mempresentasikan setiap indikatornya dalam mengukur kualitas sistem. Berikut adalah penjelasan dari setiap indikator yang ada pada model [58]:

a. Functionality (Fungsionalitas)

   Merupakan kemampuan perangkat lunak untuk memenuhi kebutuhan dan persyaratan yang harus dimiliki dalam perangkat lunak tersebut untuk memenuhi harapan pengguna. Hal ini meliputi feature set, capabilities, dan security.

b. Usability (Kegunaan)

   Yaitu menggambarkan sejauh mana perangkat lunak mudah digunakan dan dipahami oleh pengguna. Hal ini meliputi faktor manusia, estetika, dan konsistensi dalam antar muka pengguna, dokumentasi pengguna dan materi pelatihan.

c. Reliability (Keandalan)

Mencakup bagaimana tingkat keandalan perangkat lunak untuk berfungsi secara konsisten tanpa mengalami gangguan atau kegagalan. Hal ini meliputi frekuensi tingkat keparahan kegagalan (failure), pemulihan (recovery), akurasi, prediksi dan waktu rata-rata antar terjadinya kegagalan.

d. Performance (Performa)

Merupakan kemampuan perangkat lunak untuk berfungsi dengan cepat, efisien, dan responsif sesuai kebutuhan pengguna. Hal ini meliputi kecepatan, efisiensi, ketersediaan, akurasi, waktu respon, waktu pemulihan, dan pemanfaatan sumber daya.

e. Supportability (Daya Dukung)

Merupakan sejauh mana perangkat lunak dapat didukung, diperbaiki, diadaptasi, dan diperluas di masa depan. Hal ini meliputi kemampuan untuk di uji, dikembangkan, adaptasi terhadap perbaruan, pemeliharaan, kompabilitas, konfigurabilitas, kemudahan servis, dan kemampuan instalasi.