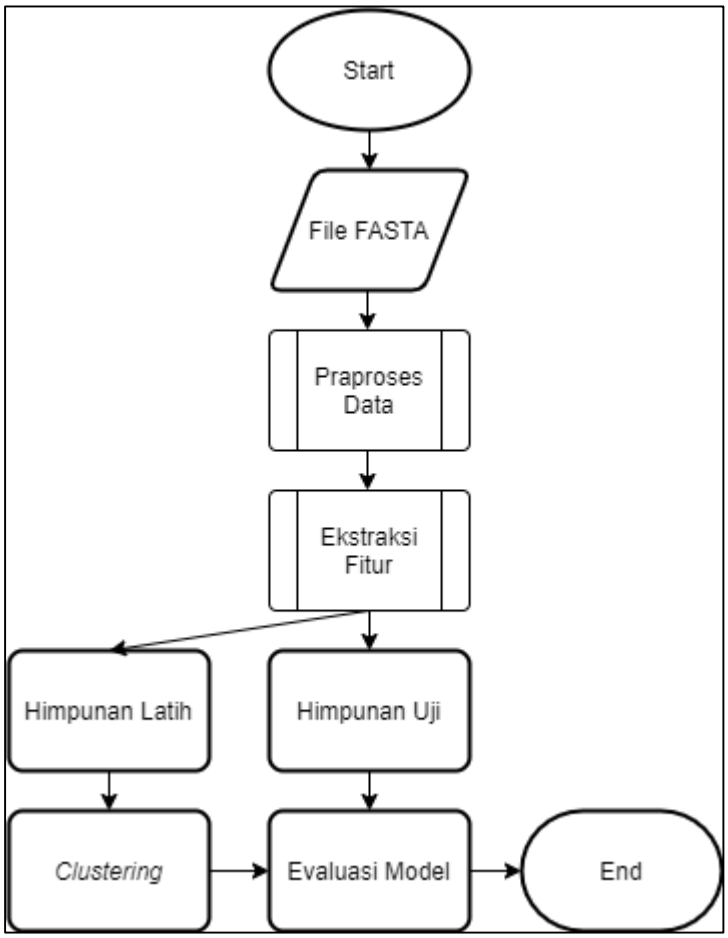
BAB 3

METODOLOGI PENELITIAN

Penelitian ini akan dilakukan secara bertahap dengan tahapan sebagai berikut.



Gambar 3.1. Gambaran alur proses penelitian

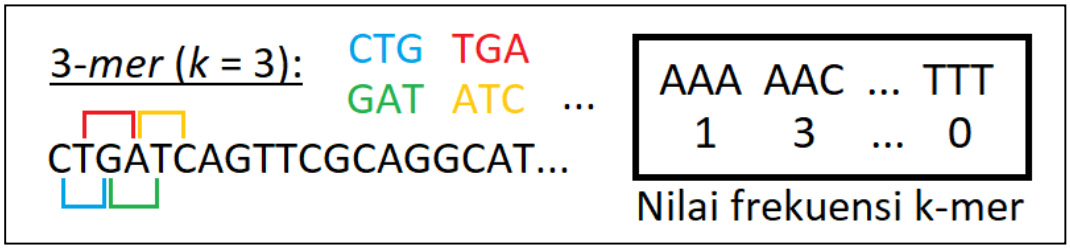
3.1. Data Genom Mikroorganisme

Data genom untuk setiap spesies mikroorganisme diperoleh dari situs NCBI (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/) dan diolah menggunakan peranti MetaSim menjadi keluaran data metagenom dalam bentuk

file format FASTA. Himpunan data yang digunakan dalam penelitian ini terdiri atas delapan puluh spesies yang berasal dari sepuluh genus, sebagaimana terlampir pada Lampiran 1. Penelitian akan dilakukan terhadap besaran panjang fragmen yang berbeda-beda, mulai dari 0,5 Kbp (*kilo base pair*; 10^3 bp), 1 Kbp, 5 Kbp, hingga 10 Kbp. Pembobotan total dari semua spesies dalam fragmen akhir diterapkan sebesar 10.000.

3.2. Ekstraksi Ciri

Hasil keluaran data FASTA terlebih dahulu dipersiapkan melalui prosedur praproses sebelum data dapat digunakan oleh model pembelajaran mesin. Prosedur praproses di sini dimulai dengan ekstraksi ciri menggunakan metode *k-mers*, di mana setiap untaian fragmen diamati untuk dicatat seberapa sering setiap kombinasi pasangan basa (*base pair*) A, T, G, dan C, sebagai komponen penentu ciri-ciri/karakteristik mikroorganisme, muncul pada untaian. Panjang kombinasi pasangan basa yang akan diamati di sini adalah *3-mer* seperti AAA, AAC, AAG, sampai dengan TTC, TTG, TTT (4^3 = 64 kombinasi), serta *4-mer* seperti AAAA, AAAC, sampai dengan TTTG, TTTT (4^4 = 256 kombinasi). Pembedaan kombinasi ini penting, karena DNA pada dasarnya merupakan materi genetika dalam hampir semua makhluk hidup, termasuk mikroorganisme dan juga manusia, di mana kombinasi yang berbeda akan menghasilkan karakteristik fisiologis yang berbeda pula (*What is DNA?: MedlinePlus Genetics*, 2021).



Gambar 3.2. Ilustrasi penghitungan frekuensi *k-mer* pada nilai *k* = 3

3.3. Normalisasi Data

Data fragmen metagenom yang sudah diekstraksi ciri kemudian dinormalisasi dengan metode *min-max scaling*. Dalam normalisasi *min-max*, semua nilai fitur diskala ulang ke dalam suatu rentang nilai yang baru dengan tetap mempertahankan perbandingan selisih antar nilai. Pada penelitian ini, rentang nilai yang digunakan adalah [0, 1].

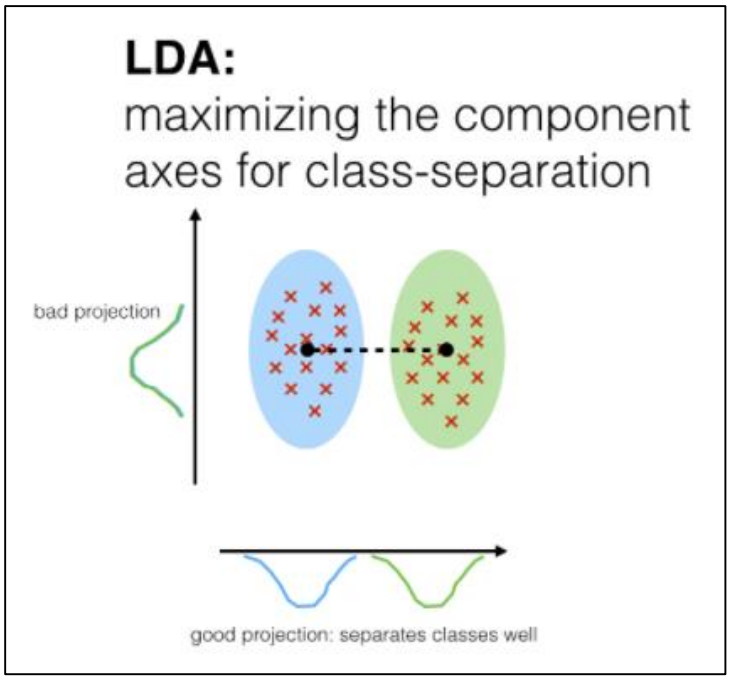
AAA	AAC	AAG	AAT	...	TTA	TTC	TTG	TTT
10	7	0	3		5	5	4	9
2	8	9	1		6	11	0	3
3	10	6	0		17	4	2	16
7	1	4	3		4	5	6	3

AAA	AAC	AAG	AAT	...	TTA	TTC	TTG	TTT
0.588235	0.411765	0	0.176471		0.294118	0.294118	0.235294	0.529412
0.117647	0.470588	0.529412	0.058824		0.352941	0.647059	0	0.176471
0.176471	0.588235	0.352941	0		1	0.235294	0.117647	0.941176
0.411765	0.058824	0.235294	0.176471		0.235294	0.294118	0.352941	0.176471

Gambar 3.3. Ilustrasi *min-max scaling*

3.4. Reduksi Dimensi – Linear Discriminant Analysis

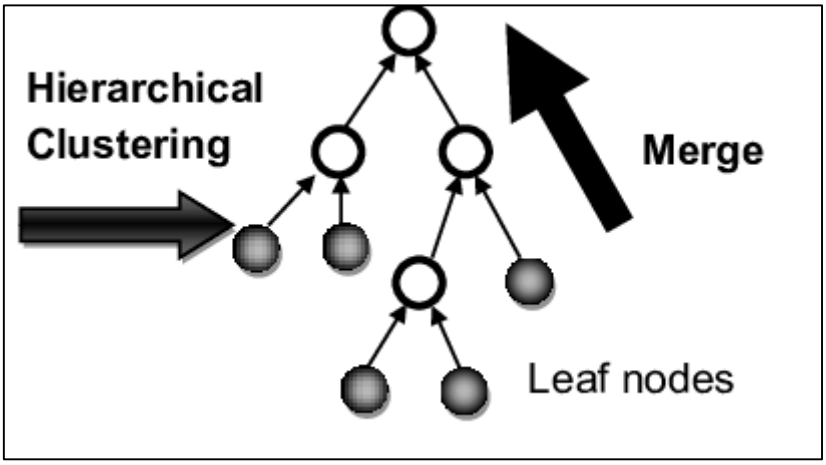
Setelah normalisasi selesai, dilakukan reduksi dimensi terhadap himpunan data sebelum siap untuk digunakan dalam model pembelajaran mesin. Reduksi dimensi bertujuan untuk menghilangkan ciri-ciri/fitur-fitur yang redundan dengan mentransformasikan ciri-ciri dari matriks data yang berdimensi lebih besar ke dalam dimensi yang lebih kecil (Tharwat dkk., 2017). Algoritma reduksi yang digunakan dalam penelitian ini adalah *Linear Discriminant Analysis* (LDA). Pada penelitian ini, algoritma LDA dijalankan dengan menempuh tahapan-tahapan yang telah dijabarkan oleh Raschka dan Mirjalili (2019) sebelumnya, terkecuali untuk standarisasi data. Ini karena data awal sudah dinormalisasikan terlebih dahulu pada subbab 3.3. di atas, sehingga prosedur LDA di sini dapat dilanjutkan tanpa harus melakukan standarisasi ulang.



Gambar 3.4. Ilustrasi praktek *linear discriminant analysis* (Raschka, 2014)

3.5. Pengelompokan – Agglomerative Hierarchical Clustering

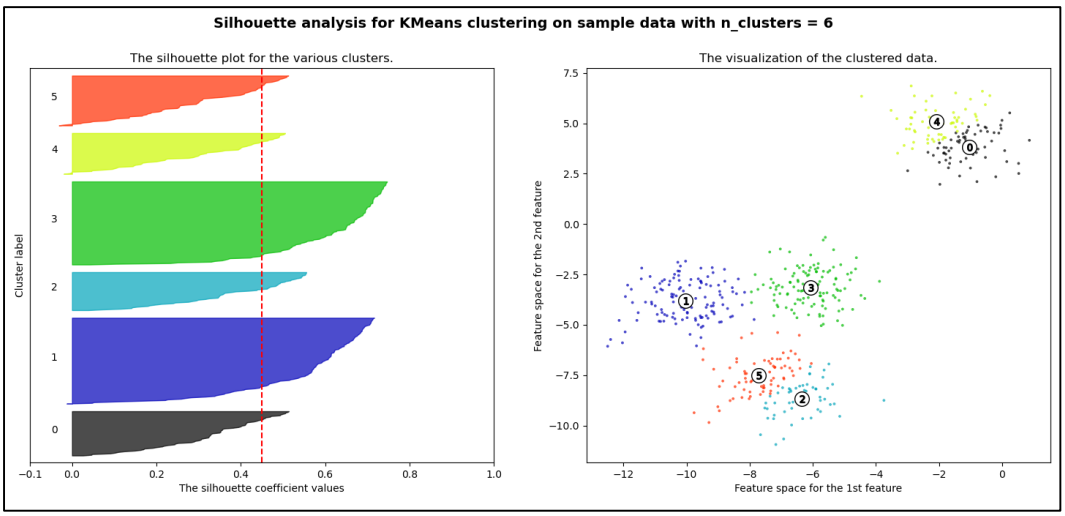
Analisis kemudian dilakukan terhadap himpunan data latih terlebih dahulu untuk melihat dan menentukan kondisi-kondisi dengan tingkat akurasi yang paling optimal. Kondisi optimal tersebut kemudian dijadikan landasan analisis terhadap himpunan data uji dalam rangka pengelompokan genus dari himpunan data itu. Algoritma yang digunakan untuk tahap pengujian ini adalah *Agglomerative Hierarchical Clustering*. Hasil pengelompokan yang diperoleh dari ini kemudian dibandingkan dengan penelitian oleh Simangusong (2015) dari segi akurasinya.



Gambar 3.6. Ilustrasi praktek *agglomerative hierarchical clustering* (Wu *et al.*, 2012)

Untuk meninjau tingkat keabsahan atau validitas dari setiap gugusan (*cluster*) data hasil pengelompokan oleh model, ada beberapa sistem tolak ukur yang dapat digunakan, termasuk *silhouette index* yang digunakan dalam penelitian ini. *Silhouette index* setiap sampel dihitung berdasarkan rata-rata jarak sampel tersebut terhadap semua sampel lain dalam gugusan yang sama dan juga jaraknya terhadap semua gugusan lain dengan mengambil gugusan terdekat dari posisi

sampel tersebut. Ilustrasi nilai *silhouette index* dapat dilihat pada Gambar 3.7 berikut.



Gambar 3.7. Grafik nilai *silhouette index* untuk semua sampel dari semua gugusan; tonjolan kecil di bagian kiri bawah beberapa gugusan menunjukkan *outlier* dari gugusan tersebut
(Pedregosa *et al.*, 2011)