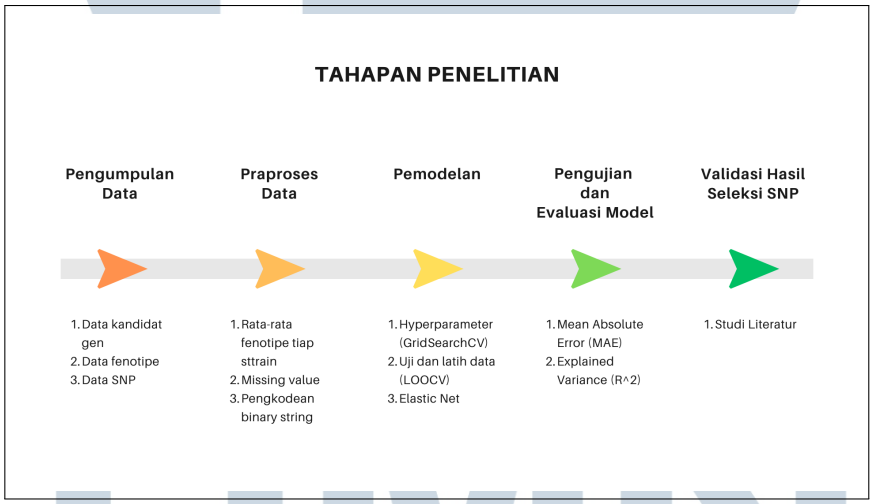
BAB 3
METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

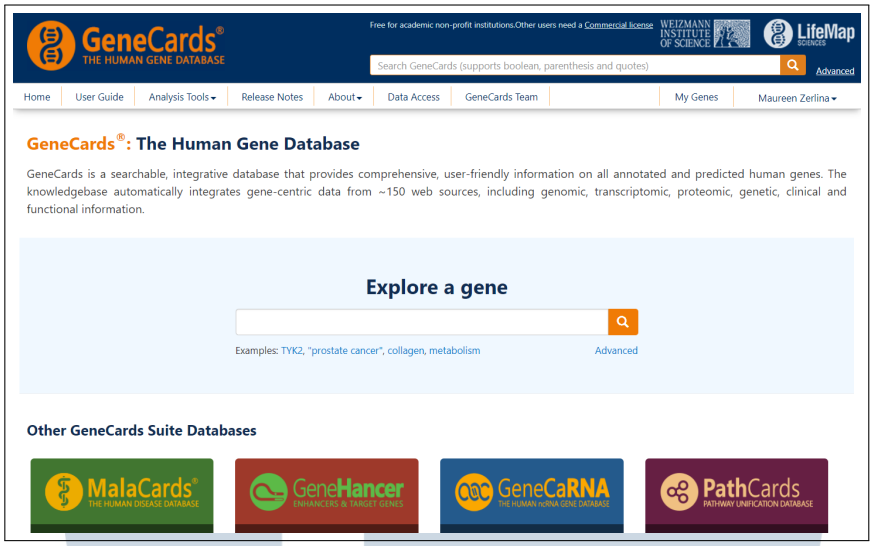
Penelitian ini diselesaikan melalui lima tahapan terstruktur yaitu pengumpulan data, praproses data, pemodelan, pengujian dan evaluasi model, dan validasi hasil seleksi SNP. Pada tahapan pertama, dikumpulkan data kandidat gen, data fenotipe *blood albumin amount*, dan data SNP. Dilanjutkan dengan tahap praproses ketiga data tersebut. Tahap ketiga yaitu pemodelan dengan menggunakan metode *Elastic Net*. Terakhir adalah tahapan pengujian serta evaluasi model, dan validasi hasil seleksi SNP. Alur tahapan penelitian disajikan pada Gambar 3.1.



Gambar 3.1. Alur tahapan penelitian

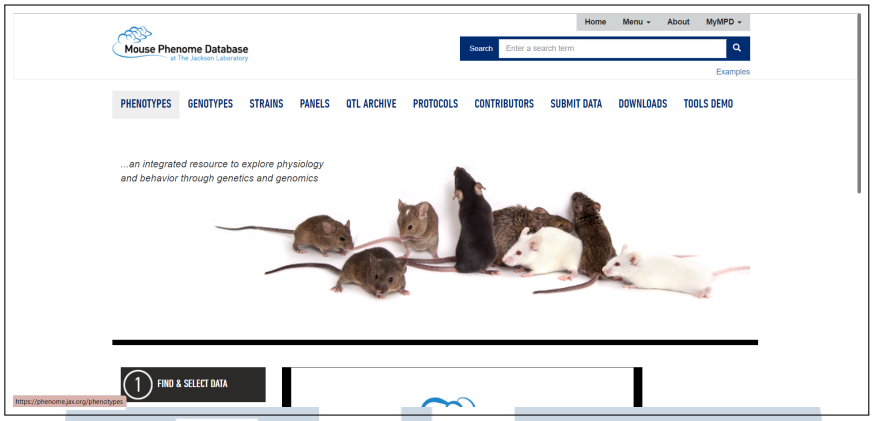
3.2 Pengumpulan Data

Data kandidat gen *blood albumin amount* yang diperlukan untuk mendapatkan data SNP, diambil melalui situs *website https://www.genecards.org/* dengan memasukan kata kunci "*blood albumin amount*".



Gambar 3.2. Tampilan *home website GeneCards*

Data kandidat gen kemudian diseleksi berdasarkan skor relevansinya dengan hanya mengambil kandidat gen yang memiliki skor relevansi $\geq$ 25. Kemudian, data kandidat gen tersebut yang digunakan dalam mengkuerikan SNP pada basis data CGD-MDA1 milik Yang W et al. [37] yang didapatkan pada *database The Mouse Phenome Database* (MPD) melalui alamat *website https://phenome.jax.org/*. Untuk data fenotipe berasal dari dataset Yuan3 milik Sinke A et al. [38] yang terdapat pengukuran untuk fenotipe *blood albumin amount*, yang juga diakses melalui situs *website* MPD. *Blood albumin amount* digunakan sebagai penanda fenotipe karena merupakan salah satu fenotipe PGK [39]. Data SNP yang diambil disesuaikan dengan ketersediaan data *strain* pada basis data fenotipe. Terdapat 31 *strain* tikus yang digunakan yaitu 129S1/SvImJ, A/J, AKR/J, BALB/cByJ, BTBR T<+>Itpr3<tf>/J, BUB/BnJ, C3H/HeJ, C57BL/6J, C57BL/10J, C57BLKS/J, C57BR/cdJ, C57L/J, CAST/EiJ, CBA/J, DBA/2J, FVB/NJ, KK/HlJ, LP/J, MRL/MpJ, NOD.B10Sn-H2<b>/J, NON/ShiLtJ, NZO/HlLtJ, NZW/LacJ, P/J, PL/J, PWD/PhJ, RIIIS/J, SJL/J, SM/J, SWR/J, dan WSB/EiJ. Data kandidat gen, data fenotipe, dan data SNP diambil pada tanggal 19 April 2023. Ketiga dataset tersebut diambil dari *website* yang cukup kredibel karena terdapat jurnal publikasi untuk setiap dataset yang digunakan pada penelitian ini. *Website genecards* dan MPD memungkinkan pengguna untuk mengunduh dataset dalam format *.csv* sehingga dapat langsung digunakan.



Gambar 3.3. Tampilan *home website The Mouse Phenome Database* (MPD)

3.3 Praproses Data

Dilakukan tiga proses pada tahap praproses data. Pertama, menghitung rata-rata nilai fenotipe, yaitu *blood albumin amount*, untuk setiap *strain*-nya. Perhitungan nilai rata-rata ini dilakukan karena hanya dibutuhkan satu nilai fenotipe yang representatif untuk setiap *strain*. Sebelum menghitung nilai rata-rata setiap *strain*, dilakukan standardisasi terlebih dahulu menggunakan fungsi *StandardScaler* dari paket *Scikit-Learn*. Kedua, memperbaiki data SNP dengan cara menghapus SNP yang mengandung *missing value* dan genotipe 'H' yang dapat menimbulkan kerancuan karena bisa diartikan menjadi basa A, basa G, basa C, atau basa T. Untuk penghapusan *missing value* dan genotipe 'H', dilakukan pelacakan index dan *string* terlebih dahulu untuk mengetahui index ke berapa yang masih mengandung *missing value* ataupun genotipe 'H', baru kemudian secara manual menghapus baris index tersebut dengan menggunakan fungsi *drop*. Ketiga, melakukan pengkodean data SNP yang direpresentasikan menggunakan *binary string* berdasarkan [40]. SNP dengan modus muncul terbanyak disebut sebagai alel mayor, sedangkan sisanya disebut dengan alel minor. Pada urutan genotipe, informasi alel dibentuk oleh variasi {A/A, A/T, A/C, A/G ... T/C, T/T}. Berdasarkan kondisi variasi genotipenya, berikut pengkodean SNP ke dalam *binary string*:

0: kedua alel merupakan mayor homozigot

1: kedua alel merupakan minor homozigot

2: kedua alel merupakan heterozigot.

Contoh pengkodean data SNP menurut Ilhan I et al. [40] dapat dilihat pada Gambar 3.4. Alel mayor homozigot diwakili oleh warna abu-abu, alel minor homozigot diwakili oleh warna hitam, dan putih untuk alel heterozigot.

	SNP1	SNP2	SNP3	SNP4	SNP5
Individu 1	A/A	T/T	T/T	G/G	A/C
Individu 2	C/G	T/A	C/C	C/G	G/G
Individu 3	A/A	A/T	T/C	C/C	C/A
Individu 4	G/C	T/T	C/C	G/G	G/G

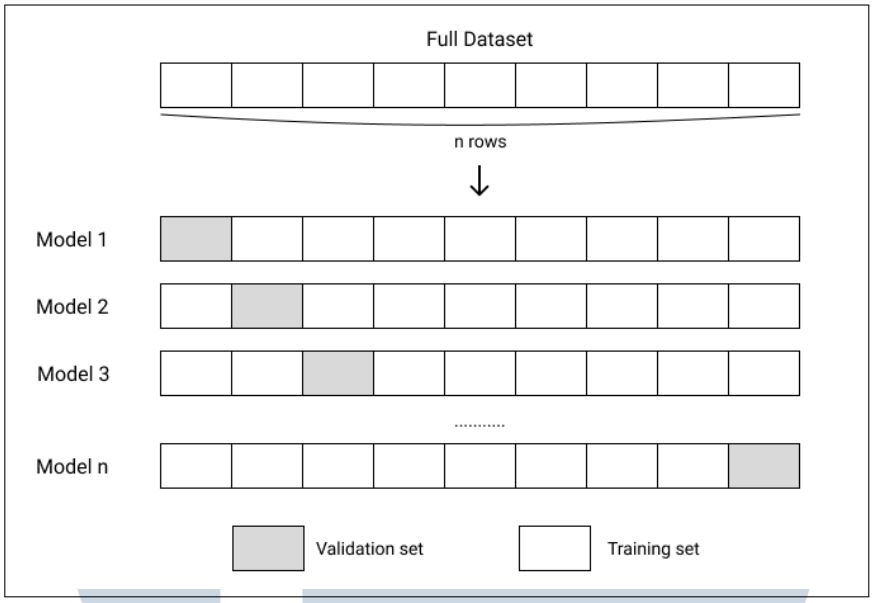
	SNP1	SNP2	SNP3	SNP4	SNP5
Individu 1	0	0	1	0	2
Individu 2	2	2	0	2	0
Individu 3	0	2	2	1	2
Individu 4	2	0	0	0	0

Gambar 3.4. Gen SNP sebelum diubah dan sesudah diubah ke dalam *binary string*

Setelah dilakukan pengkodean, nilai rata-rata *blood albumin* yang telah dihitung sebelumnya pada dataset fenotipe, digabungkan pada dataset SNP yang telah dikodekan ini berdasarkan *strain*-nya.

3.4 Pemodelan

Dengan menggunakan teknik *Grid Search Cross Validation*, dilakukan pencarian terlebih dahulu *hyperparameter* yang paling sesuai dengan model yang digunakan. *Hyperparameter* yang paling sesuai kemudian akan digunakan untuk pelatihan model. Sedangkan teknik *Leave One Out Cross Validation* (LOOCV) yang digunakan untuk pelatihan dan seleksi model. Dengan menggunakan teknik LOOCV, untuk setiap *training set* hanya satu titik data yang digunakan sebagai data uji, dan data lain yang tersisa digunakan sebagai data latih untuk semua sampel pada data. Oleh karena itu, jika terdapat 100 titik data, maka terdapat juga 100 model untuk setiap titik data yang digunakan sebagai data uji. Hal inilah dari LOOCV yang dapat mengatasi kelemahan penggunaan data dengan jumlah sampel yang sedikit karena tidak membuat partisi antara data latih dan data uji.



Gambar 3.5. Cara kerja teknik LOOCV

Metode yang digunakan untuk pemodelan *machine learning* pada tahap ini adalah *Elastic Net* dengan menggunakan paket *Scikit-Learn* yang menyediakan *linear_model* sebagai modul yang memiliki kelas *ElasticNet*. Metode ini dapat secara langsung menyeleksi fitur yang memiliki nilai *importance* yang signifikan pada saat pemodelan. Sedangkan fitur yang tidak signifikan akan diberi nilai koefisien regresi dari persamaan model = 0, sehingga fitur tersebut tidak terpilih. Himpunan SNP siginfikan yang terseleksi yang dianggap sebagai SNP yang berpengaruh terhadap fenotipe yang telah ditentukan.

3.5 Pengujian dan Evaluasi Model

Tahap pengujian dilaksanakan untuk memprediksi nilai kuantitatif variabel target, yaitu *blood albumin amount* dengan menggunakan model terbaik yang telah didapatkan pada tahap sebelumnya saat mencari nilai *hyperparameter* dengan teknik *Leave One Out Cross Validation* (LOOCV). Kemudian untuk tahap evaluasi model, menggunakan *Mean Absolute Error* (MAE) serta koefisien determinasi (R^2). Nilai MAE mempresentasikan rata-rata kesalahan absolut antara hasil prediksi dengan nilai sebenarnya, singkatnya MAE mengukur tingkat keakuratan dari hasil perhitungan metode Elastic Net dalam penelitian ini. Sedangkan pengujian (R^2) bertujuan untuk mengukur seberapa baik garis regresi. Semakin nilai (R^2) mendekati nilai 1, maka dapat diartikan bahwa variasi pada nilai albumin

dapat dijelaskan dengan baik oleh SNP sebagai prediktor yang diseleksi oleh model.

3.6 Validasi Hasil Seleksi SNP

Tahap validasi SNP yang telah diseleksi sebelumnya dilakukan dengan memetakan SNP ke gen asalnya pada data gen. Kemudian, berdasarkan penelitian terdahulu dilakukan studi literatur apakah ada penelitian yang sudah berhasil mengungkapkan asosiasi antara gen yang telah didapatkan dengan fenotipe *blood albumin amount* maupun PGK secara umum sebagai penyakit. Pencarian literatur menggunakan sumber rujukan artikel ilmiah yang kredibel seperti PubMed (https://pubmed.ncbi.nlm.nih.gov/).