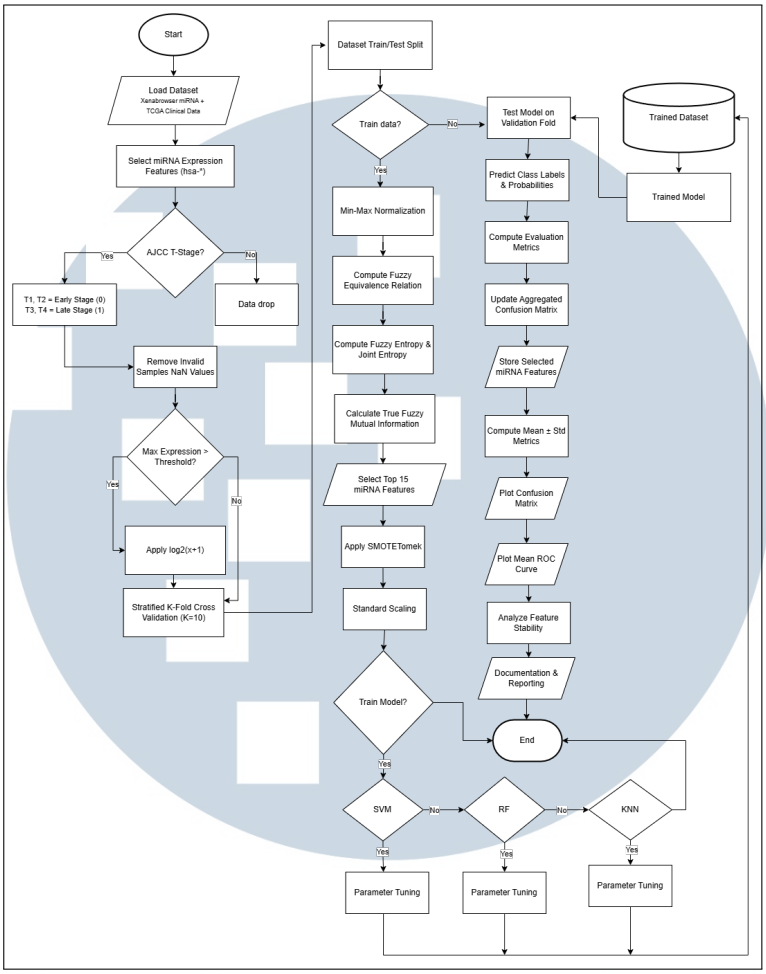
# BAB 3
# METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi penelitian yang digunakan untuk memprediksi stadium kanker lambung berbasis data microRNA (miRNA) dengan metode seleksi fitur berbasis *Fuzzy Mutual Information* (FMI) dan algoritma *supervised machine learning*. Seluruh tahapan dirancang agar sesuai dengan karakteristik data miRNA yang berdimensi tinggi, tidak seimbang, serta mengandung potensi redundansi fitur, dan diimplementasikan secara ketat dalam skema *cross-validation* untuk mencegah kebocoran data (*data leakage*).

## 3.1 Metodologi Penelitian

Penelitian ini disusun secara sistematis melalui beberapa tahapan yang saling berkaitan untuk menghasilkan model klasifikasi stadium kanker lambung berbasis data ekspresi miRNA. Tahapan penelitian dimulai dari pengumpulan dan praproses data, dilanjutkan dengan seleksi fitur berbasis *True Fuzzy Mutual Information*, penanganan ketidakseimbangan kelas, pelatihan model *machine learning*, evaluasi performa model, hingga dokumentasi hasil penelitian. Untuk memperjelas alur kerja penelitian secara menyeluruh, tahapan metodologi penelitian ini juga direpresentasikan dalam bentuk *flowchart*. Flowchart digunakan untuk menggambarkan urutan proses mulai dari pengumpulan data, praproses, seleksi fitur, pelatihan model klasifikasi, yaitu SVM, *Random Forest*, dan KNN, hingga evaluasi kinerja secara visual dan sistematis. Dengan adanya *flowchart*, hubungan antar tahapan serta aliran data pada setiap proses dapat dipahami dengan lebih jelas, sehingga memudahkan pembaca dalam mengikuti logika dan implementasi penelitian secara keseluruhan.




Gambar 3.1. Flowchart Metodologi Penelitian

Secara umum, alur penelitian meliputi pengumpulan data dari dataset publik, pemetaan label stadium kanker, normalisasi dan transformasi data, pemilihan fitur miRNA yang relevan menggunakan pendekatan *fuzzy*, penanganan ketidakseimbangan kelas dengan teknik *resampling*, pelatihan model klasifikasi, serta evaluasi menggunakan skema *Stratified K-Fold Cross-Validation*. Setiap tahapan dirancang untuk memastikan bahwa model yang dihasilkan memiliki kinerja yang stabil, objektif, dan mampu melakukan generalisasi dengan baik.

### 3.2 Pengumpulan dan Persipan Dataset

Data yang digunakan dalam penelitian ini merupakan data ekspresi microRNA (miRNA) yang diperoleh dari XenaBrowser dan data klinik dari The Cancer Genome Atlas (TCGA). Dataset ini mencakup data ekspresi miRNA serta informasi klinis pasien kanker lambung, termasuk *AJCC pathologic T-stage* yang digunakan sebagai dasar pelabelan kelas.

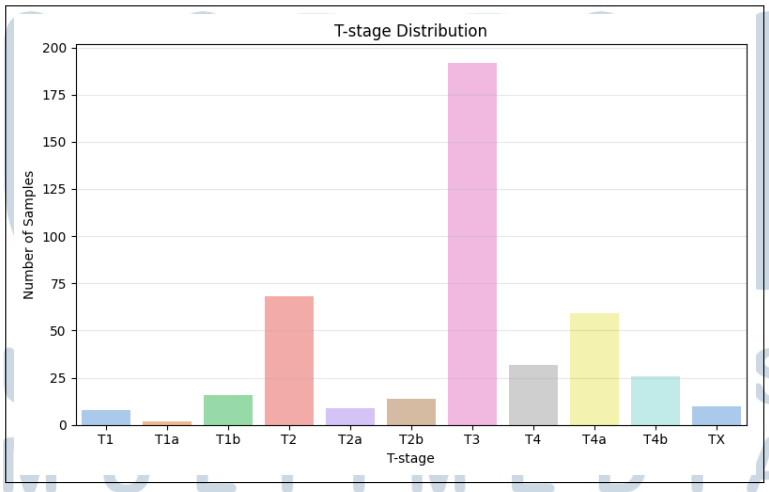
Tahap persiapan data diawali dengan pemilihan fitur miRNA, yaitu seluruh kolom yang merepresentasikan ekspresi miRNA dengan awalan penamaan “hsa”. Sebelum pengelompokan kelas, dilakukan analisis distribusi variabel target *AJCC pathologic T-stage* untuk mengetahui

sebaran data awal. Dari keseluruhan data TCGA yang tersedia, terdapat 436 sampel yang memiliki informasi T-stage valid. Distribusi rinci stadium T ditunjukkan pada Tabel 3.1.

Tabel 3.1. Distribusi data berdasarkan AJCC pathologic T-stage pada dataset TCGA kanker lambung.

| T-stage | Jumlah Sampel | Persentase (%) |
|---|---|---|
| T1 | 8 | 1.8 |
| T1a | 2 | 0.5 |
| T1b | 16 | 3.7 |
| T2 | 68 | 15.6 |
| T2a | 9 | 2.1 |
| T2b | 14 | 3.2 |
| T3 | 192 | 44.0 |
| T4 | 32 | 7.3 |
| T4a | 59 | 13.5 |
| T4b | 26 | 6.0 |
| TX | 10 | 2.3 |
| **Total** | **436** | **100** |

Sumber: Data TCGA Gastric Cancer




Gambar 3.2. T-stage Distribution

Selanjutnya, dilakukan pemetaan stadium kanker menjadi dua kelas, yaitu stadium awal dan stadium lanjut. Sampel dengan label yang tidak valid atau tidak lengkap dikeluarkan dari dataset. Penanganan nilai yang hilang pada data ekspresi miRNA, digunakan nilai median dari masing-masing fitur. Untuk keperluan klasifikasi biner diatas, dua kelas klinis, yaitu stadium awal

atau *early stage* terdiri dari gabungan T1, T1a, T1b, T2, T2a, dan T2b. Stadium lanjut atau *late stage* terdiri atas T3, T4, T4a, dan T4b. Sampel dengan label TX dikeluarkan dari proses klasifikasi karena tidak merepresentasikan stadium T yang jelas. Pengelompokan ini dilakukan berdasarkan pedoman AJCC dan praktik klinis yang umum digunakan dalam penelitian kanker lambung. Ringkasan hasil penggabungan kelas ditampilkan pada Tabel 3.2.

Tabel 3.2. Distribusi dataset berdasarkan pengelompokan stadium awal dan stadium lanjut.

| **Kategori Stadium** | **Jumlah Sampel** | **Persentase (%)** |
|---|---|---|
| Stadium Awal (T1 + T2) | 117 | 26.8 |
| Stadium Lanjut (T3 + T4) | 309 | 70.9 |
| **Total** | **426** | **100** |

Sumber: Data TCGA Gastric Cancer (setelah pembersihan data)

Apabila nilai ekspresi miRNA memiliki rentang yang besar, dilakukan transformasi logaritmik menggunakan $\log_2(x+1)$ untuk mengurangi skewness distribusi data. Tahapan ini bertujuan untuk menstabilkan variansi data sebelum dilakukan analisis lebih lanjut. Berdasarkan analisis distribusi *T-stage*, terlihat bahwa jumlah sampel pada stadium lanjut jauh lebih dominan dibandingkan stadium awal. Ketidakseimbangan distribusi kelas ini berpotensi memengaruhi kinerja model klasifikasi apabila tidak ditangani dengan tepat. Oleh karena itu, pada tahap selanjutnya diterapkan teknik penyeimbangan data menggunakan metode SMOTETomek untuk mengurangi bias model terhadap kelas mayoritas.

Dataset kemudian dibagi menjadi tiga subset utama untuk keperluan pembelajaran dan evaluasi model. Subset *training* digunakan untuk melatih model dan melakukan seleksi fitur, dengan memastikan bahwa distribusi kelas seimbang setelah penerapan SMOTETomek. Subset *validation* digunakan untuk menyesuaikan parameter model (*hyperparameter tuning*) dan memilih konfigurasi terbaik sebelum evaluasi akhir, sehingga mengurangi risiko *overfitting*. Subset *testing* digunakan untuk menilai performa akhir model secara independen, memberikan estimasi akurasi, sensitivitas, spesifisitas, F1-score, dan AUC yang tidak bias. Pengelompokan stadium menjadi dua kelas utama, yaitu stadium awal dan stadium lanjut, juga bertujuan untuk meningkatkan stabilitas model serta relevansi klinis dari hasil prediksi.

### 3.3 Seleksi Fitur Berbasis True Fuzzy Mutual Information

Seleksi fitur dilakukan menggunakan pendekatan *True Fuzzy Mutual Information* untuk mengidentifikasi miRNA yang paling relevan terhadap klasifikasi stadium kanker yang dikembangkan berdasarkan teori *fuzzy information* [13]. Pendekatan ini dirancang untuk menangkap hubungan nonlinier serta ketidakpastian yang umum terdapat pada data biologis berdimensi tinggi seperti ekspresi miRNA. Berbeda dengan *mutual information* klasik yang bersifat diskrit dan keras (*crisp*), pendekatan TFMI menggunakan relasi *fuzzy* memungkinkan representasi hubungan antar sampel secara kontinu. Dengan demikian, metode ini lebih robust terhadap *noise* serta variasi nilai ekspresi miRNA. Tahapan seleksi fitur terdiri dari normalisasi data, pembentukan relasi ekuivalensi

*fuzzy*, perhitungan entropi *fuzzy*, entropi gabungan, hingga perhitungan nilai *True Fuzzy Mutual Information* untuk setiap fitur miRNA.

Pseudocode pada Kode 3.1 menggambarkan alur seleksi fitur berbasis *True Fuzzy Mutual Information* yang digunakan dalam penelitian ini. Setiap fitur miRNA dievaluasi berdasarkan tingkat ketergantungannya terhadap label stadium kanker dengan mempertimbangkan ketidakpastian data melalui relasi ekuivalensi *fuzzy*. Fitur kemudian diurutkan berdasarkan nilai TFMI tertinggi dan dipilih sejumlah k fitur terbaik sebagai masukan model klasifikasi.

```
Input:
    X : matrix ekspresi miRNA (n_samples x n_features)
    y : label kelas stadium kanker (Early / Late)
    k : jumlah fitur terpilih

Output:
    SelectedFeatures : daftar miRNA terpilih

# Step 1: Normalisasi Data
X_norm <- MinMaxNormalization(X)

# Step 2: Inisialisasi skor TFMI
TFMI_scores <- empty list

# Step 3: Hitung TFMI untuk setiap fitur miRNA
for each feature f in X_norm.columns:

    # 3.1 Bentuk relasi ekuivalensi fuzzy untuk fitur f
    R_f <- ComputeFuzzyEquivalenceRelation(X_norm[f])

    # 3.2 Bentuk relasi ekuivalensi fuzzy untuk label kelas
    R_y <- ComputeFuzzyEquivalenceRelation(y)

    # 3.3 Hitung fuzzy entropy masing-masing relasi
    H_f <- FuzzyEntropy(R_f)
    H_y <- FuzzyEntropy(R_y)

    # 3.4 Hitung fuzzy joint entropy
    H_fy <- FuzzyJointEntropy(R_f, R_y)

    # 3.5 Hitung True Fuzzy Mutual Information
    TFMI_f <- H_f + H_y - H_fy

    # 3.6 Simpan skor TFMI
    TFMI_scores.append(TFMI_f)

# Step 4: Urutkan fitur berdasarkan skor TFMI
RankedFeatures <- SortDescending(TFMI_scores)

# Step 5: Pilih k fitur terbaik
SelectedFeatures <- RankedFeatures[0:k]

return SelectedFeatures
```

Kode 3.1: Pseudocode seleksi fitur miRNA menggunakan True Fuzzy Mutual Information



### 3.3.1 Normalisasi Data Min-Max Scaling

Sebelum memasuki tahapan seleksi fitur berbasis *fuzzy*, seluruh data ekspresi miRNA dinormalisasi menggunakan metode *Min-Max Scaling*. Normalisasi ini bertujuan untuk

menyamakan skala seluruh fitur ke dalam rentang [0,1], sehingga setiap fitur memiliki kontribusi yang sebanding dalam proses perhitungan jarak dan derajat keanggotaan *fuzzy*.

Pada data genomik, khususnya ekspresi miRNA, perbedaan rentang nilai antar fitur dapat sangat signifikan. Apabila kondisi ini tidak ditangani, fitur dengan skala nilai yang lebih besar akan mendominasi proses perhitungan relasi *fuzzy* dan berpotensi menyebabkan bias dalam pengukuran ketergantungan fitur terhadap kelas target. Oleh karena itu, normalisasi *Min-Max* dipilih karena mampu mempertahankan distribusi relatif data serta sesuai untuk integrasi dengan fungsi keanggotaan berbasis jarak [20].

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.1}$$

di mana:

- $x$ merupakan nilai asli dari suatu fitur miRNA,
- $x_{\min}$ dan $x_{\max}$ masing-masing adalah nilai minimum dan maksimum fitur tersebut,
- $x'$ adalah nilai hasil normalisasi dalam rentang [0,1].

Normalisasi ini dilakukan secara independen untuk setiap fitur miRNA agar struktur variasi biologis antar fitur tetap terjaga sebelum proses analisis *fuzzy* dilakukan.

### 3.3.2 Relasi Ekuivalensi Fuzzy (Fuzzy Equivalence Relation)

Setelah proses normalisasi, tahap selanjutnya adalah pembentukan relasi ekuivalensi *fuzzy* antar sampel. Relasi ini bertujuan untuk merepresentasikan tingkat kemiripan antar sampel berdasarkan nilai ekspresi suatu fitur miRNA secara kontinu, tidak bersifat biner seperti pada pendekatan klasik.

Dalam penelitian ini, relasi ekuivalensi *fuzzy* dihitung menggunakan fungsi keanggotaan berbasis Gaussian [21]. Pemilihan fungsi Gaussian didasarkan pada kemampuannya dalam merepresentasikan kedekatan antar nilai secara halus dan stabil terhadap variasi data, yang sangat sesuai untuk karakteristik data biologis yang bersifat noisy dan tidak pasti [11].

$$R(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^2}{2\sigma^2}\right) \tag{3.2}$$

di mana:

- $x_i$ dan $x_j$ adalah nilai fitur miRNA dari dua sampel yang dibandingkan,
- $\sigma$ merupakan parameter lebar (*spread*) dari fungsi Gaussian,
- $R(x_i, x_j) \in [0, 1]$ menyatakan tingkat kemiripan *fuzzy* antara dua sampel.

Nilai relasi yang mendekati 1 menunjukkan tingkat kemiripan yang tinggi, sedangkan nilai yang mendekati 0 menunjukkan perbedaan yang signifikan antar sampel.

### 3.3.3 Entropi Fuzzy (Fuzzy Entropy)

*Entropy fuzzy* digunakan untuk mengukur tingkat ketidakpastian atau keragaman informasi yang terkandung dalam suatu fitur miRNA berdasarkan relasi ekuivalensi *fuzzy*. Semakin tinggi nilai entropi, semakin besar ketidakpastian yang dimiliki fitur tersebut dalam merepresentasikan pola data.

Untuk setiap sampel, terlebih dahulu dihitung nilai rata-rata derajat keanggotaan *fuzzy* terhadap seluruh sampel lainnya. Nilai ini kemudian digunakan untuk menghitung entropi fuzzy fitur secara keseluruhan [22].

$$H_f = -\frac{1}{n}\sum_{i=1}^{n}\log_2(\phi_i) \tag{3.3}$$

dengan:

- $n$ adalah jumlah total sampel,
- $\phi_i$ merupakan rata-rata derajat keanggotaan *fuzzy* sampel ke-$i$ terhadap seluruh sampel lainnya.

Entropi fuzzy yang rendah mengindikasikan bahwa fitur memiliki struktur informasi yang lebih teratur dan berpotensi lebih relevan untuk proses klasifikasi.

### 3.3.4 Entropi Fuzzy Label (Fuzzy Entropy of Class Label)

Selain entropi pada fitur, penelitian ini juga menghitung entropi *fuzzy* pada label kelas target, yaitu stadium kanker lambung (*Early* dan *Late* stage). Entropi ini digunakan sebagai acuan ketidakpastian kelas sebelum dikaitkan dengan fitur miRNA.

Relasi *fuzzy* pada label dibentuk menggunakan relasi kesetaraan keras (*crisp equivalence*), di mana dua sampel dianggap memiliki kemiripan penuh jika berasal dari kelas yang sama, dan nol jika berasal dari kelas yang berbeda [6].

$$H(Y) = -\frac{1}{n}\sum_{i=1}^{n}\log_2(\phi_i^Y) \tag{3.4}$$

di mana $\phi_i^Y$ merepresentasikan derajat keanggotaan *fuzzy* sampel ke-$i$ terhadap kelas target.

### 3.3.5 Entropi Gabungan *Fuzzy* (Fuzzy Joint Entropy)

Entropi gabungan *fuzzy* mengukur tingkat ketidakpastian bersama antara suatu fitur miRNA dan label kelas. Relasi gabungan ini dibentuk dengan mengombinasikan relasi *fuzzy* fitur dan relasi *fuzzy* label menggunakan operator minimum sebagai representasi irisan *fuzzy* [12].

$$H(f,Y) = -\frac{1}{n}\sum_{i=1}^{n}\log_2(\phi_i^{fY}) \tag{3.5}$$

di mana $\phi_i^{fY}$ merupakan derajat keanggotaan gabungan antara fitur dan label kelas untuk sampel ke-$i$.

### 3.3.6 True Fuzzy Mutual Information (TFMI)

Nilai *True Fuzzy Mutual Information* (TFMI) digunakan sebagai ukuran utama dalam proses seleksi fitur. TFMI mengukur seberapa besar informasi yang dibagikan antara suatu fitur miRNA dan label stadium kanker dengan mempertimbangkan ketidakpastian dan sifat kontinu data melalui pendekatan *fuzzy*. [12, 13]

$$TFMI(f,Y) = H(f) + H(Y) - H(f,Y) \tag{3.6}$$

Nilai TFMI yang tinggi menunjukkan bahwa fitur miRNA tersebut memiliki hubungan informasi yang kuat dengan kelas stadium kanker. Oleh karena itu, fitur-fitur dengan nilai TFMI tertinggi dipilih sebagai subset fitur optimal untuk tahap pelatihan model *machine learning* selanjutnya.

## 3.4 Penanganan Ketidakseimbangan Kelas

Distribusi kelas pada data klinis kanker lambung umumnya bersifat tidak seimbang, di mana jumlah sampel pada stadium lanjut (*Late stage*) jauh lebih dominan dibandingkan stadium awal (*Early stage*). Ketidakseimbangan kelas ini dapat menyebabkan model *machine learning* menjadi bias terhadap kelas mayoritas, sehingga menurunkan kemampuan model dalam mengenali pola pada kelas minoritas dan menghasilkan performa prediksi yang tidak representatif [14].

Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan teknik *hybrid resampling* SMOTETomek, yang merupakan kombinasi dari metode *Synthetic Minority Over-sampling Technique* (SMOTE) dan *Tomek Links*. Pendekatan ini bertujuan tidak hanya untuk meningkatkan jumlah sampel pada kelas minoritas, tetapi juga untuk membersihkan batas keputusan antar kelas dari sampel yang bersifat ambigu atau berpotensi menjadi *noise* [15].

### 3.4.1 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE merupakan metode *oversampling* yang menghasilkan sampel sintetis baru pada kelas minoritas dengan memanfaatkan hubungan kedekatan antar sampel dalam ruang fitur. Berbeda dengan *random oversampling* yang hanya menduplikasi data, SMOTE membentuk sampel baru melalui interpolasi linier antara sampel minoritas dan tetangga terdekatnya (*k-nearest neighbors*) [23].

Secara matematis, pembentukan sampel sintetis SMOTE dapat dinyatakan sebagai berikut:

$$x_{\text{new}} = x_i + \lambda \cdot (x_{nn} - x_i), \quad \lambda \in [0,1] \tag{3.7}$$

di mana:

- $x_i$ adalah sampel pada kelas minoritas
- $x_{nn}$ adalah salah satu tetangga terdekat dari $x_i$
- $\lambda$ adalah bilangan acak dalam interval $[0, 1]$
- $x_{\text{new}}$ merupakan sampel sintetis yang dihasilkan

Dengan mekanisme ini, SMOTE mampu memperluas distribusi kelas minoritas secara lebih alami dan kontinu, sehingga membantu model mempelajari karakteristik kelas minoritas dengan lebih baik.

### 3.4.2 Tomek Links

Meskipun SMOTE efektif dalam meningkatkan keseimbangan kelas, proses *oversampling* dapat memperkenalkan sampel sintetis yang berada terlalu dekat dengan kelas mayoritas. Oleh karena itu, digunakan metode Tomek Links sebagai tahap *undersampling* untuk membersihkan batas kelas [23].

Tomek Link didefinisikan sebagai pasangan dua sampel $(x_i, x_j)$ dari kelas yang berbeda yang saling menjadi tetangga terdekat satu sama lain. Keberadaan pasangan ini menunjukkan area tumpang tindih (*overlapping*) antar kelas. Dalam penelitian ini, sampel dari kelas mayoritas yang terlibat dalam Tomek Link dihapus untuk memperjelas batas keputusan antar kelas.

### 3.4.3 Kombinasi SMOTETomek

SMOTETomek menggabungkan keunggulan SMOTE dan Tomek *Links* dalam satu pipeline resampling. Tahapan kerjanya dapat dirangkum sebagai berikut:

1. Menghasilkan sampel sintetis pada kelas minoritas menggunakan SMOTE.
2. Mengidentifikasi pasangan Tomek *Links* pada data hasil oversampling.
3. Menghapus sampel mayoritas yang terlibat dalam Tomek *Links* untuk mengurangi ambiguitas dan *noise*.

Pendekatan ini menghasilkan dataset pelatihan yang lebih seimbang sekaligus memiliki batas kelas yang lebih jelas, sehingga meningkatkan stabilitas dan performa model klasifikasi [15].

### 3.4.4 Penerapan pada Pipeline Penelitian

Dalam penelitian ini, teknik SMOTETomek hanya diterapkan pada data pelatihan di setiap fold *Stratified K-Fold Cross-Validation*. Pendekatan ini bertujuan untuk mencegah terjadinya *data leakage* antara data pelatihan dan data pengujian, serta memastikan bahwa evaluasi performa model tetap valid dan tidak bias [24].

Dengan penerapan SMOTETomek, diharapkan model *machine learning* mampu mempelajari pola diskriminatif antara stadium awal dan stadium lanjut kanker lambung secara lebih seimbang dan robust, terutama dalam konteks data genomik berdimensi tinggi seperti ekspresi miRNA.

### 3.5 Machine Learning Model

Pada penelitian ini, tiga algoritma *machine learning* digunakan untuk melakukan klasifikasi stadium kanker berdasarkan ekspresi miRNA, yaitu *Support Vector Machine* (SVM), *Random Forest* (RF), dan *K-Nearest Neighbors* (KNN). Setiap model dipilih berdasarkan kemampuan mereka dalam menangani dataset tidak seimbang dan potensi multikolinearitas antar fitur. Tabel 3.3 memperlihatkan parameter utama yang digunakan untuk membangun ketiga model *machine learning* dalam penelitian ini. Untuk SVM, kernel diubah menjadi RBF agar mampu menangkap hubungan non-linear antar fitur, dengan C=1 untuk menyeimbangkan regularisasi dan generalisasi, gamma diset ke *scale*, serta *class weight* seimbang untuk mengatasi ketidakseimbangan kelas. Pada *Random Forest*, jumlah *tree* ditingkatkan dari default 100 menjadi 200 untuk meningkatkan stabilitas model, dan *class weight* disesuaikan agar kelas minoritas tetap diperhatikan. Sementara pada KNN, jumlah tetangga ditingkatkan menjadi 7 dan bobot *distance-based* digunakan untuk memberikan pengaruh lebih besar pada tetangga yang lebih dekat, dengan jarak Euclidean sebagai metrik utama. Penentuan parameter ini dilakukan melalui eksperimen awal dan literatur terkait agar model dapat memberikan performa optimal tanpa overfitting.

Tabel 3.3. Parameter default dan nilai yang digunakan untuk tuning model machine learning

| **Model** | **Parameter** | **Default** | **Yang Digunakan** |
|---|---|---|---|
| SVM (RBF) | Kernel | linear | *Radial Basis Function* |
| | C (Regularisasi) | 1.0 | 1 |
| | Gamma | auto | scale |
| | Class weight | None | balanced |
| Random Forest | Jumlah *trees* | 100 | 200 |
| | *Criterion* | Gini | Gini |
| | Class weight | None | balanced |
| | Random state | None | 42 |
| K-Nearest Neighbors | Jumlah tetangga ($k$) | 5 | 7 |
| | Bobot | uniform | distance |
| | Metode jarak | Minkowski (p=2) | Euclidean |

Catatan: Tabel menunjukkan perbandingan antara nilai default dan parameter yang digunakan setelah proses tuning untuk membangun model akhir pada dataset klasifikasi stadium kanker.

#### 3.5.1 Support Vector Machine (SVM)

SVM digunakan dengan kernel *Radial Basis Function (RBF)* yang mampu menangkap hubungan non-linear antar fitur. Parameter *regularization C* diset ke 1 untuk menyeimbangkan kompleksitas model dan kemampuan generalisasi, sedangkan *gamma* menggunakan nilai *scale* untuk mengatur pengaruh setiap sampel terhadap pembentukan *decision boundary* [25]. SVM juga menggunakan *class weight* seimbang (*balanced*) untuk menangani ketidakseimbangan kelas pada dataset.

### 3.5.2 Random Forest (RF)

RF adalah metode *ensemble* berbasis *bagging* yang membangun banyak *decision tree* secara independen dan mengambil keputusan melalui *majority voting*. Model ini dikonfigurasi dengan 200 *trees* dan *class weight* seimbang untuk mengurangi bias terhadap kelas mayoritas. Keunggulan RF adalah kemampuan untuk menangkap interaksi kompleks antar fitur serta memberikan informasi penting tentang kontribusi setiap fitur terhadap klasifikasi [26].

### 3.5.3 K-Nearest Neighbors (KNN)

KNN adalah algoritma *instance-based learning* yang melakukan klasifikasi berdasarkan kedekatan jarak antar sampel. Parameter $k$ ditetapkan sebanyak 7 tetangga terdekat dan pemberian bobot *distance-based* memastikan sampel yang lebih dekat memiliki pengaruh lebih besar pada prediksi akhir [27]. KNN tidak membuat asumsi distribusi data dan sensitif terhadap skala, sehingga normalisasi fitur menjadi langkah penting sebelum pelatihan.

## 3.6 Metrik Evaluasi Kinerja Model

Kinerja model klasifikasi dievaluasi menggunakan sejumlah metrik yang relevan dengan konteks medis, yaitu *accuracy*, *precision*, *recall* (sensitivitas), *specificity*, *F1-score*, dan *Area Under the Curve (AUC)*. Pemilihan metrik ini bertujuan untuk memberikan gambaran komprehensif terhadap performa model, baik dari aspek keseluruhan prediksi maupun kemampuan model dalam membedakan kelas minoritas dan mayoritas. Definisi formal dari setiap metrik adalah sebagai berikut [28].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.8}$$

$$Precision = \frac{TP}{TP+FP} \tag{3.9}$$

$$Recall = \frac{TP}{TP+FN} \tag{3.10}$$

$$Specificity = \frac{TN}{TN+FP} \tag{3.11}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.12}$$

Di samping metrik di atas, AUC digunakan untuk menilai kemampuan model dalam membedakan kelas positif dan negatif pada berbagai ambang keputusan. Nilai AUC memberikan indikasi seberapa

baik model dapat memisahkan kelas, dengan nilai 1 menunjukkan pemisahan sempurna dan 0,5 menunjukkan performa sebanding dengan tebakan acak.

## 3.7 Analisis Confusion Matrix dan Kurva ROC

### 3.7.1 Confusion Matrix

*Confusion matrix* dianalisis untuk mengevaluasi distribusi prediksi benar dan salah pada masing-masing kelas. Analisis dilakukan baik dalam bentuk matriks agregat maupun matriks ternormalisasi (*normalized confusion matrix*). Pendekatan ini memungkinkan identifikasi pola kesalahan klasifikasi, terutama untuk stadium lanjut yang memiliki prioritas klinis lebih tinggi. Dengan demikian, evaluasi berbasis *confusion matrix* tidak hanya menilai performa keseluruhan, tetapi juga memberikan informasi granular mengenai kesalahan spesifik yang dilakukan model.

### 3.7.2 Kurva ROC

Kurva ROC (*Receiver Operating Characteristic*) dihitung untuk masing-masing fold pada *cross-validation*, kemudian dirata-ratakan untuk menilai stabilitas performa model. Kurva ROC rata-rata memberikan gambaran menyeluruh mengenai kemampuan diskriminatif model dalam membedakan kelas pada seluruh iterasi. Nilai AUC yang diperoleh dari kurva ini menjadi indikator utama kualitas prediksi model, khususnya dalam konteks klasifikasi yang tidak seimbang.

## 3.8 Dokumentasi dan Analisis Stabilitas Fitur

Tahap akhir metodologi meliputi dokumentasi dan analisis stabilitas fitur. Dokumentasi mencakup pelaporan performa model, visualisasi *confusion matrix*, kurva ROC, serta analisis frekuensi kemunculan fitur miRNA yang terpilih pada seluruh fold dan lintas model. Analisis frekuensi fitur ini bertujuan untuk mengidentifikasi miRNA yang secara konsisten muncul sebagai prediktor penting, sehingga memiliki potensi untuk dijadikan kandidat biomarker stadium kanker lambung. Dengan pendekatan ini, penelitian tidak hanya menghasilkan model klasifikasi yang optimal, tetapi juga memberikan insight biologis mengenai fitur-fitur yang paling relevan dengan progresi penyakit.