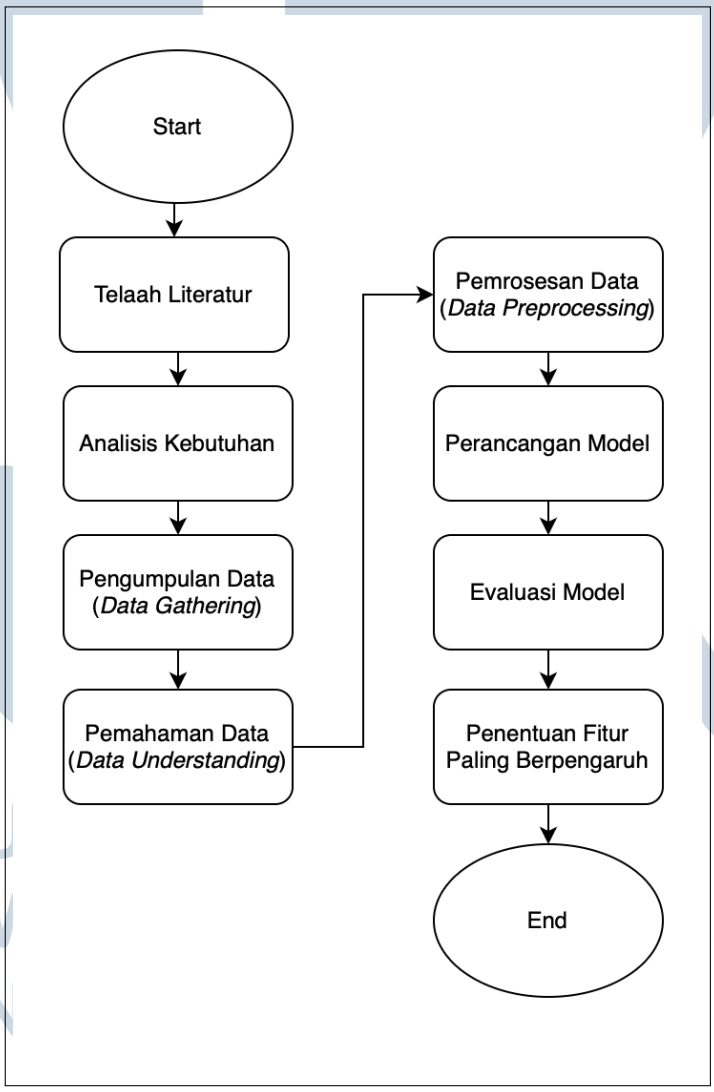
BAB 3
METODOLOGI PENELITIAN

3.1 Metodologi Penelitian

Metodologi penelitian merupakan langkah ilmiah yang digunakan untuk memperoleh data dan menjelaskan proses penelitian, di mana dalam studi ini diterapkan metode komputasional berbasis data sekunder dengan pendekatan eksperimental melalui pengujian dua algoritma *machine learning* untuk menyelesaikan kasus klasifikasi.



Gambar 3.1. *Flowchart* Metodologi Penelitian

Gambar 3.1 merupakan *flowchart* yang menunjukkan alur proses dan metodologi yang diterapkan dalam membangun model analisis kebiasaan tidur dan prediksi risiko insomnia menggunakan metode *Decission Tree* dan *Logistic Regression*.

3.1.1 Telaah Literatur

Tahap Telaah Literatur mencakup analisis mendalam terhadap teori-teori dan hasil penelitian sebelumnya yang relevan dengan topik, seperti metode dan penerapan algoritma *Decissiong Tree* dan *Logistic Regression*, evaluasi model yang menggunakan *confussion matrix*, serta faktor-faktor yang memengaruhi kebiasaan tidur. Referensi yang digunakan berasal dari penelitian dalam lima tahun terakhir.

3.2 Analisis Kebutuhan

Pada tahap analisis kebutuhan, dilakukan identifikasi terhadap elemen-elemen yang diperlukan untuk mendukung pelaksanaan penelitian. Berdasarkan hasil telaah literatur, analisis ini difokuskan pada pencapaian tujuan penelitian, khususnya terkait manfaat dan *output* yang diharapkan dari penelitian ini.

3.2.1 Pengumpulan Data *(Data Gathering)*

Tahap pengumpulan data melibatkan proses pencarian dan pengumpulan data yang relevan dengan fokus penelitian, yakni untuk keperluan analisis serta prediksi kemungkinan insomnia. Data yang dibutuhkan harus sesuai dengan tujuan penelitian agar dapat memberikan hasil yang akurat dan mendukung pengujian model yang digunakan.

Dalam penelitian ini, data yang digunakan merupakan data sekunder yang diperoleh dari sumber tepercaya seperti *Kaggle*. *Dataset* tersebut berisi informasi mengenai kebiasaan tidur serta kualitas tidur dari 1000 individu pada tahun 2024, serta variabel-variabel pendukung yang memengaruhi risiko insomnia, seperti denyut jantung *(heart rate variability)*, temperatur badan *(body temperature)*, gerakan saat tidur *(movement during sleep)*, durasi tidur *(sleep duration hours)*, kualitas tidur *(sleep quality score)*, konsumsi kafein *(caffeine intake)*, tingkat stres *(stress level)*, perubahan jam tidur *(bedtime consistency)*, dan paparan cahaya *(light exposure hours)*.

3.2.2 Pemahaman Data *(Data Understanding)*

Dataset yang digunakan dalam penelitian ini berasal dari laman *Kaggle* dengan judul *Sleep and Health Metrics*, yang termasuk dalam kategori *Health*. *Dataset* tersebut berbentuk *file* dengan nama *"wearable tech sleep quality 1"* dan berformat .csv dan berukuran 152,85 KB. Contoh isi dari *dataset* ini ditampilkan pada Tabel 3.1.

Tabel 3.1. Data Kebiasaan Tidur serta Kualitas Tidur

HRV	BT	MDS	SDH	SQC	CI	SL	BC	LEH
79.93	37.19	1.324	4.638	1.0	107.6	2.771	0.657	7.933
67.23	36.96	1.855	6.209	1.0	104.6	3.738	0.144	6.992
82.95	36.52	1.207	6.879	10.0	0.0	3.115	0.642	7.655
...	...	...	...	...	...	...	...	...
86.27	36.79	0.760	8.210	3.291	80.76	5.087	0.394	4.765
45.38	36.85	0.532	7.098	2.186	84.09	6.911	0.770	8.692
74.54	36.35	2.164	8.770	1.983	92.58	4.031	0.0	6.739

Tabel tersebut menampilkan cuplikan data dari *dataset* yang digunakan, dengan penjelasan masing-masing kolom disajikan sebagai berikut.

- ***HRV (Heart Rate Variability)***: Variabilitas detak jantung yang disimulasikan berdasarkan perbedaan waktu antar detak jantung.
- ***BT (Body Temperature)***: Suhu tubuh yang dihasilkan secara artifisial dalam satuan derajat Celsius.
- ***MDS (Movement During Sleep)***: Data sintetis mengenai jumlah gerakan tubuh saat tidur.
- ***SDH (Sleep Duration Hours)***: Total durasi tidur dalam jam yang diperoleh melalui simulasi.
- ***SQC (Sleep Quality Score)***: Skor sintetis yang merepresentasikan kualitas tidur.

- ***CI (Caffeine Intake)***: Jumlah konsumsi kafein yang disimulasikan dalam satuan miligram.
- ***SL (Stress Level)***: Indeks tingkat stres berdasarkan data simulasi.
- ***BC (Bedtime Consistency)***: Konsistensi rutinitas waktu tidur yang disimulasikan, dalam skala 0–1; nilai lebih rendah menunjukkan ketidakkonsistenan yang lebih tinggi.
- ***LEH (Light Exposure Hours)***: Jumlah paparan cahaya sintetis selama siang hari, mencerminkan paparan cahaya alami secara umum.

Dataset ini dirancang untuk mensimulasikan berbagai skenario dan kondisi yang mungkin terjadi, sehingga dapat menjadi landasan yang solid bagi analisis prediktif dan penelitian data. Melalui penggunaan data sintetis, *dataset* ini mampu merepresentasikan beragam variasi dan interaksi dalam metrik terkait tidur dan kesehatan secara menyeluruh.

3.3 Praproses Data *(Data Preprocessing)*

Pada tahap ini dilakukan proses pra-pemrosesan data, yang meliputi pembersihan data, penanganan data yang tidak lengkap, augmentasi data, serta transformasi data agar siap digunakan dalam proses pemodelan. Tujuan dari tahapan ini adalah untuk memastikan data dalam kondisi bersih, terstruktur, dan layak pakai. Berikut merupakan penjabaran dari langkah-langkah pra-pemrosesan data yang telah dilakukan.

3.3.1 Eksplorasi Data *(Data Exploration)*

Tahap eksplorasi data merupakan proses analisis awal terhadap *dataset* untuk memahami karakteristik dan struktur data yang digunakan. Dalam tahap ini, dilakukan berbagai kegiatan seperti pengenalan data, normalisasi, serta visualisasi guna mengidentifikasi pola atau hubungan yang mungkin ada. Berikut adalah tahapan-tahapan yang dilakukan dalam proses eksplorasi data.

A. ***Data Understanding***

Proses *data understanding* merupakan langkah awal dalam tahapan eksplorasi data yang bertujuan untuk memahami struktur, isi, dan karakteristik dari *dataset* yang digunakan. *Dataset* yang dipakai dalam

penelitian ini diperoleh dari *platform Kaggle*, dengan judul *"Sleep and Health Metrics"*. *Dataset* tersebut terdiri dari 1000 baris data dengan 9 fitur numerik utama, yaitu: *Heart Rate Variability*, *Body Temperature*, *Movement During Sleep*, *Sleep Duration Hours*, *Sleep Quality Score*, *Caffeine Intake* (mg), *Stress Level*, *Bedtime Consistency*, dan *Light Exposure Hours*. Karakteristik *dataset* ini mengategorikannya sebagai *data cross-sectional*, karena setiap baris data merepresentasikan observasi independen dari metrik kebiasaan tidur dan aktivitas harian pada satu titik waktu tertentu per individu, tanpa adanya variabel waktu eksplisit (tanggal atau *timestamp*) yang menunjukkan urutan kronologis antar baris. Dengan demikian, data ini sangat sesuai untuk tujuan klasifikasi risiko insomnia, di mana setiap observasi dianggap sebagai *snapshot* kondisi seseorang untuk diprediksi.

Langkah awal dilakukan dengan menampilkan lima data teratas (*head*) untuk melihat isi data secara umum. Selanjutnya, dilakukan pengecekan informasi struktur *dataset* menggunakan fungsi *df.info()* yang menampilkan jumlah entri, jumlah data *non-null*, dan tipe data tiap kolom.

Data ini kemudian diproses menjadi variabel target biner bernama 'Insomnia', yang diperoleh dari kolom *Sleep Quality Score*, dengan ketentuan: jika skor kurang dari 5 maka diberi label 1 (berisiko insomnia), dan jika skor 5 atau lebih maka diberi label 0 (tidak berisiko insomnia). Hal ini mengubah permasalahan menjadi bentuk klasifikasi biner.

Deskripsi statistik kemudian dihitung untuk memahami distribusi nilai dari tiap fitur. Statistik deskriptif ini mencakup nilai minimum, maksimum, rata-rata, dan standar deviasi. Hasilnya membantu menentukan apakah ada fitur dengan skala yang terlalu berbeda atau nilai ekstrem *(outlier)*.

B. ***Data Augmentation***

Data augmentation merupakan proses penambahan data latih secara artifisial dengan tujuan untuk meningkatkan jumlah dan keberagaman data yang tersedia, terutama ketika jumlah data awal terbatas. Dalam penelitian ini, teknik augmentasi yang digunakan adalah model *Vector Autoregression* (VAR), yang efektif digunakan untuk data deret waktu (*time series*) multivariat.

Model VAR membentuk hubungan antar fitur berdasarkan lag waktu dan menghasilkan simulasi data baru yang mengikuti pola distribusi data asli.

Proses ini dilakukan dengan cara melatih model VAR pada data awal, lalu menyimulasikan sejumlah langkah ke depan untuk menghasilkan data sintetis. Hasil dari proses ini digabungkan kembali dengan data asli hingga mencapai jumlah total 5000 baris data.

Teknik augmentasi ini bertujuan untuk memperbaiki performa model klasifikasi dengan memberikan lebih banyak variasi pola, sekaligus mencegah *overfitting* akibat jumlah data yang terlalu kecil.

C. ***Data Normalization***

Proses *data normalization* bertujuan untuk menyetarakan skala dari seluruh fitur numerik agar tidak ada satu fitur yang mendominasi pembelajaran model karena skala yang lebih besar. Dalam penelitian ini digunakan metode *Standard Scaler*, sebuah teknik yang mentransformasikan data ke dalam distribusi standar (*mean* = 0, standar deviasi = 1).

D. ***Data Visualization***

Tahap *data visualization* merupakan proses mengubah informasi dalam bentuk *dataframe* menjadi representasi visual seperti grafik atau plot. Tahapan ini berperan penting dalam mempermudah pemahaman data melalui tampilan visual, sehingga pola, tren, dan hubungan antar data dapat dikenali lebih mudah dibandingkan hanya dengan analisis statistik. Visualisasi ini biasanya disajikan dalam bentuk grafik, diagram, atau peta, yang membantu dalam mengidentifikasi pola tersembunyi, tren data, maupun anomali secara lebih efektif.

3.4 Perancangan Kode dan Model

Tahap perancangan kode dan model mencakup beberapa langkah penting, seperti seleksi fitur, pengaturan parameter model, serta validasi model. Seleksi fitur bertujuan untuk memilih variabel-variabel yang paling relevan dalam proses pemodelan. Penyesuaian parameter dilakukan guna menemukan konfigurasi terbaik agar model bekerja secara optimal. Sementara itu, proses validasi diperlukan untuk memastikan bahwa model yang dibangun mampu menghasilkan prediksi yang akurat dan dapat dipercaya.

3.5 Evaluasi Model

Tahap evaluasi model merupakan langkah krusial dalam siklus hidup pengembangan model *machine learning* untuk memastikan bahwa model yang dibangun memiliki kinerja yang optimal dan dapat diandalkan dalam melakukan prediksi. Pada tahap ini, model yang telah dilatih akan diuji coba menggunakan data yang belum pernah dilihat sebelumnya *(data test)* untuk menilai kemampuannya dalam menggeneralisasi pola dari data pelatihan.

Evaluasi kinerja model klasifikasi dalam penelitian ini dilakukan dengan menggunakan beberapa metrik keakuratan yang umum dan relevan, seperti, *accuracy*, *precision*, *recall*, *F1-score,* dan *confusion matrix*.

Analisis metrik-metrik ini bertujuan untuk membantu dalam mengidentifikasi kelebihan serta kekurangan model dalam memprediksi data, serta memberikan rekomendasi untuk peningkatan lebih lanjut. Metrik-metrik ini memberikan pandangan yang komprehensif tentang performa model dari berbagai sudut pandang, sehingga dapat membantu untuk mengidentifikasi kekurangan dan perbaikan dari model yang diimplementasikan.

3.6 *Hyperparameter Tuning* dengan RandomizedSearchCV

Dalam penelitian ini, proses pencarian *hyperparameter* terbaik untuk masing-masing algoritma dilakukan menggunakan `RandomizedSearchCV` dari pustaka `scikit-learn`. Pendekatan ini dipilih karena mampu mengeksplorasi kombinasi parameter secara acak dalam jumlah iterasi terbatas, sehingga efisien secara komputasi dibandingkan `GridSearchCV`.

3.6.1 Logistic Regression

Untuk model *Logistic Regression*, dilakukan pencarian *hyperparameter* dengan menggunakan `Pipeline` yang terdiri dari dua langkah utama, yaitu normalisasi menggunakan `StandardScaler`, dan klasifikasi menggunakan `LogisticRegression` dengan solver `saga`. Parameter yang ditelusuri meliputi:

- `penalty`: l1, l2, elasticnet, none
- `C`: 0.01, 0.1, 1, 10, 100
- `l1_ratio`: 0, 0.25, 0.5, 0.75, 1 (hanya relevan jika penalty = elasticnet)

Validasi dilakukan menggunakan `TimeSeriesSplit` dengan 5 fold untuk menjaga kontinuitas data berbasis waktu. Skor evaluasi yang digunakan adalah `accuracy`. Proses pencarian dilakukan sebanyak 20 iterasi secara paralel.

A Proses Pembelajaran dan Prediksi Logistic Regression

Proses pembelajaran dan prediksi dari algoritma *Logistic Regression* dijelaskan melalui diagram alir pada Gambar 3.2. Tahapan utama dari proses ini adalah sebagai berikut:

1. Inisialisasi bobot w secara acak dan bias $b = 0$.
2. Mendefinisikan fungsi aktivasi sigmoid sebagai:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

3. Untuk setiap *epoch* dalam proses pelatihan:
   (a) Untuk setiap sampel pelatihan:
   - Hitung nilai z:

$$z = w \cdot X + b$$

   - Hitung nilai prediksi:

$$\hat{y} = \sigma(z)$$

   - Hitung galat (error):

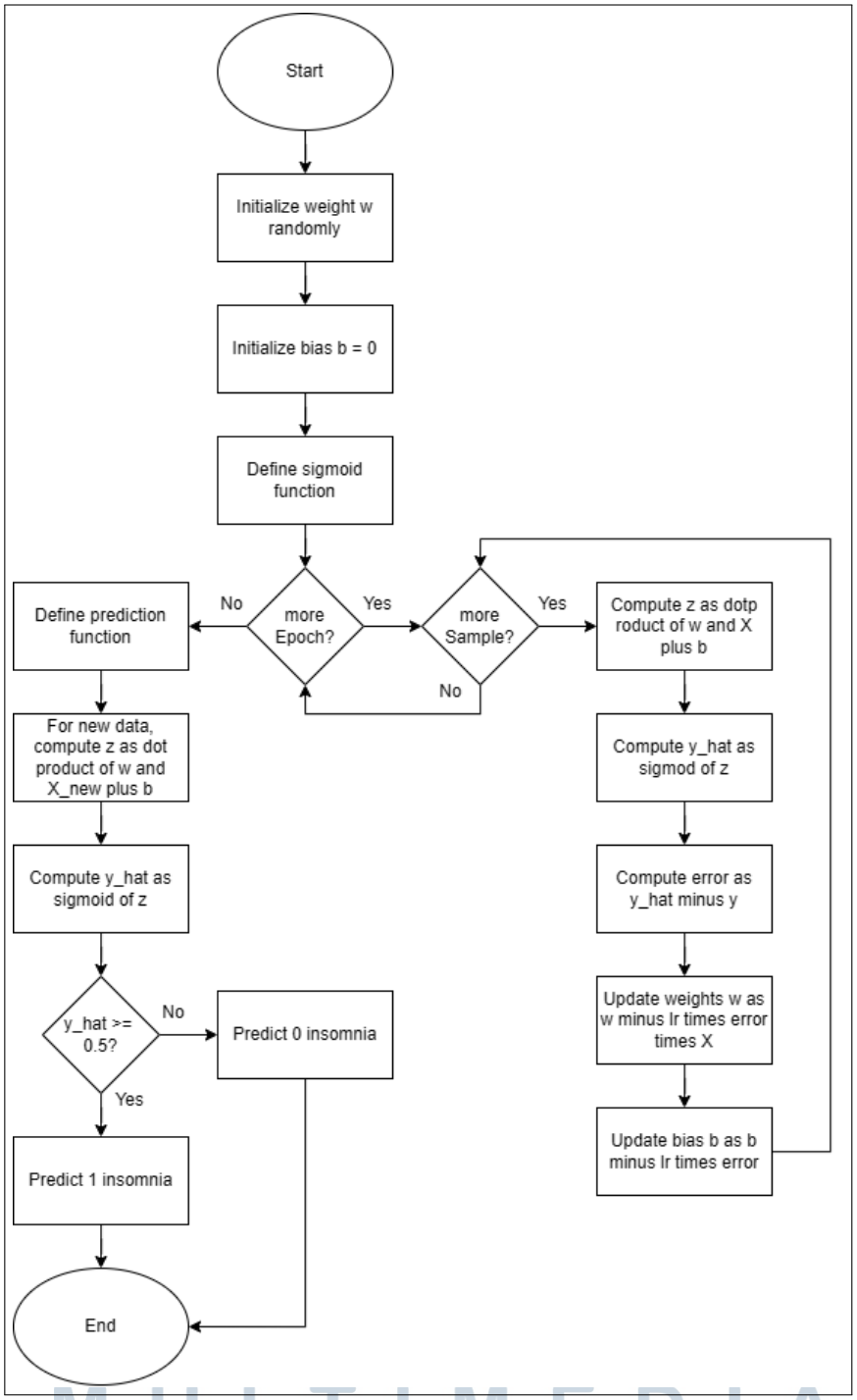
$$\text{error} = \hat{y} - y$$

   - Perbarui bobot dan bias:

$$w = w - \alpha \cdot \text{error} \cdot X$$
$$b = b - \alpha \cdot \text{error}$$

4. Setelah pelatihan selesai, digunakan fungsi prediksi untuk data baru:
   - Hitung $z = w \cdot X_{\text{new}} + b$
   - Hitung $\hat{y} = \sigma(z)$
   - Lakukan klasifikasi:

$$\text{Jika } \hat{y} \geq 0.5 \Rightarrow \text{Insomnia (1), selain itu } \Rightarrow \text{Tidak Insomnia (0)}$$



Gambar 3.2. Flowchart Proses Logistic Regression

3.6.2 Decision Tree

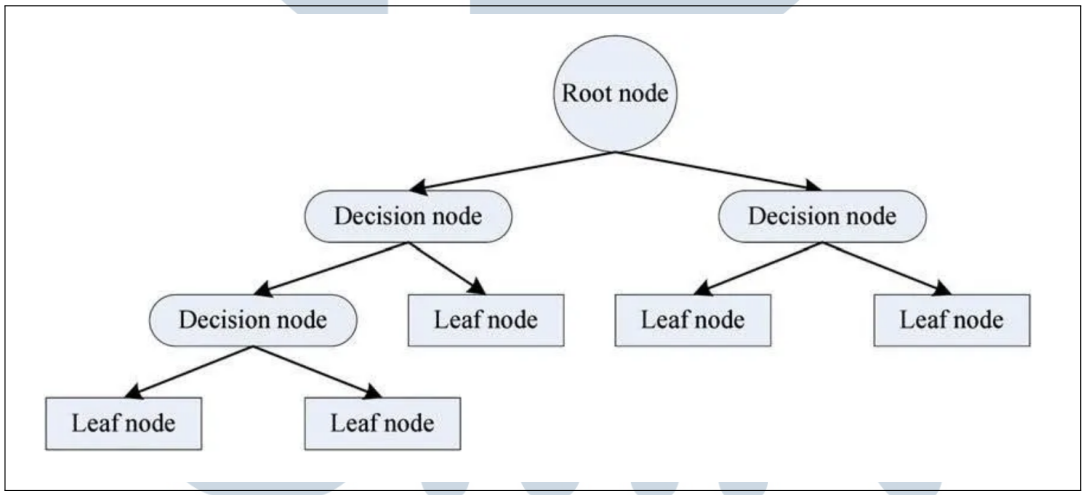
Pada model *Decision Tree*, pipeline juga digunakan, meskipun `StandardScaler` tidak berpengaruh terhadap model berbasis pohon keputusan.

Parameter yang ditelusuri mencakup:

- `criterion`: gini, entropy
- `max_depth`: 3, 7, 15, 25
- `min_samples_split`: 2, 5, 10, 20
- `min_samples_leaf`: 1, 2, 5, 10
- `max_features`: sqrt, log2, None

Seperti pada model sebelumnya, pencarian dilakukan sebanyak 1000 iterasi menggunakan `RandomizedSearchCV`, dengan skema validasi `TimeSeriesSplit` sebanyak 5 fold dan metrik evaluasi `F1 Score`.

A Proses Pembelajaran dan Prediksi Decision Tree



Gambar 3.3. *Decission Tree Classifier*
Sumber: Medium [52]

Gambar 3.3 memvisualisasikan operasi alur keputusan hierarkis untuk klasifikasi risiko insomnia. Logika model ini berpusat pada pertanyaan fundamental di *root node* mengenai asupan kafein: ”Apakah *Caffeine Intake* kurang dari atau sama dengan 53.7”. Pertanyaan ini berfungsi sebagai pemisah data utama, yang mengarahkan data ke dua jalur analisis yang berbeda. Untuk data dengan asupan kafein rendah, model melanjutkan analisis dengan memeriksa *Sleep Duration Hours*, di mana durasi tidur yang memadai menjadi indikator kuat

untuk kesimpulan ”*No Insomnia*”. Sebaliknya, pada jalur untuk asupan kafein tinggi, model menggunakan *Movement During Sleep* sebagai kriteria sekunder, dengan tingkat pergerakan yang tinggi secara signifikan meningkatkan probabilitas klasifikasi ”Insomnia”. Setelah pemisahan-pemisahan awal ini, *decision tree* terus menyempurnakan klasifikasinya dengan mempertimbangkan fitur-fitur lain seperti *Stress Level* pada level yang lebih dalam, hingga setiap alur data mencapai *leaf node* yang menentukan prediksi akhir.

3.6.3 Hasil Tuning

Setelah proses tuning selesai, diperoleh kombinasi parameter terbaik untuk masing-masing model yang kemudian digunakan untuk melakukan prediksi pada data validasi. Akurasi validasi dari masing-masing model dicatat untuk dibandingkan pada tahap evaluasi akhir.

- **Logistic Regression**: Akurasi Validasi = `\accuracy_score(y_val, y_pred_lr)`
- **Decision Tree**: Akurasi Validasi = `\accuracy_score(y_val, y_pred_dt)`

Informasi detail mengenai parameter terbaik ditampilkan pada bagian hasil dan pembahasan.

3.7 Penentuan Fitur Paling Berpengaruh

Setelah tahap pembangunan dan pelatihan model, langkah selanjutnya dalam metodologi adalah mengidentifikasi fitur-fitur pada *dataset* yang paling signifikan dalam memengaruhi prediksi model. Penentuan fitur paling berpengaruh *(feature importance)* yang digunakan untuk memahami karakteristik data yang dominan dalam menentukan kualitas tidur (risiko insomnia), serta untuk memberikan interpretasi terhadap bagaimana model membuat keputusan. Pemahaman ini juga dapat menjadi dasar untuk optimasi model di masa depan atau untuk memberikan wawasan lebih lanjut terkait faktor-faktor pemicu risiko insomnia.

Dalam penelitian ini, metode *Permutation Importance* digunakan untuk menilai kontribusi setiap fitur terhadap kinerja model. Metode ini bekerja dengan mengukur penurunan skor model (misalnya, akurasi) ketika nilai-nilai dari suatu fitur diacak secara acak (di-*permute*). Fitur yang menyebabkan penurunan

skor paling besar ketika diacak dianggap sebagai fitur yang paling penting atau berpengaruh. Keunggulan metode ini adalah kemampuannya untuk mengukur pentingnya fitur secara global untuk model apa pun, serta dapat menangani fitur yang saling berkorelasi.

Proses penentuan fitur paling berpengaruh dilakukan pada data pengujian *(test set)* untuk memastikan bahwa penilaian dilakukan pada data yang tidak terlibat dalam proses pelatihan model. Hasil dari analisis ini kemudian akan digunakan untuk membandingkan fitur-fitur penting antara model *Decision Tree* dan *Logistic Regression*, serta untuk mengidentifikasi pola umum dari fitur-fitur yang berkontribusi pada klasifikasi kondisi risiko 'Insomnia'.