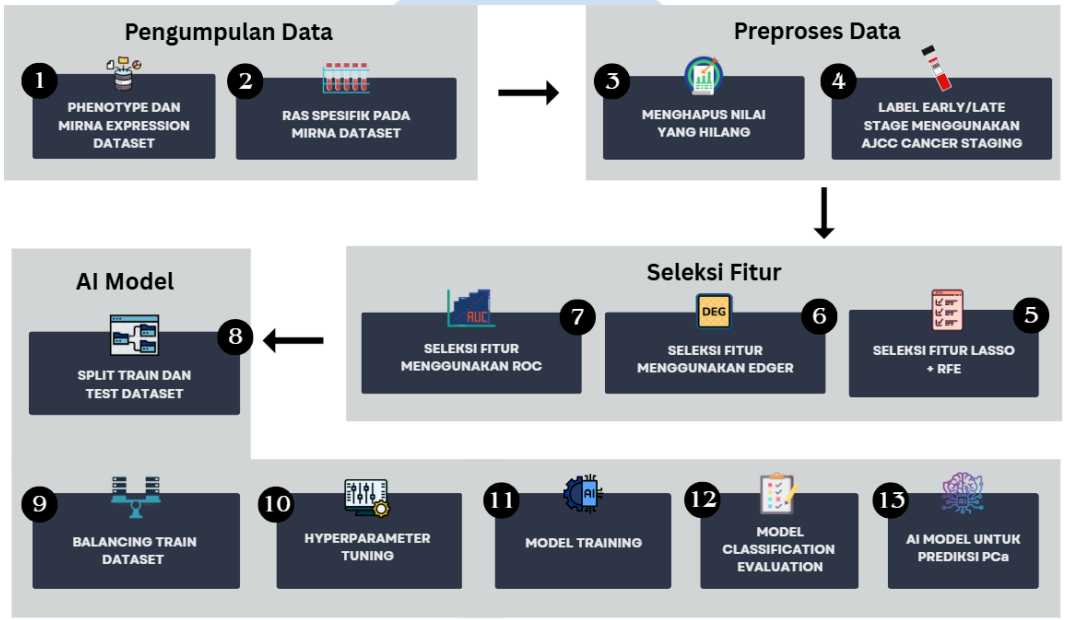
BAB III
METODE PENELITIAN

Penelitian ini menggunakan rancangan metode sesuai pada Gambar 3.1. Penelitian ini menggunakan metode yang terdiri dari pengumpulan data, preproses data, pemilihan fitur dan pembuatan model menggunakan algoritma Xgboost. Perangkat yang digunakan selama menjalankan proses adalah Windows 11 OS, dengan 16GB RAM, *processor* 12th Gen Intel® Core™ i5-12500H (16 CPUs), 2.5Hz dan NVIDIA GeForce RTX 3050 4GB.



Gambar 3. 1 Alur Penelitian Prediksi Kanker Prostat

3.1.Pengumpulan Data

Pada pengumpulan data, data yang digunakan berupa data stem loop – miRNA expression dan GDC TCGA-PRAD phenotype yang diambil pada 3 Febuari 2025. Pada data stem loop – miRNA expression berisi nilai ekspresi miRNA dari pasien yang dapat digunakan untuk melakukan identifikasi terhadap kanker prostat. Dataset miRNA expression yang digunakan sudah di normalisasi dengan unit log2(RPM+1). Selain itu, data GDC TCGA-PRAD phenotype berisi data klinik dari pasien seperti nilai T, nilai M, PSA, *Gleason score* dan *age*.

3.2 Preproses Data

Pada tahap preproses data, dataset ekspresi miRNA dipisahkan berdasarkan ras menggunakan dataset fenotip sebagai acuan. Data yang dipisahkan, dilakukan proses *cleaning* terhadap nilai kosong. Setelah melakukan proses awal pada preproses data, pemberian label dapat dilakukan untuk kanker tahap 1 hingga 4 yang kemudian dikelompokkan menjadi *early stage* dan *late stage*, tahap 1 dan 2 akan masuk pada *early stage* lalu tahap 3 dan 4 akan masuk pada *late stage*. Label yang dilakukan berdasarkan dari nilai T, N, M, *gleason score* dan PSA. Aturan yang digunakan untuk pengelompokkan label dapat dilihat pada Tabel 3.1. Aturan yang digunakan seperti nilai T, N, M, *gleason score* dan PSA didapat dari dataset fenotip. Paket seperti Pandas versi 2.2.2 dan Numpy versi 1.26.4 digunakan untuk melakukan preproses data. Data di proses menggunakan scikit-learn versi 1.5.1 untuk dilakukan label.

Tabel 3. 1 Syarat pemberian kanker tahap 1-4 [30]

Group	T	N	M	PSA	Gleason
I	T1a–c	N0	M0	PSA < 10	Gleason ≤ 6
	T2a	N0	M0	PSA < 10	Gleason ≤ 6
	T1–2a	N0	M0	PSA X	Gleason X
IIA	T1a–c	N0	M0	PSA < 20	Gleason 7
	T1a–c	N0	M0	PSA ≥10 < 20	Gleason ≤ 6
	T2a	N0	M0	PSA ≥10 < 20	Gleason ≤ 6
	T2a	N0	M0	PSA < 20	Gleason 7
	T2b	N0	M0	PSA < 20	Gleason ≤ 7
	T2b	N0	M0	PSA X	Gleason X
IIB	T2c	N0	M0	Any PSA	Any Gleason

	T1–2	N0	M0	PSA ≥ 20	Any Gleason
	T1–2	N0	M0	Any PSA	Gleason ≥ 8
III	T3a–b	N0	M0	Any PSA	Any Gleason
IV	T4	N0	M0	Any PSA	Any Gleason
	Any T	N1	M0	Any PSA	Any Gleason
	Any T	Any N	M1	Any PSA	Any Gleason

3.3 Seleksi Fitur

Setelah mendapatkan data yang sudah memiliki label *early* dan *late stage*, seleksi fitur dilakukan untuk mendapatkan fitur miRNA yang lebih signifikan dan berkorelasi terhadap kanker. RFE dan *Lasso* digunakan untuk mencari fitur yang signifikan dengan memasang jumlah fitur yang di inginkan pada RFE. Selain itu, seleksi fitur menggunakan Teknik bioinformatika juga digunakan yaitu EdgeR untuk mendapatkan fitur miRNA signifikan. Pemilihan fitur menggunakan EdgeR berfokus pada nilai dari logFC dan nilai *Pvalue*. Metode seleksi fitur yang digunakan untuk mendapatkan fitur miRNA signifikan dapat dilihat pada Tabel 3.2. Scikit-learn versi 1.5.1 merupakan Paket yang digunakan untuk RFE dan *Lasso* dan EdgeR mengunakan paket versi 4.0.16. ROC juga digunakan untuk memilih fitur miRNA yang signifikan dari dari awal. Ekspresi miRNA yang sudah didapat akan dilakukan pengecekan menggunakan website *miR2Disease, miRCancer* dan *dbDEMC (Database of Differentially Expressed miRNAs in human Cancers)*.

Tabel 3. 2 Metode Pemilihan Fitur

No.	Metode Seleksi Fitur	Kriteria
1	EdgeR	PValue < 0.01 dan logFC > 0.4
2	Lasso + RFE	Hasil terbaik dari 15, 18, dan 19 fitur miRNA
3	ROC	AUC > 0.6

3.4 Pembuatan Model

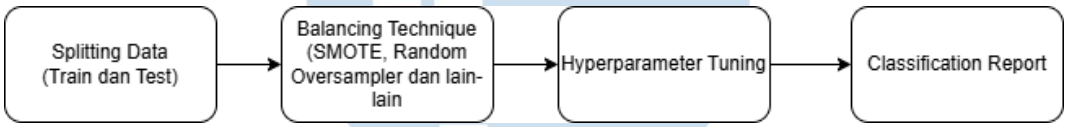
Pada pembuatan model, data displit untuk training dan tes dengan ratio 80/20, 70/30 dan 60/40. Dataset yang tidak seimbang diurus menggunakan *balancing technique* dengan rasio 1:3. Dalam menyeimbangan data, hanya data pada *training* yang diseimbangkan untuk menghindari data palsu pada tes. *Balancing technique* yang digunakan pada penelitian ini adalah *Randomoversampler, SMOTE, SMOTEEN* dan *BorderlineSMOTE*. Paket yang digunakan untuk menyeimbangkan data adalah imbalance-learn dengan versi 0.12.3. Pada Model terdapat beberapa skenario yang berfokus pada seleksi fitur menggunakan RFE *Lasso*, ROC dan EdgeR. Dalam pembuatan model *hyperparameter* tuning seperti *grid* search digunakan dengan *cross validation* untuk mendapatkan hasil yang lebih optimal. Selain itu, *hyperparameter* seperti max depth, learning rate juga digunakan untuk pembuatan model. Hasil yang didapat setelah melewati semua skenario akan dilakukan validasi menggunakan *AUC Score* Skenario yang digunakan dalam Pembangunan model dapat dilihat pada Tabel 3.3.

Tabel 3. 3 Skenario Pembuatan Model Menggunakan Xgboost

Feature Selection	*Splitting Ratio*	*Balancing Technique*
EdgeR dengan kriteria PValue < 0.01 and logFC > 0.4	80:20 70:30 60:40	*RandomOversampler* *SMOTEEN* *KMeans SMOTE* *Borderline SMOTE*
Lasso + RFE dengan hasil terbaik dari 15, 18 dan 19 fitur	80:20 70:30 60:40	*RandomOversampler* *SMOTEEN* *KMeans SMOTE* *Borderline SMOTE*
ROC dengan kriteria > 0.6	80:20 70:30 60:40	*RandomOversampler* *SMOTEEN*

		KMeans SMOTE *Borderline SMOTE*

Model yang didapat ditunjukkan hasilnya pada *classification report* dan *confusion matrix* yang memiliki matriks akurasi, presisi, *recall* dan *f1-score*. Setiap parameter memiliki fungsi masing-masing yang sangat berguna untuk menentukan apakah hasil yang digunakan sudah bagus dan sesuai dengan kriteria. Proses pembangunan model dapat dilihat pada Gambar 3. Model yang sudah didapat akan dilakukan tes menggunakan data yang berbeda yaitu data ras kulit hitam untuk mengetahui signifikansi dari perbedaan ras.



Gambar 3. 2 Pipeline Pembangunan Model