# BAB III
# METODE PENELITIAN

## 3.1 Objek Penelitian

Penelitian ini berfokus pada perancangan dan pengembangan sebuah sistem rekomendasi aktivitas fisik yang menggabungkan metode *machine learning* dan *rule-based reasoning*. Sistem ini bertujuan untuk memprediksi kategori intensitas aktivitas fisik pengguna, seperti ringan (*Light*), sedang (*moderate*), dan berat (*vigorous*), berdasarkan data pola hidup seperti usia, berat badan, tinggi badan, detak jantung rata-rata, durasi latihan, dan frekuensi olahraga. Hasil prediksi intensitas ini kemudian digabungkan dengan aturan fisiologis dan logika pakar untuk menentukan jenis latihan yang aman dan sesuai bagi pengguna, seperti *Cardio*, *Strength training*, *HIIT*, atau yoga, dengan mempertimbangkan kondisi khusus seperti *BMI* tinggi atau kebiasaan sehari-hari. Sistem yang dikembangkan ini dirancang agar rekomendasi yang diberikan bersifat aman, terpersonalisasi, dan berlandaskan bukti ilmiah, serta dapat diakses dengan mudah melalui aplikasi web yang dibangun menggunakan *framework Streamlit*.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Age | Gender | Weight (kg) | Height (m) | Max_BPM | Avg_BPM | Resting_BPM | Session_Duration (hou | Calories_Burn | Workout_Type | Fat_Percent | Water_Intake | Workout_Freq | Experience_Level | BMI |
| 2 | 56 | Male | 88.3 | 1.71 | 180 | 157 | 60 | 1.69 | 1313 | Yoga | 12.6 | 3.5 | 4 | 3 | 30.2 |
| 3 | 46 | Female | 74.9 | 1.53 | 179 | 151 | 66 | 1.3 | 883 | HIIT | 33.9 | 2.1 | 4 | 2 | 32 |
| 4 | 32 | Female | 68.1 | 1.66 | 167 | 122 | 54 | 1.11 | 677 | Cardio | 33.4 | 2.3 | 4 | 2 | 24.71 |
| 5 | 25 | Male | 53.2 | 1.7 | 190 | 164 | 56 | 0.59 | 532 | Strength | 28.8 | 2.1 | 3 | 1 | 18.41 |
| 6 | 38 | Male | 46.1 | 1.79 | 188 | 158 | 68 | 0.64 | 556 | Strength | 29.2 | 2.8 | 3 | 1 | 14.39 |
| 7 | 56 | Female | 58 | 1.68 | 168 | 156 | 74 | 1.59 | 1116 | HIIT | 15.5 | 2.7 | 5 | 3 | 20.55 |
| 8 | 36 | Male | 70.3 | 1.72 | 174 | 169 | 73 | 1.49 | 1385 | Cardio | 21.3 | 2.3 | 3 | 2 | 23.76 |
| 9 | 40 | Female | 69.7 | 1.51 | 189 | 141 | 64 | 1.27 | 895 | Cardio | 30.6 | 1.9 | 3 | 2 | 30.57 |
| 10 | 28 | Male | 121.7 | 1.94 | 185 | 127 | 52 | 1.03 | 719 | Strength | 28.9 | 2.6 | 4 | 2 | 32.34 |
| 11 | 28 | Male | 101.8 | 1.84 | 169 | 136 | 64 | 1.08 | 808 | Cardio | 29.7 | 2.7 | 3 | 1 | 30.07 |
| 12 | 41 | Male | 120.8 | 1.67 | 188 | 146 | 54 | 0.82 | 593 | HIIT | 20.5 | 3 | 2 | 1 | 43.31 |
| 13 | 53 | Male | 51.7 | 1.7 | 175 | 152 | 72 | 1.15 | 865 | HIIT | 23.6 | 3.5 | 3 | 2 | 17.89 |
| 14 | 57 | Male | 112.5 | 1.61 | 195 | 165 | 61 | 1.24 | 1013 | Cardio | 22.1 | 2.7 | 3 | 2 | 43.4 |
| 15 | 41 | Male | 94.5 | 2 | 179 | 136 | 69 | 1.18 | 794 | HIIT | 27.6 | 3.7 | 3 | 1 | 23.62 |
| 16 | 20 | Male | 117.7 | 1.81 | 196 | 161 | 54 | 1.35 | 1195 | Yoga | 26.4 | 3.3 | 3 | 2 | 35.93 |
| 17 | 39 | Female | 42.5 | 1.75 | 181 | 131 | 52 | 1.13 | 740 | Strength | 26.2 | 2.1 | 2 | 1 | 13.88 |
| 18 | 19 | Female | 64 | 1.53 | 166 | 167 | 58 | 1.33 | 1111 | HIIT | 29.8 | 2.3 | 3 | 2 | 27.34 |
| 19 | 41 | Female | 43.8 | 1.77 | 182 | 165 | 58 | 1.19 | 884 | Cardio | 31.9 | 1.6 | 3 | 1 | 13.98 |
| 20 | 47 | Female | 66.8 | 1.75 | 199 | 146 | 56 | 1.13 | 742 | Strength | 32.8 | 2.5 | 3 | 2 | 21.81 |

Gambar 3.1 Contoh *Dataset Gym Members Exercise Tracking*

Gambar 3.1 menunjukkan contoh *dataset Gym Members Exercise Tracking* yang digunakan dalam penelitian ini. *Dataset* tersebut memuat data usia, jenis kelamin, berat badan, tinggi badan, detak jantung, durasi latihan, jenis aktivitas fisik, frekuensi latihan, tingkat pengalaman, serta nilai *BMI*. *Dataset* ini menjadi objek penelitian utama yang digunakan sebagai dasar proses *preprocessing*, pelabelan intensitas aktivitas fisik, dan pelatihan model *machine learning*.

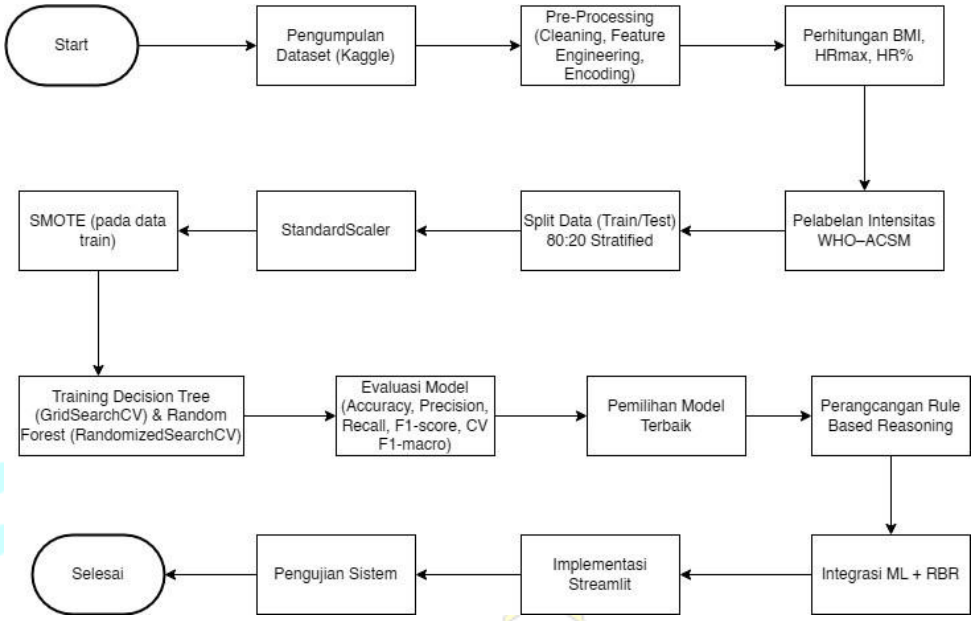
### 3.2 Prosedur Penelitian

Penelitian ini dilakukan melalui beberapa tahap yang terstruktur, dimulai dari pengumpulan *dataset* latihan fisik dari *platform Kaggle*, kemudian dilanjutkan dengan tahap pra-pemrosesan data yang mencakup pembersihan data, *feature engineering* berupa penghitungan *BMI*, serta *encoding* variabel kategorikal. Selanjutnya dilakukan pelabelan intensitas aktivitas fisik ke dalam tiga kategori yaitu *Light*, *Moderate*, dan *Vigorous* berdasarkan pedoman fisiologis dari WHO dan ACSM menggunakan nilai *HR_pct* dan *weekly_min* yang dihitung pada tahap sebelumnya.

Setelah data berlabel terbentuk, *dataset* dibagi menggunakan metode *stratified split* dengan rasio 80:20 untuk menjaga proporsi kelas pada data latih dan data uji. Selanjutnya, fitur numerik dinormalisasi menggunakan *StandardScaler*, kemudian diterapkan *SMOTE* pada data latih untuk mengatasi ketidakseimbangan kelas minoritas. Kedua proses ini dilakukan setelah pembagian data guna mencegah terjadinya data *leakage*.

Model kemudian dilatih menggunakan algoritma *Decision Tree* dan *Random Forest* dengan optimasi parameter masing-masing melalui *GridSearchCV* dan *RandomizedSearchCV*, serta dievaluasi menggunakan 5-*fold cross-validation*. Evaluasi dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk mengukur performa model secara menyeluruh.

Model terbaik yang diperoleh digunakan sebagai dasar untuk membuktikan bahwa pendekatan *machine learning* dengan tahapan pemrosesan data yang tepat mampu mengklasifikasikan intensitas aktivitas fisik secara akurat. Meskipun demikian, kedua algoritma yang digunakan dalam penelitian ini tetap diimplementasikan ke dalam sistem agar hasil prediksi keduanya dapat dibandingkan secara langsung. Hasil prediksi tersebut diintegrasikan dengan sistem berbasis aturan (*Rule-Based Reasoning*) untuk membentuk sistem rekomendasi aktivitas fisik yang tidak hanya mengandalkan kemampuan prediktif model, tetapi juga mempertimbangkan aturan keselamatan dan kondisi pengguna. Alur keseluruhan proses penelitian ini dapat dilihat pada Gambar 3.2.




Gambar 3.2 Prosedur Penelitian

Gambar 3.2 menjelaskan urutan dalam penelitian, mulai dari tahap pengumpulan data hingga pengujian sistem. Setiap kotak dalam *flowchart* merepresentasikan satu tahapan proses yang harus dilalui secara berurutan, memastikan bahwa model yang dihasilkan telah melalui validasi yang ketat mulai dari penanganan ketidakseimbangan kelas, normalisasi fitur, hingga *tuning hyperparameter* sehingga siap digunakan dalam sistem rekomendasi aktivitas fisik.

### 3.3 Sumber Data dan *Dataset*

*Dataset* yang digunakan dalam penelitian ini berasal dari *platform Kaggle*, yaitu *Gym Members Exercise Tracking Dataset* (2024), yang tersedia dalam format *Comma-Separated Values* (CSV). Dataset ini terdiri dari 973 data dengan 15 atribut yang mencakup profil fisik, parameter fisiologis, serta aktivitas latihan pengguna.

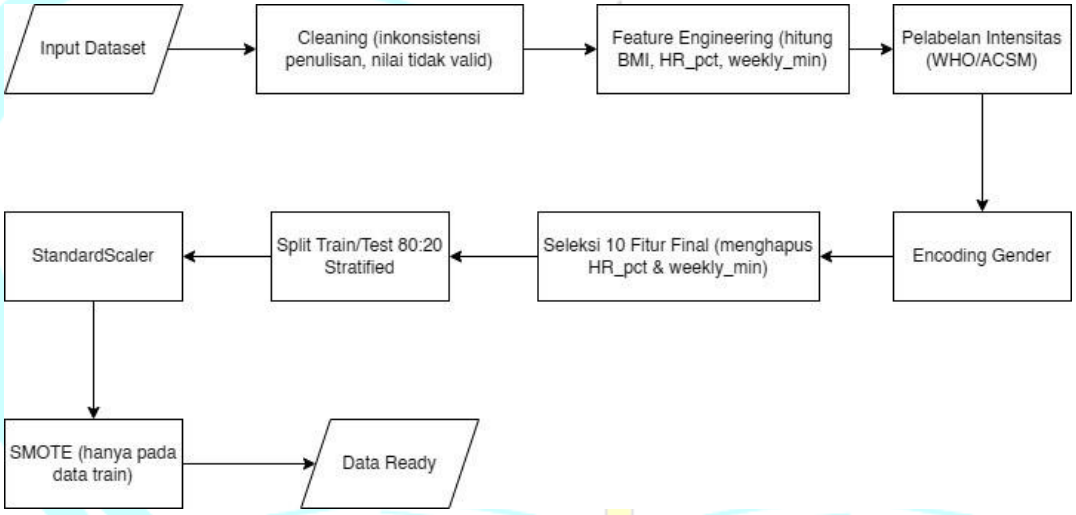
Berdasarkan pemeriksaan awal, dataset telah berada dalam kondisi bersih dan terstruktur dengan baik, sehingga tidak ditemukan nilai kosong (*missing value*) maupun inkonsistensi data yang signifikan. Oleh karena itu, tahap preprocessing tidak mencakup proses pembersihan data (*data cleaning*) yang kompleks, melainkan difokuskan pada transformasi data, *feature engineering*, dan persiapan data untuk pemodelan *machine learning*.

Pada tahap seleksi fitur, dipilih 10 atribut utama yang berhubungan langsung dengan intensitas latihan, yaitu usia, jenis kelamin, berat badan, tinggi badan, durasi latihan, frekuensi latihan, detak jantung rata-rata, detak jantung maksimal, kalori

terbakar, dan BMI. Sementara itu, variabel *HR_pct* dan *weekly_min* yang dihasilkan pada tahap *feature engineering* hanya digunakan untuk pelabelan intensitas dan tidak dimasukkan ke dalam model untuk menghindari *data leakage*. Selanjutnya, data diproses melalui tahap pra-pemrosesan sebelum digunakan dalam pembangunan model *machine learning*.

### 3.4 *Preprocessing* Data

Tahap pra-pemrosesan ini dilakukan untuk memastikan data yang akan digunakan berkualitas baik sebelum masuk ke proses pemodelan. Proses *preprocessing* data dapat dilihat pada gambar 3.3.




Gambar 3.3 Proses *Prepocessing Data*

Gambar 3.3 merupakan alur tahapan proses *preprocessing data*. Proses dimulai dari pembersihan data terhadap inkonsistensi penulisan dan nilai tidak valid. Selanjutnya dilakukan *feature engineering* untuk menghitung *BMI* dari berat dan tinggi badan, serta menghitung *HR_pct* dan *weekly_min* yang digunakan sebagai dasar penentuan label intensitas berdasarkan pedoman ACSM dan WHO. Setelah label intensitas terbentuk, dilakukan *encoding* pada variabel kategorikal jenis kelamin, kemudian dipilih 10 fitur final yang bebas *data leakage* dengan mengecualikan *HR_pct* dan *weekly_min* dari fitur model. Data kemudian dibagi secara *stratified* dengan rasio 80:20. Setelah pembagian, diterapkan *StandardScaler* pada seluruh fitur numerik untuk menyamakan skala antar variabel. Khusus pada data latih, diterapkan *SMOTE* untuk menyeimbangkan distribusi kelas minoritas

*Light* dengan menghasilkan sampel sintetis, sehingga model tidak bias terhadap kelas mayoritas.

Untuk memastikan parameter fisiologis dihitung secara tepat, penelitian ini menggunakan rumus:

$$HR_{max} = 220 - \text{usia}$$

$$HR\% = \frac{HR_{aktual}}{HR_{max}} \times 100\% \qquad [3.1]$$

$$BMI = \frac{\text{berat (kg)}}{\text{tinggi (m)}^2}$$

Ketiga rumus ini menjadi fondasi dalam proses *Labelling* intensitas yang digunakan sebagai target model *machine learning*.

### 3.5 Variabel Penelitian

Variabel penelitian dipisahkan menjadi variabel *Input*, variabel turunan, serta variabel target. Seluruh variabel yang digunakan tercantum pada Tabel 3.1.

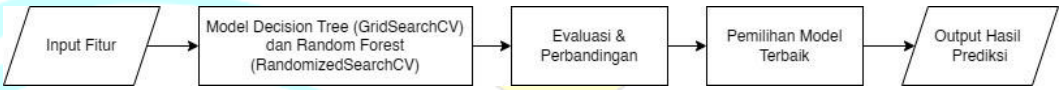
Tabel 3.1 Variabel Penelitian

| Variabel | Jenis | Deskripsi |
|---|---|---|
| *Age* | *Input* | Usia pengguna |
| *Gender* | *Input* | Jenis kelamin (*encoded*: 0/1) |
| *Weight (kg)* | *Input* | Berat badan pengguna |
| *Height (m)* | *Input* | Tinggi badan pengguna |
| *Session_Duration* | *Input* | Durasi latihan per sesi (jam) |
| *Workout_Frequency* | *Input* | Frekuensi latihan per minggu |
| *Avg_BPM* | *Input* | Detak jantung rata-rata saat latihan |
| *Max_BPM* | *Input* | Detak jantung maksimal saat latihan |
| *Calories_Burned* | *Input* | Kalori terbakar per sesi latihan |
| *BMI* | Turunan | Indeks massa tubuh, dihitung dari berat dan tinggi badan |
| *HR_pct* | Turunan (Pelabelan Target) | Persentase HR terhadap *HRmax*, digunakan sebagai dasar pelabelan intensitas, tidak dimasukkan sebagai fitur model |

| Variabel | Jenis | Deskripsi |
|---|---|---|
| *weekly_min* | Turunan (Pelabelan Target) | Total menit latihan per minggu, digunakan sebagai dasar pelabelan intensitas, tidak dimasukkan sebagai fitur model |

### 3.6 Perancangan dan Pelatihan Model *Machine learning*

Tahapan pelatihan model ditunjukkan dalam Gambar 3.4.

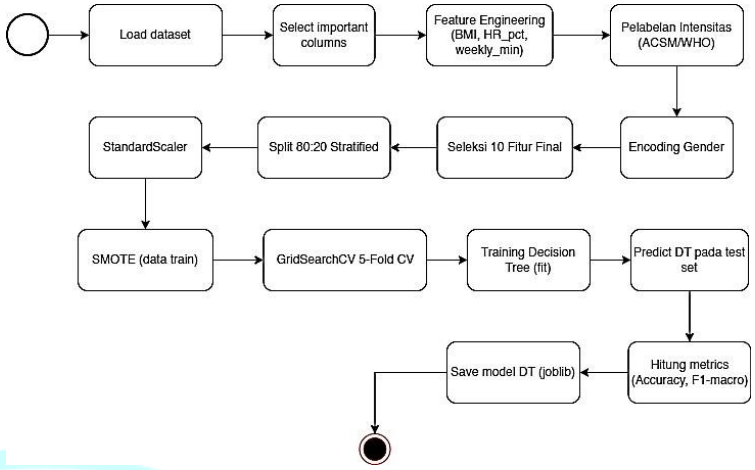



Gambar 3.4 Arsitektur Model *Machine learning*

Arsitektur perancangan model *machine learning* dalam penelitian ini dapat dilihat pada Gambar 3.4. Data masukan berupa 10 fitur yang telah melalui proses *preprocessing* digunakan secara bersamaan untuk melatih dua model secara paralel dan independen, yaitu *Decision Tree* dengan optimasi *GridSearchCV* dan *Random Forest* dengan *RandomizedSearchCV*. Kedua model dievaluasi menggunakan metrik yang sama, yaitu akurasi dan *CV F1-macro*, sehingga perbandingan yang dihasilkan bersifat adil. Hasil evaluasi kemudian dibandingkan untuk menentukan model terbaik yang akan diintegrasikan ke dalam sistem rekomendasi aktivitas fisik.

#### 3.6.1 *Decision Tree*

Algoritma *Decision Tree* bekerja dengan membangun sebuah struktur pohon yang berdasarkan pada fitur-fitur yang memberikan pemisahan terbaik pada data. Keunggulan utama dari model ini adalah kemampuannya untuk dijelaskan dan dipahami dengan mudah, sehingga sering digunakan sebagai model dasar atau *Baseline* dalam penelitian. Arsitektur yang digunakan pada *Decision Tree* dalam penelitian ini dapat dilihat pada Gambar 3.5.

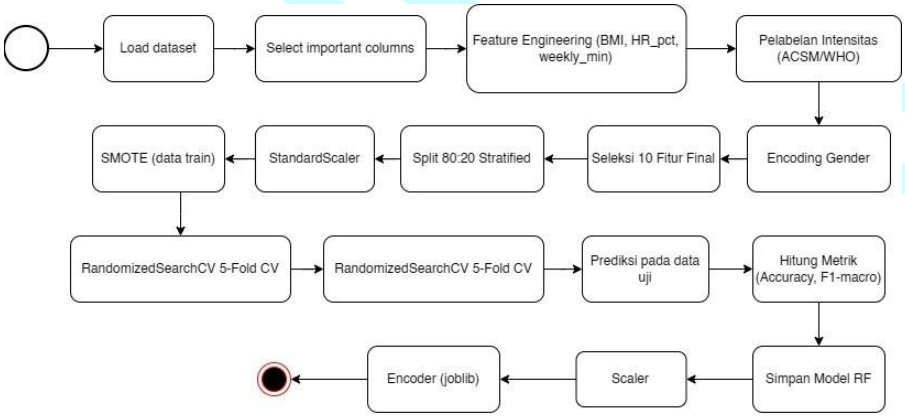



Gambar 3.5 Arsitektur *Decision Tree*

Gambar 3.5 menunjukkan alur proses pelatihan model *Decision Tree* secara menyeluruh, dimulai dari tahap pemuatan data, preprocessing, hingga penyimpanan model. Pada tahap optimasi, metode *GridSearchCV* digunakan pada data latih yang telah melalui proses *SMOTE* untuk memperoleh kombinasi parameter terbaik. Selanjutnya, model dievaluasi menggunakan *cross-validation* pada data original (tanpa *SMOTE*) agar hasil evaluasi dapat menggambarkan kemampuan generalisasi model secara lebih objektif. Setelah proses pelatihan dan evaluasi selesai, model terbaik disimpan untuk digunakan dalam sistem rekomendasi aktivitas fisik.

### 3.6.2 Model *Random Forest*

*Random Forest* dipilih sebagai kandidat model utama karena stabilitas prediksi dan efektivitasnya terhadap fitur fisiologis. Model ini dapat mengatasi hubungan *non-linear* dan interaksi fitur kompleks. Arsitektur yang digunakan pada *Random Forest* dalam penelitian ini dapat dilihat pada Gambar 3.6.




Gambar 3.6 Arsitektur *Random Forest*

Gambar 3.6 menunjukkan alur proses pelatihan model *Random Forest* yang dimulai dari tahap pemuatan *dataset*, seleksi kolom penting, *feature engineering*, hingga proses pelabelan intensitas berdasarkan pedoman ACSM/WHO. Setelah itu, data melalui proses *encoding*, seleksi fitur akhir, pembagian data secara *stratified* dengan rasio 80:20, normalisasi menggunakan *StandardScaler*, serta penyeimbangan data latih menggunakan *SMOTE*. Pada tahap optimasi model, digunakan metode *RandomizedSearchCV* dengan *5-fold cross-validation* untuk mencari kombinasi parameter terbaik secara lebih efisien. Setelah model terbaik diperoleh, dilakukan prediksi pada data uji dan evaluasi performa menggunakan metrik seperti *accuracy* dan *F1-macro*. Model yang telah dilatih kemudian disimpan bersama *scaler* dan *encoder* dalam format joblib agar dapat digunakan kembali pada sistem rekomendasi aktivitas fisik.

### 3.7 Evaluasi Model

Evaluasi model dilakukan untuk mengukur kemampuan generalisasi algoritma *Decision Tree* dan *Random Forest* terhadap data baru yang belum pernah digunakan selama proses pelatihan. Pengujian menggunakan data uji sebanyak 195 sampel (20% dari total dataset) yang telah dipisahkan sejak awal, sehingga hasil evaluasi dapat menggambarkan performa model secara objektif.

Pengukuran performa dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai ketepatan prediksi model pada setiap kelas intensitas. Selain itu, dilakukan juga *5-fold cross-validation* pada data original (tanpa *SMOTE*) untuk mengukur stabilitas dan konsistensi model. Hasil *cross-validation* dilaporkan dalam bentuk rata-rata *macro F1-score* untuk melihat keseimbangan performa model pada seluruh kelas.

Pemilihan model terbaik didasarkan pada nilai *macro F1-score* tertinggi, karena metrik ini lebih mampu menggambarkan performa model secara merata pada setiap kelas, termasuk kelas minoritas seperti *Light*, yang penting dalam sistem rekomendasi aktivitas fisik agar pengguna tidak menerima rekomendasi intensitas yang terlalu tinggi.

### 3.8 Perancangan *Rule-based reasoning*

Perancangan *Rule-based reasoning* dapat dilihat pada tabel 3.2.

Tabel 3.2 Aturan *Rule-based reasoning* berdasarkan ACSM (2025)

| Kondisi | Aturan | Rekomendasi |
|---|---|---|
| Usia ≥ 60 tahun | Hindari latihan intensitas tinggi | *Cardio* / *Yoga* / *Strength* ringan |
| Memiliki riwayat cedera | Hindari latihan berisiko tinggi sesuai area cedera | *Yoga* / *Strength* ringan |
| Prediksi intensitas *Light* | Prioritaskan latihan adaptasi | *Yoga / Cardio* |
| *BMI* normal (18,5–27), tanpa cedera, frekuensi ≥ 4x/minggu | Diperbolehkan latihan intensitas tinggi | *HIIT* |
| Frekuensi ≥ 3x/minggu & aktivitas aktif | Diperbolehkan latihan progresif | *HIIT / Strength* |
| Frekuensi ≤ 1x/minggu atau aktivitas rendah | Kategori pemula atau pasif | Yoga |
| Kondisi umum tanpa batasan khusus | Aktivitas standar | *Cardio* |

Pada Tabel 3.2, *Rule-Based Reasoning* digunakan sebagai mekanisme penyaringan terhadap hasil prediksi *Machine Learning* untuk memastikan rekomendasi aktivitas fisik tetap aman dan sesuai kondisi pengguna. Aturan disusun berdasarkan parameter usia, *BMI*, frekuensi olahraga, tingkat aktivitas, serta riwayat cedera. Pengguna dengan kondisi khusus seperti lansia atau memiliki cedera akan diberikan rekomendasi aktivitas dengan risiko rendah, sedangkan pengguna dengan kondisi fisik yang lebih baik dan rutin berolahraga dapat memperoleh rekomendasi latihan dengan intensitas lebih tinggi seperti *HIIT*. Dengan pendekatan ini, sistem mampu menghasilkan rekomendasi yang lebih personal dan aman.