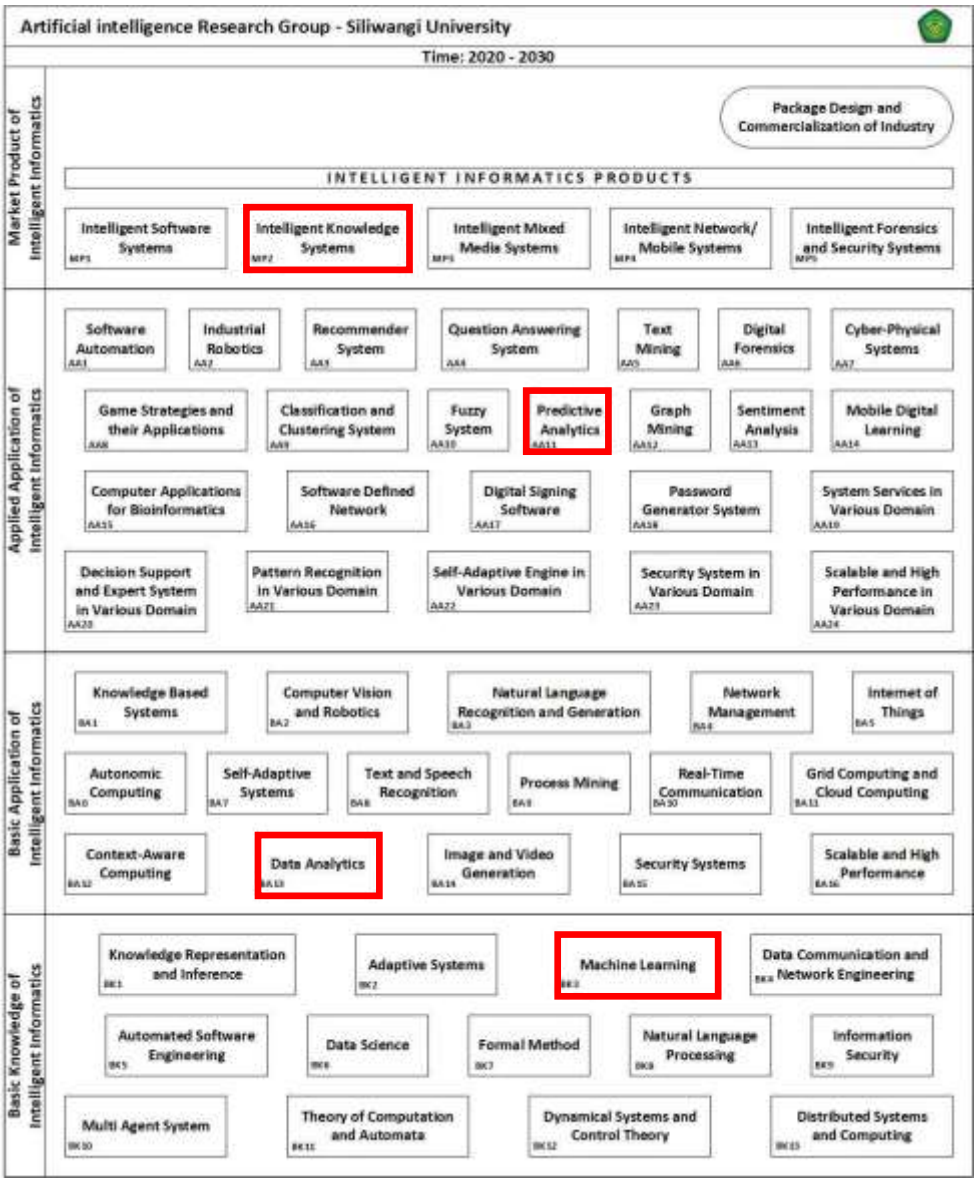
BAB III

METODOLOGI PENELITIAN

3.1 Peta Jalan (Road Map) Penelitian

Penelitian yang dijelaskan dalam proposal ini sejalan dengan *Roadmap Artificial Intelligence Research Group* – Universitas Siliwangi tahun 2020 – 2030 yang dapat dilihat pada Gambar 3.1

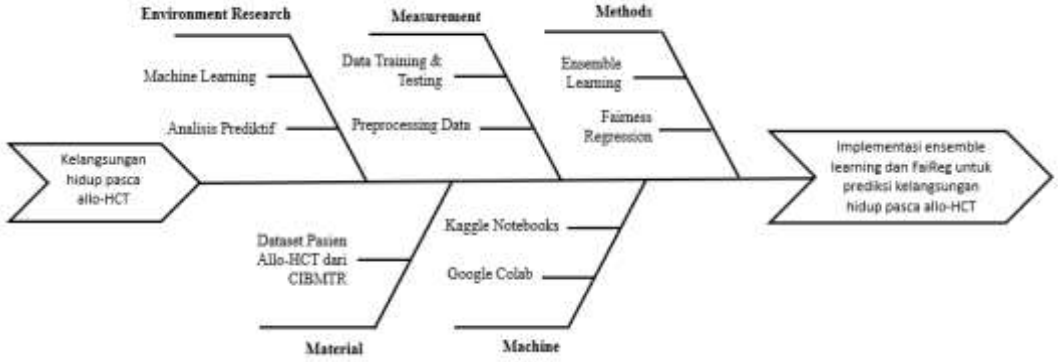


Gambar 3. 1 Roadmap Penelitian

Gambar 3.1 menunjukkan spesifikasi dari setiap lapisan yang ditandai dengan kotak merah. Penelitian yang diusulkan berkaitan dengan bidang *machine learning, data analytics*, *predictive analytics* untuk mencapai tujuan *intelligent knowledge systems*. Penelitian ini melakukan pemodelan yang berorientasi dalam memprediksi kelangsungan hidup pasien pasca allo-HCT untuk menjadi teknologi analisis prediktif dalam medis.

Penelitian ini dirancang untuk mendukung pencapaian tujuan pada bidang yang dipilih dalam *roadmap* penelitian. Dataset yang digunakan melalui proses akuisisi dan pra-pemrosesan yang mengarah kepada *data analytics*. Implementasi teknik *fairness regression* selaras dengan *predictive analytics* untuk meningkatkan keandalan sistem dalam hal keadilan. Pengembangan dan pelatihan model dilakukan menggunakan *ensemble learning* yang berada dalam fokus *machine learning*. Evaluasi model memastikan bahwa model memiliki performa yang baik dan adil. Seluruh proses tersebut memastikan penelitian berkontribusi langsung pada tujuan *intelligence knowledge systems*

Tujuan dari penelitian direpresentasikan menggunakan *Fishbone Diagram*. *Fishbone diagram* merupakan alat pemetaan pikiran yang akan menunjukkan hubungan sebab akibat antar berbagai faktor yang mempengaruhi suatu masalah. Faktor penyebab ini meliputi 4M+1E, yaitu penggunaan data (*man/measurement*), peralatan yang digunakan (*machine*), informasi yang digunakan (*material*), prosedur yang digunakan (*method*), dan kondisi yang terjadi (*environment*). Penggunaan *fishbone diagram* memungkinkan penentuan solusi dari setiap permasalahan agar tepat sasaran (Holifahtus Sakdiyah dkk., 2022).



Gambar 3. 2 *Fishbone Diagram*

Gambar 3.2 menunjukkan *fishbone diagram* untuk penelitian ini. Tujuan utama dari penelitian ini, yaitu implementasi model yang dikembangkan menggunakan *bagging* dan *boosting* dengan teknik *fairness regression* untuk prediksi kelangsungan hidup pasien pasca allo-HCT. Bagian "ekor" dari diagram mewakili objek penelitian, yaitu kelangsungan hidup pasien pasca allo-HCT, sedangkan pada bagian "kepala" dilakukan implementasi *bagging dan boosting* dengan teknik *fairness regression* untuk prediksi kelangsungan hidup pasien pasca allo-HCT. Bagian "tulang" dari diagram diuraikan sebagai berikut:

a. *Environment Research,* merupakan area penelitian yang akan menjadi fokus penelitian, di mana performa model *ML* dalam analisis prediktif menentukan tercapainya tujuan penelitian.
b. *Material,* sumber data yang digunakan dalam penelitian ini adalah dataset pasien penerima allo-HCT dari CIBMTR. Kualitas dan representasi dataset mempengaruhi kemampuan generalisasi model.
c. *Measurement,* merupakan tahapan persiapan data sebelum dilakukan pengujian. Tahap persiapan meliputi *pre-processing* untuk meningkatkan kualitas data dan

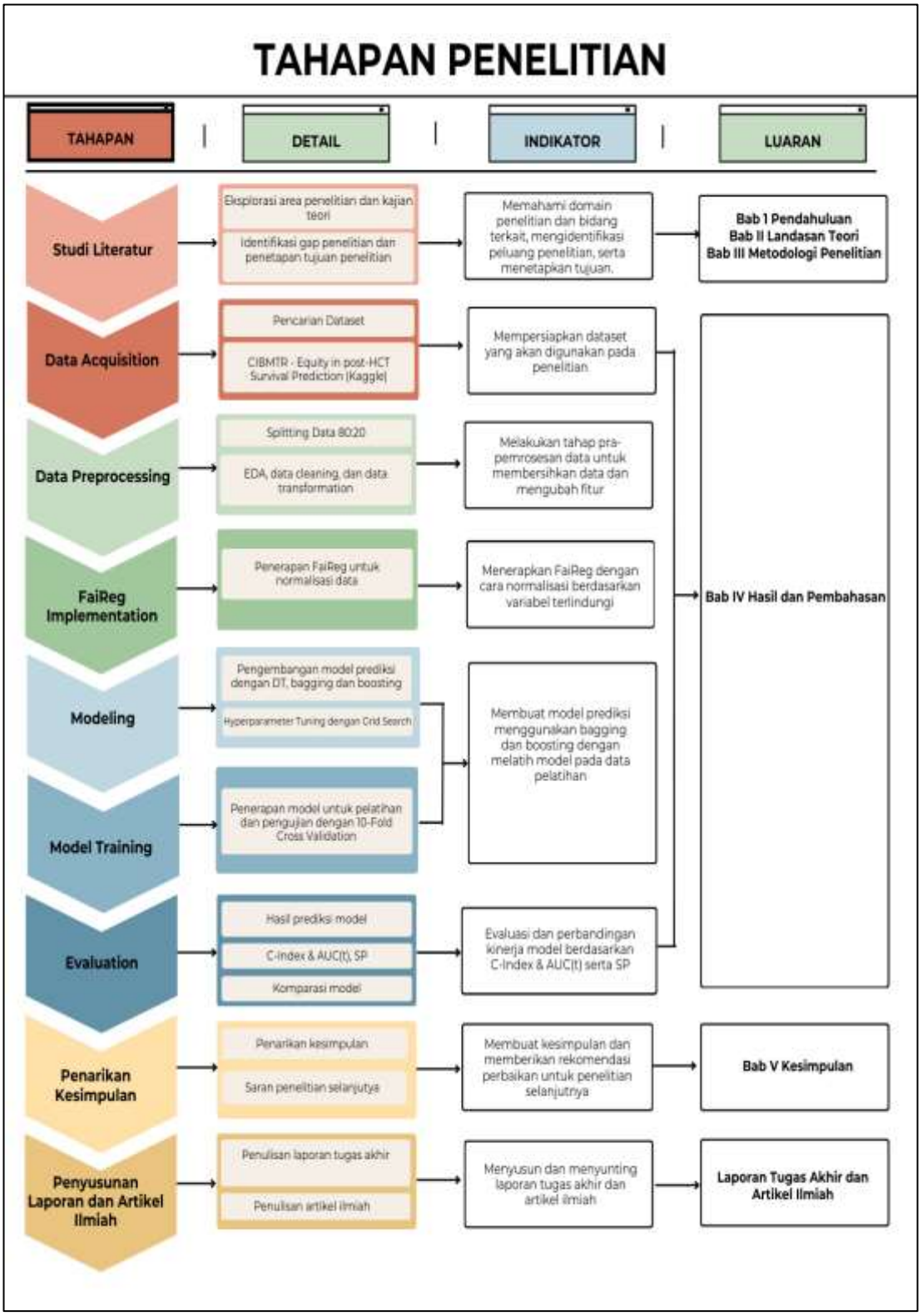
memastikan kompatibilitas data. Pembagian data menjadi data *training* dan *data testing* memastikan ketersediaan data untuk evaluasi model.

d. *Machine,* merupakan perangkat atau alat yang digunakan dalam penelitian, yaitu dengan memanfaatkan *Kaggle Notebook* dan *Google Colab* sebagai platform komputasi.

e. *Methods,* merupakan teknik atau algoritma yang digunakan dalam penelitian ini, mencakup pemanfaatan *ensemble learning* dengan teknik *fairness regression*. Strategi kombinasi metode *bagging* dan *boosting* dengan FaiReg menjadi faktor keberhasilan dalam mencapai performa model yang adil pada setiap kelompok ras.

3.2 Tahapan Penelitian

Tahapan penelitian yang diterapkan dimulai dengan tahap studi literatur, akuisisi data (*data acquisition*), pra-pemrosesan data (*data preprocessing*), implementasi *FaiReg* (*FaiReg implementation*), pemodelan (*modeling*), evaluasi model (*model evaluation*) dan penarikan kesimpulan. Rangkaian tahapan penelitian ini dapat diilustrasikan pada Gambar 3.3



Gambar 3. 3 Tahapan Penelitian

3.2.1 Studi Literatur

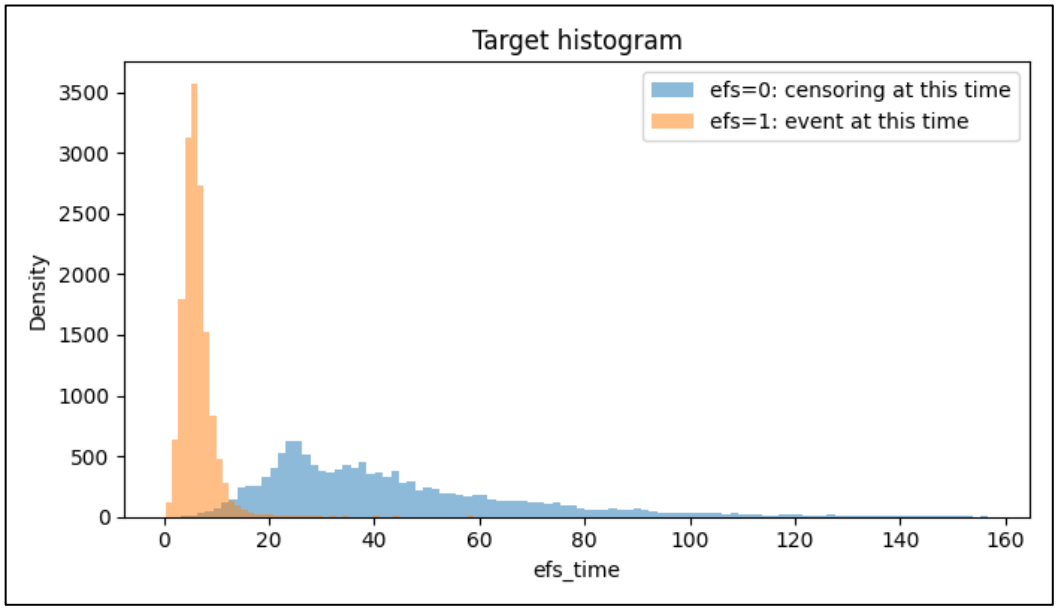
Studi literatur berisi proses pengkajian berbagai sumber informasi, termasuk jurnal penelitian, buku, situs web resmi dan laporan penelitian terkait dengan topik allo-HCT, analisis survival, *ensemble learning*, *machine learning*, dan *fairness regression*. Analisis terhadap berbagai literatur bertujuan untuk memperoleh pemahaman yang mendalam mengenai kerangka kerja konseptual, teknologi dan metodologi yang relevan dengan penelitian ini.

3.2.2 *Data Acquisition*

Dataset yang digunakan bersumber dari *Kaggle Competition CIBMTR - Equity in post-HCT Survival Predictions* yang diadakan oleh *Center for International Blood & Marrow Transplant* (CIBMTR) untuk keperluan penelitian (https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions). Data ini digeneralisasi menggunakan SurvivalGAN yang mencerminkan kelompok besar data CIBMTR nyata. Data dengan representasi yang sama di seluruh kelompok ras dibuat secara sintetis menggunakan *synthcity*. Keseluruhan proses dilakukan secara langsung oleh CIBMTR, sehingga tersedia dataset yang siap digunakan.

Dataset terdiri dari 59 variabel yang terkait dengan HCT, mencakup berbagai karakteristik demografis dan medis penerima dan donor. Target utama pada data adalah kelangsungan hidup bebas kejadian (*event-free survival*), yang diwakili oleh variabel *efs* dan waktu kelangsungan hidup bebas kejadian *efs_time*. Kedua variabel tersebut secara bersamaan digunakan untuk analisis *time-to-event* yang disensor.

Variabel *efs* menunjukkan status biner pasien berdasarkan adanya kejadian "*event*" dan bebas kejadian sampai akhir pengamatan "*censoring*". Proporsi *efs* pada data sebanyak 53,9% mengalami *event* dan 46,1% dinyatakan *censoring*. Variabel *efs_time* menunjukkan distribusi yang menyebar luas pada pasien *censoring*, sementara pasien yang mengalami *event* lebih terkonsentrasi dalam jangka waktu pendek. Distribusi tersebut menunjukkan sebagian besar kejadian terjadi pada waktu awal. Gambar 3.4 menyajikan histogram distribusi *efs_time* berdasarkan status *efs*.



Gambar 3. 4 Histogram Distribusi *efs_time* berdasarkan Status *efs*

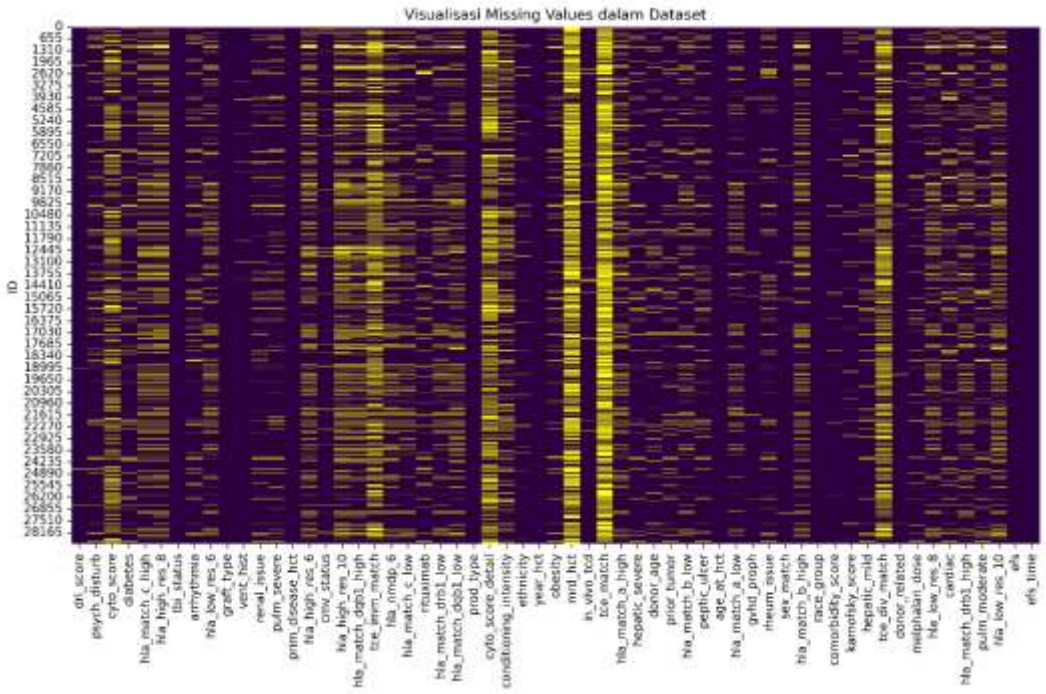
Variabel yang menyatakan kelompok ras memiliki representasi data yang hampir sama akibat penerapan *synthhcity*. Kelompok ras terdiri dari *Asian*, *White*, *Black*, *American Indian*, *Native Hawaiian or other Pacific Islander* dan *More than one race*. Setiap kelompok ras memiliki proporsi data 16,3-16,8%.

3.2.3 *Data Preprocessing*

Data preprocessing memiliki peran penting dalam mempersiapkan dataset sebelum pemrosesan lebih lanjut. *Data pre-processing* yang dilakukan pada penelitian ini dimulai dengan Exploratory Data Analysis (EDA), dilanjutkan dengan *data cleaning* dan *data transformation*. *Data cleaning* bertujuan untuk meningkatkan kualitas dataset melalui imputasi *missing values* dan teknik deteksi *outliers*. *Data transformation* bertujuan untuk memastikan kompatibilitas data pada algoritma pelatihan melalui penggabungan fitur dan transformasi fitur kategorikal.

3.2.3.1 *Data Cleaning*

Data cleaning dimulai dari pengecekan *missing values*. Bagian berwarna terang menunjukkan lokasi *missing values*, sementara bagian gelap menunjukkan data lengkap. Visualisasi *missing values* dapat dilihat pada Gambar 3.5.



Gambar 3. 5 Visualisasi *Missing Values* dalam Dataset

Missing *values* terdapat hampir pada seluruh variabel, kecuali *graft_type*, *prod_type*, *prim_disease_hct*, *tbi_status*, *year_hct*, *age_at_hct*, dan *race_group* serta kedua variabel target *efs* dan *efs_time*. *Missing value* tertinggi terdapat pada variabel *tce_match* sejumlah 18.996 data, dimana persentase total *missing values* dalam dataset adalah sebesar 11.16%. Penghapusan *missing values* tidak dilakukan karena berisiko menghilangkan informasi yang signifikan. Sehingga, penanganan *missing values* dilakukan melalui imputasi nilai "*missing*" (kategorikal) dan "-1" (numerik).

Deteksi *outliers* melibatkan aturan *interquartile range* (IQR). IQR membagi dataset menjadi empat bagian dengan rentang Q1, Q2, dan Q3. *Outliers* pada data teridentifikasi jika data berada di luar *range* Q1 − 1.5(Q3 − Q1) dan Q3 + 1.5(Q3 − Q1), di mana Q1 dan Q3 masing-masing merepresentasikan kuartil pertama dan ketiga (Yu dkk., Fan dkk., 2021). Meskipun metode berbasis *Decision Tree* cukup *robust* terhadap *outliers*, pemeriksaan ini tetap dilakukan untuk memastikan kualitas data.

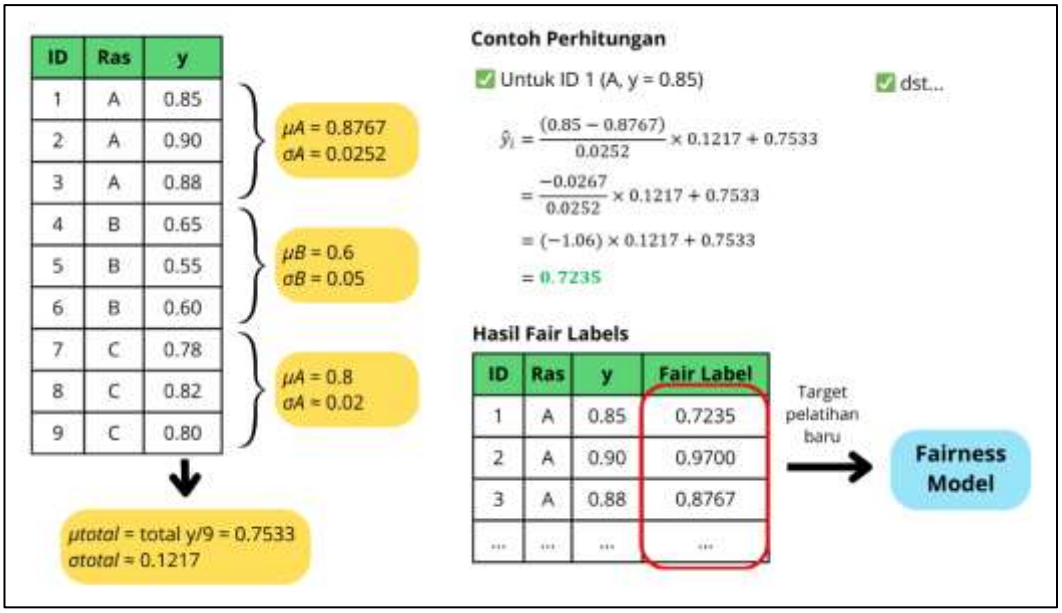
3.2.3.2 *Data Transformation*

Penggabungan fitur dilakukan pada variabel *efs* dan *efs_time* menggunakan metode Kaplan Meier yang menghasilkan variabel baru berupa probabilitas risiko. Variabel probabilitas risiko (*y*) merupakan nilai kontinu yang digunakan sebagai target utama prediksi. Selanjutnya, transformasi fitur dilakukan dengan metode *label encoding* pada fitur kategorikal serta optimasi tipe data untuk efisiensi memori. *Label encoding* bertujuan untuk mengubah data kategorikal menjadi representasi numerik agar dapat digunakan dalam pelatihan.

Proses encoding dilakukan dengan mempelajari label yang terdapat pada data pelatihan, kemudian setiap kategori dikonversi menjadi representasi numerik. Tipe data kemudian diubah ke *int32* dan dikembalikan sebagai tipe *category* untuk efisiensi memori. Optimasi tipe data juga berlaku pada fitur numerik dengan mengubah *float64* menjadi *float32* dan *int64* menjadi *int32*. Proses ini menghasilkan data yang kompatibel untuk pengembangan model.

3.2.4 *FaiReg Implementation*

FaiReg diimplementasikan dengan melakukan normalisasi label target berdasarkan statistik kelompok ras agar distribusi nilai target lebih seragam. Contoh perhitungan FaiReg dalam normalisasi label target ditunjukkan pada Gambar 3.6.



Gambar 3.6 Contoh Perhitungan FaiReg dalam Normalisasi Label

Normalisasi diawali dengan menghitung rata-rata (μ_{ci}) dan standar deviasi (σ_{ci}) untuk setiap kelompok ras. Rata-rata dan standar deviasi yang dihasilkan sebanyak jumlah kelompok. Selanjutnya, dilakukan perhitungan rata-rata (μ) dan standar deviasi keseluruhan (σ) pada target. Transformasi target menjadi *fair labels*

dihitung berdasarkan Persamaan 2.3. *Fair labels* menjadi target baru yang digunakan dalam pelatihan model dengan fungsi kerugian berdasarkan Persamaan 2.4.

3.2.5 *Modeling*

Pengembangan model dilakukan dengan model *baseline*: DT, metode *bagging*: RF, serta metode *boosting*: AdaBoost, GBM, XGBoost, LightGBM, dan CatBoost. Metode *bagging* dipilih karena kemampuannya dalam mengurangi *overfitting* dan meningkatkan stabilitas model. Metode *boosting* memiliki generalisasi yang baik dan berfokus pada kesalahan sebelumnya untuk meningkatkan akurasi.

DT merupakan metode populer yang intuitif dan mudah diinterpretasikan yang seringkali menjadi dasar bagi metode yang lebih kompleks seperti *bagging* dan *boosting* (Plaia dkk., 2022). RF mampu menangani data berdimensi tinggi dan fleksibel untuk mendeteksi pola yang kompleks. RF sepenuhnya nonparametrik, sehingga cenderung mencapai kinerja baik tanpa perlu penyetelan parameter (Rothacher & Strobl, 2024). AdaBoost memberikan bobot lebih tinggi pada kesalahan sebelumnya dan menggabungkan prediksi *weak learners* menjadi *strong learners*. GBM menggunakan pendekatan lebih fleksibel dengan meminimalkan korelasi terhadap gradien negatif dari fungsi *loss* keseluruhan. XGBoost mengoptimalkan efisiensi melalui regularisasi dibandingkan GBM standar. LightGBM menggunakan strategi GOSS dan FEB untuk mempercepat pelatihan. CatBoost menangani data kategorikal tanpa memerlukan banyak *preprocessing* (Sahin, 2022).

Pencarian *hyperparameter* terbaik pada setiap algoritma dilakukan dengan menggunakan GridSearchCV untuk mengeksplorasi kombinasi dari setiap *hyperparameter* dan mengevaluasi performanya. Pemilihan *hyperparameter* dan rentang nilai ditunjukkan pada Tabel 3.1.

Tabel 3. 1 *Hyperparameter* dan Rentang Nilai untuk *Grid Search*

Algoritma	***Hyperparameter***	***Parameters***
Decision Tree (Rian Oktafiani dkk., 2024)	‘max_depth’	3, 4, 5, 6, 7
Random Forest (Suryadi dkk., 2024)	‘max_depth’	None, 10, 20, 40
	‘n_estimators’	50, 100, 200, 400
AdaBoost (Mubaarok dkk., 2024)	‘max_depth’	5, 10, 15
	‘n_estimator’	200, 300
	‘learning_rate’	0.01, 0.1, 0.2
GBM (Z. Fan dkk., 2023)	‘max_depth’	3, 5
	‘n_estimator’	1000, 2000
	‘learning_rate’	0.02, 0.03
	‘subsample’	0.8
XGBoost (Z. Fan dkk., 2023)	‘max_depth	3, 5
	‘n_estimator’	1000, 2000
	‘learning_rate’	0.02, 0.03
	‘subsample	0.8
	‘colsample_bytree	0.5
LightGBM (Z. Fan dkk., 2023)	‘max_depth	3, 5
	‘n_estimator’	1000, 2000
	‘learning_rate’	0.02, 0.03
	‘subsample’	0.8
	‘colsample_bytree’	0.5
CatBoost (Hartono dkk., 2024)	‘iterations’	1000, 2000, 3000
	‘learning_rate’	0.1, 0.05, 0.01

Penelitian ini menguji dua skema berdasarkan konfigurasi yang berbeda untuk mengevaluasi kinerja model. Rincian dari masing-masing skema pengujian ditunjukkan pada Tabel 3.2.

Tabel 3. 2 Skema Pengujian

Skema	**Model**	**Algoritma**	***Cross Validation***	**Evaluasi Performa**	**Evaluasi *Fairness***
Tanpa FaiReg	1	*Baseline* (DT)	10-Fold	C-Index & AUC(t)	SP (PCC & MI)
	2	RF			
	3	AdaBoost			
	4	GBM			
	5	XGBoost			
	6	LightGBM			
	7	CatBoost			
Dengan FaiReg	1	*Baseline* (DT)	10-Fold	C-Index & AUC(t)	SP (PCC & MI)
	2	RF			
	3	AdaBoost			
	4	GBM			
	5	XGBoost			
	6	LightGBM			
	7	CatBoost			

Skema 1 menggunakan model dasar tanpa implementasi FaiReg. Model 1 dalam skema ini merupakan *baseline* yang menggunakan algoritma *Decision Tree*. Model 2 menggunakan Random Forest yang mewakili metode *bagging*, sedangkan Model 3 hingga Model 7 menerapkan algoritma berbasis *boosting*, seperti AdaBoost, Gradient Boosting, XGBoost, LightGBM, dan CatBoost. Skema 2 mengimplementasikan metode FaiReg pada ketujuh model yang sama untuk mengevaluasi pengaruh FaiReg terhadap performa dan keadilan model.

Evaluasi performa dilakukan menggunakan dua metrik utama, yaitu Concordance Index (C-Index) dan AUC(t) sebagai bentuk time-dependent ROC. Sementara itu, evaluasi *fairness* dilakukan menggunakan metrik *Statistical Parity (SP)* yang diukur melalui *Pearson Correlation Coefficient (PCC)* dan *Mutual Information (MI)*. Berdasarkan kedua skema, terdapat 14 model yang akan melalui pelatihan dan pengujian yang dijalankan menggunakan CPU. Validasi selama pelatihan menggunakan *K-Fold Cross-Validation* yang bertujuan untuk memastikan model dapat bekerja dengan baik pada data baru yang tidak pernah dilihat sebelumnya. Pemilihan *k* didasarkan pada rekomendasi penelitian (Yates dkk., 2023) yang menyatakan bahwa $k = 10$ sering digunakan karena memberikan keseimbangan yang baik antara bias dan varians.

3.2.6 *Evaluation*

Evaluasi kinerja model mencakup penilaian model keseluruhan serta penilaian implementasi teknik *fairness regression*. Performa model dievaluasi dengan C-Index dan AUC(t) terhadap masing-masing kelompok dan terhadap prediksi keseluruhan. Penilaian *fairness* melibatkan nilai PCC dan MI pada setiap model.

3.2.7 Penarikan Kesimpulan

Hasil dari evaluasi kinerja model akan digunakan untuk penarikan kesimpulan. Penentuan keandalan model dalam prediksi serta keadilan pada variabel terlindungi merupakan langkah penting pada tahap ini. Pengambilan kesimpulan berupa informasi mengenai kinerja model serta saran perbaikan dalam penelitian selanjutnya.