KLASIFIKASI RISIKO INSOMNIA BERDASARKAN POLA TIDUR DAN GEJALA PSIKOLOGIS MENGGUNAKAN LOGISTIC REGRESSION DAN DECISSION TREE

Bilal
Department of Informatics Engineering
STMIK Tazkia
Bogor. Indonesia
bilalfarhani635@gmail.com

Thoriqurrahman Akrami
Department of Informatics Engineering
STMIK Tazkia
Bogor. Indonesia
thoriqurrahmana@gmail.com

Abstrak

Insomnia merupakan salah satu gangguan tidur yang paling umum dan memiliki dampak signifikan terhadap kualitas hidup, kesehatan mental, serta produktivitas individu. Penelitian ini bertujuan untuk mengembangkan model klasifikasi risiko insomnia dengan memanfaatkan pendekatan *Machine Learning* berbasis data pola tidur dan gejala psikologis. Dataset yang digunakan berasal dari "*Dataset of Insomniac and Normal People*" yang terdiri atas 30 sampel dan 11 fitur utama, di antaranya *Total_sleep_time (hour)*, *Satisfaction_of_sleep*, dan *Recent_psychological_attack*, dengan variabel target *Disorder* (0 = normal, 1 = insomnia). Tahapan pra-pemrosesan data meliputi penanganan *missing values* (tidak ditemukan), transformasi variabel kategorikal menggunakan *LabelEncoder*, serta standardisasi fitur numerik melalui *StandardScaler*. Dataset kemudian dibagi ke dalam subset pelatihan dan pengujian dengan rasio 70:30. Empat algoritma klasifikasi, yaitu *Logistic Regression*, *Decision Tree*, *Random Forest*, dan *Support Vector Machine* (*SVM*), digunakan dalam eksperimen dan dievaluasi melalui metrik akurasi, *classification report*, *confusion matrix*, serta *ROC Curve* dengan nilai *Area Under the Curve* (*AUC*). Hasil penelitian menunjukkan bahwa model *Logistic Regression* memperoleh akurasi sempurna sebesar 100% pada data uji, mengungguli model lain seperti *Decision Tree*, *Random Forest*, dan *SVM* yang masing-masing mencatatkan akurasi sebesar 89%. *Logistic Regression* juga menunjukkan kinerja klasifikasi yang konsisten dan presisi tinggi untuk kedua kelas. Meskipun demikian, keterbatasan jumlah sampel menimbulkan potensi *overfitting* dan mengurangi generalisasi model terhadap populasi lebih luas. Studi ini menegaskan potensi implementasi *Machine Learning* sebagai sistem pendukung keputusan dalam deteksi dini gangguan tidur seperti insomnia, namun validasi lanjutan dengan dataset yang lebih besar dan beragam tetap diperlukan.

Kata kunci : *Insomnia, klasifikasi, machine learning, logistic regression, decision tree*

INSOMNIA RISK CLASSIFICATION BASED ON SLEEP PATTERNS AND PSYCHOLOGICAL SYMPTOMS USING LOGISTIC REGRESSION AND DECISION TREE

Abstract

Insomnia is one of the most common sleep disorders and has a significant impact on quality of life, mental health, and individual productivity. This study aims to develop an insomnia risk classification model using a *Machine Learning* approach based on sleep pattern and psychological symptom data. The dataset used comes from the "*Dataset of Insomniac and Normal People*" which consists of 30 samples and 11 main features, including *Total_sleep_time (hour)*, *Satisfaction_of_sleep*, and *Recent_psychological_attack*, with the target variable *Disorder* (0 = normal, 1 = insomnia). Data pre-processing stages include handling *missing values* (not found), transformation of categorical variables using *LabelEncoder*, and

Naskah dikirim: dd-mm-yyyy; direvisi dd-mm-yyyy; diterima: dd-mm-yyyy

standardization of numeric features using *StandardScaler*. The dataset is then divided into training and testing subsets with a ratio of 70:30. Four classification algorithms, namely *Logistic Regression*, *Decision Tree*, *Random Forest*, and *Support Vector Machine* (*SVM*), are used in the experiment and evaluated through *accuracy metrics*, *classification report*, *confusion matrix*, and *ROC Curve* with *Area Under the Curve* (*AUC*) values. The results showed that the *Logistic Regression* model achieved perfect accuracy of 100% on the test data, outperforming other models such as *Decision Tree*, *Random Forest*, and *SVM*, which each recorded an accuracy of 89%. *Logistic Regression* also demonstrated consistent classification performance and high precision for both classes. However, the limited sample size creates the potential for *overfitting* and reduces the model's generalizability to the broader population. This study confirms the potential of implementing *Machine Learning* as a decision support system in the early detection of sleep disorders such as insomnia, but further validation with larger and more diverse datasets is still needed.

Keywords: *Insomnia, classification, machine learning, logistic regression, decission tree*

1. Pendahuluan

Probabilistik dan statistik merupakan fondasi utama dalam pembelajaran mesin (*machine learning*), terutama pada pendekatan *supervised learning*. Probabilistik digunakan untuk memodelkan ketidakpastian dalam data dan hasil prediksi, sedangkan statistik berperan dalam menganalisis distribusi data, mengukur hubungan antar variabel, dan mengestimasi parameter model. Pemahaman yang baik terhadap kedua bidang ini memungkinkan perancangan model prediktif yang lebih akurat dan dapat diandalkan.

Insomnia adalah gangguan tidur yang ditandai dengan kesulitan dalam memulai, mempertahankan, atau mendapatkan kualitas tidur yang optimal (Alomedika, 2023). Gangguan ini tidak hanya berdampak pada kondisi psikologis dan kognitif seseorang, tetapi juga meningkatkan risiko penyakit kardiovaskular, gangguan metabolik, serta menurunkan produktivitas kerja (Sri Susanty Budiman, Vol. 24, Article 2385, 2024). Laporan *World Health Organization* (WHO) menunjukkan bahwa prevalensi insomnia secara global terus meningkat, terutama pada populasi usia produktif dan lansia, menjadikannya masalah kesehatan masyarakat yang signifikan.

Secara konvensional, diagnosis insomnia dilakukan melalui wawancara klinis, observasi langsung, atau kuesioner subjektif seperti *Pittsburgh Sleep Quality Index* (PSQI). Namun, metode ini memiliki keterbatasan seperti bias persepsi pasien dan waktu analisis yang relatif panjang. Oleh karena itu, dibutuhkan pendekatan komputasional yang objektif, efisien, dan adaptif untuk mendukung proses diagnosis.

Penelitian ini memanfaatkan data dari *Dataset of Insomniac and Normal People* (Saeed Gharehbaghi, 2021), yang terdiri atas fitur-fitur seperti total waktu tidur, tingkat kepuasan tidur, frekuensi tidur larut malam, frekuensi terbangun, kebiasaan tidur siang, tingkat kantuk di siang hari, lama masalah tidur berlangsung, riwayat serangan psikologis, ketakutan memulai tidur, dan status gangguan tidur (0 = normal, 1 = insomnia).

Untuk memodelkan hubungan antara fitur-fitur prediktor (input) dengan status insomnia, digunakan **Linear Regression** sebagai model linear dasar. Secara matematis, persamaan model dapat dinyatakan sebagai:

$$F(x)=x1w1+x2w2+\cdots+xnwn+x0w0$$

di mana xix_ixi adalah nilai fitur ke-iii, wiw_iwi adalah bobot atau koefisien yang merepresentasikan kontribusi fitur tersebut, dan x0w0x_0 w_0x0w0 merupakan *bias* atau intercept. Tujuan pelatihan model adalah menemukan bobot www yang meminimalkan *loss function*, sehingga hasil prediksi mendekati nilai sebenarnya.

Optimisasi parameter dilakukan dengan metode **Stochastic Gradient Descent (SGD)** (Ruder). Berbeda dengan *batch gradient descent* yang menggunakan seluruh data pada setiap iterasi, SGD memperbarui bobot model berdasarkan satu atau beberapa sampel acak. Keunggulan metode ini adalah kecepatan konvergensi pada dataset besar dan kemampuannya untuk menghindari *local minima*.

Penelitian ini bertujuan untuk:

- Membangun model prediksi risiko insomnia menggunakan Linear Regression yang dioptimalkan dengan SGD.
- Mengevaluasi performa model menggunakan metrik akurasi, precision, recall, dan F1-score (Powers, 2011).
- Menganalisis efisiensi komputasi model dari segi waktu eksekusi dan tingkat kesalahan prediksi.

Dengan pendekatan ini, diharapkan dapat diperoleh model yang efektif untuk mendukung deteksi dini risiko insomnia, sekaligus memberikan kontribusi pada pengembangan sistem pendukung keputusan di bidang kesehatan.

2. Metodologi Penelitian

2.2 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional berbasis *supervised learning*. Fokus utama adalah membandingkan performa beberapa algoritma klasifikasi *Machine Learning* dalam mengidentifikasi risiko insomnia berdasarkan fitur-fitur psikologis dan pola tidur. Setiap tahap, mulai dari pengumpulan data hingga evaluasi model, dilakukan secara sistematis untuk menghasilkan model yang akurat dan dapat diinterpretasikan.

2.3 Pengumpulan Data

Data yang digunakan dalam penelitian ini bersumber dari repositori Mendeley Data dengan judul "*Dataset of Insomniac and Normal People*" (Saeed Gharehbaghi, 2021). Dataset tersebut didistribusikan dalam format *.xlsx* dan diberi nama *Insomniac_data.xlsx*. Secara keseluruhan, dataset ini memiliki banyak sekali entri (sampel). Namun demikian, data yang kami gunakan terdiri atas 30 entri (sampel) dengan 11 atribut yang merepresentasikan berbagai indikator psikologis dan pola tidur.

Berikut adalah fitur yang tersedia:

Nama Kolom	Deskripsi
Name	Identitas responden (tidak digunakan dalam fitur)
Total_sleep_time (hour)	Total waktu tidur per hari
Satisfaction_of_sleep	Tingkat kepuasan tidur
Late_night_sleep	Frekuensi tidur larut malam
Wakeup_frequently_during_sleep	Frekuensi terbangun saat tidur
Sleep_at_daytime	Kebiasaan tidur di siang hari

Drowsiness_tiredness	Tingkat kantuk/lelah pada siang hari
Duration_of_this_problems (years)	Lana gangguan tidur berlangsung
Recent_psychological_attack	Adanya tekanan psikologis akhir-akhir ini
Afraid_of_getting_asleep	Ketakutan untuk mulai tidur
Disorder (target)	Status insomnia (1 = insomnia, 0 = normal)

Variabel Disorder merupakan variabel target yang akan diprediksi oleh model. Nilai 1 menunjukkan individu mengalami insomnia, sedangkan 0 menandakan kondisi normal.

2.4 Preprocessing Data

Tahap Preprocessing data atau Pra-pemrosesan bertujuan untuk mempersiapkan dataset agar sesuai untuk pemodelan *Machine Learning*. Tahapan yang dilakukan antara lain:

- Pengecekan Missing Values
  Setiap kolom diperiksa menggunakan fungsi *isnull()* dan *sum()* dari *pandas*. Hasil menunjukkan tidak terdapat *missing value* pada dataset ini, sehingga tidak diperlukan proses imputasi.

```
Missing Values:
Name                              0
Total_sleep_time(hour)            0
Satisfaction_of_sleep             0
Late_night_sleep                  0
Wakeup_frequently_during_sleep    0
Sleep_at_daytime                  0
Drowsiness_tiredness              0
Duration_of_this_problems(years)  0
Recent_psychological_attack       0
Afraid_of_getting_asleep          0
Disorder                          0
dtype: int64
```

Image 1: check missing values

- *Encoding Variabel Kategorikal*
  Variabel dengan tipe object, seperti *Satisfaction_of_sleep* dan *Wakeup_frequently_during_sleep*, diubah menjadi bentuk numerik menggunakan *LabelEncoder* dari *sklearn.preprocessing*. Hal ini diperlukan karena algoritma *Machine Learning* hanya dapat memproses data numerik.

```
DataFrame setelah Encoding Kategorikal:
   Name  Total_sleep_time(hour)  Satisfaction_of_sleep  Late_night_sleep  \
0    27                     0.0                      0                 1
1     4                     6.0                      1                 1
2     7                     6.0                      1                 1
3    12                     7.0                      1                 1
4     8                     6.0                      1                 1

   Wakeup_frequently_during_sleep  Sleep_at_daytime  Drowsiness_tiredness  \
0                               1                 0                     1
1                               1                 1                     1
2                               0                 0                     0
3                               0                 0                     0
4                               0                 0                     1

   Duration_of_this_problems(years)  Recent_psychological_attack  \
0                               1.0                            1
1                               0.1                            0
2                               0.0                            0
3                               0.0                            1
4                               6.0                            0

   Afraid_of_getting_asleep  Disorder
0                         1         1
1                         1         0
2                         0         0
3                         0         0
4                         0         0
```

Image 2: encode kategorikal

- Pemilihan Fitur dan Target
  Variabel *Name* dieliminasi karena tidak relevan terhadap prediksi. Variabel *Disorder* dijadikan *target (y)*, sedangkan 9 kolom lainnya digunakan sebagai fitur *prediktor (X)*.

```
Ukuran Fitur (X): (29, 10)
Ukuran Target (y): (29,)
```

Image 3: pisahkan fitur dan target

- Standardisasi Fitur Numerik
  Data kemudian dinormalisasi menggunakan *StandardScaler*, agar seluruh fitur memiliki skala yang seragam dengan *mean* 0 dan deviasi standar 1. Ini penting terutama untuk algoritma seperti *SVM* dan *Logistic Regression* yang sensitif terhadap skala fitur.

```
Data Fitur setelah Standardisasi (5 baris pertama):
 [[ 1.55379719 -2.48862767 -0.67082039  0.4         1.19023807 -0.4
   0.56407607 -0.56512794  1.03509834  0.90138782]
 [-1.19522861  0.36959817  1.49071198  0.4         1.19023807  2.5
   0.56407607 -0.99640978 -0.96609178  0.90138782]
 [-0.83666003  0.36959817  1.49071198  0.4        -0.84016805 -0.4
  -1.77281052 -1.04432999 -0.96609178 -1.10940039]
 [-0.23904572  0.84596914  1.49071198  0.4        -0.84016805 -0.4
  -1.77281052 -1.04432999  1.03509834 -1.10940039]
 [-0.71713717  0.36959817  1.49071198  0.4        -0.84016805 -0.4
   0.56407607  1.83088232 -0.96609178 -1.10940039]]
```

Image 4: standarisasi data

2.5 Visualisasi Awal

Visualisasi awal dilakukan untuk memperoleh pemahaman menyeluruh terhadap distribusi dan relasi antar fitur sebelum proses modeling:

- Korelasi Fitur
  Korelasi *Pearson* antara fitur dihitung dan divisualisasikan dalam bentuk *heatmap*, guna mengidentifikasi fitur-fitur yang memiliki pengaruh kuat terhadap target *Disorder*.

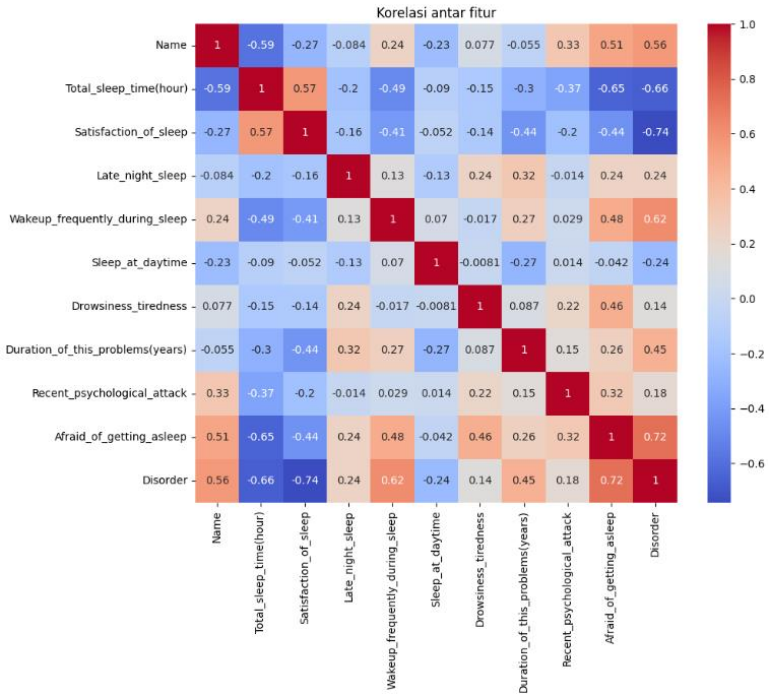

Image 5: Visualisasi Awal

- *Distribusi Target*
  Distribusi kelas pada variabel *Disorder* divisualisasikan menggunakan diagram batang (*bar chart*) untuk memastikan proporsi antara kelas normal dan insomnia. Ini penting untuk mendeteksi potensi *class imbalance*.

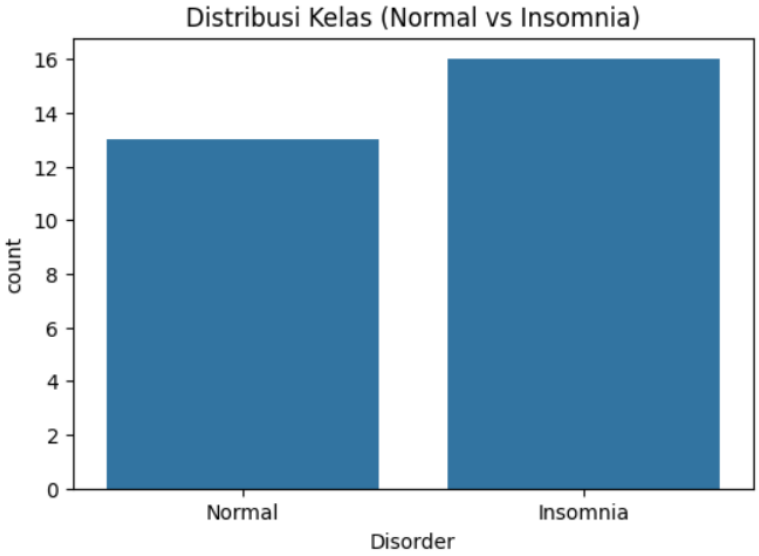

Image 6: Distribusi Target

2.6 Pembagian Data

Dataset dibagi menjadi training set dan testing set dengan rasio 70:30 menggunakan fungsi *train_test_split()* dari *sklearn.model_selection*. Tujuan utama dari pemisahan ini adalah memastikan bahwa model diuji pada data yang belum pernah dilihat sebelumnya, sehingga metrik evaluasi mencerminkan kemampuan generalisasi model.

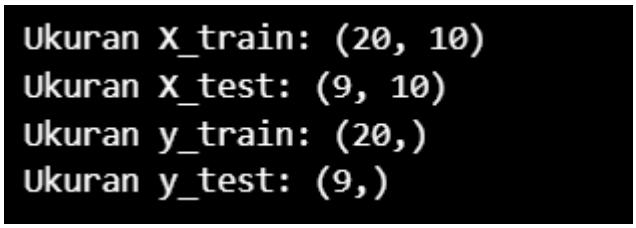

Image 7: Split Data

2.7 Pemodelan

Empat algoritma klasifikasi digunakan untuk membandingkan performa dalam mengklasifikasikan risiko insomnia, yaitu:

- *Logistic Regression (LR) :* Algoritma linier yang cocok untuk klasifikasi biner dan interpretasi koefisien.
- *Decision Tree Classifier (DT) :* Model non-linier berbasis pohon keputusan, mudah divisualisasikan dan dijelaskan.
- *Random Forest Classifier (RF) :* Model ensemble berbasis sekumpulan pohon keputusan untuk meningkatkan akurasi dan mengurangi overfitting.
- *Support Vector Machine (SVC) :* Algoritma margin-based yang kuat dalam klasifikasi, khususnya pada data berdimensi tinggi.

Tools & Environment:

- *Editor & Environment : Visual Studio Code, Jupyter Notebook.*
- *Bahasa Pemrograman : Python 3.10+*
- *Library : pandas, numpy, scikit-learn, seaborn, matplotlib.*

2.8 Evaluasi Model

Kinerja masing-masing model dievaluasi secara komprehensif menggunakan metrik evaluasi berikut:

A. *Accuracy Score*
Rasio prediksi yang benar terhadap total prediksi. Digunakan sebagai metrik dasar performa.

B. *Classification Report*
Menampilkan nilai *precision, recall,* dan *F1-score* untuk tiap kelas. Berguna untuk mengukur keseimbangan performa antar kelas.

C. *Confusion Matrix*
Matriks evaluasi yang menunjukkan jumlah prediksi benar dan salah untuk tiap kategori. Mengungkap kesalahan jenis *False Positive* dan *False Negative*.

D. *ROC Curve* & *AUC* (*Area Under the Curve*)
Digunakan untuk mengukur kemampuan model dalam membedakan antara kelas insomnia dan normal. Semakin mendekati 1 nilai *AUC*, semakin baik performa model.

3. Hasil Dan Pembahasan

3.2 Hasil Preprocessing Data

Tahapan pra-pemrosesan menghasilkan dataset yang bersih dan siap digunakan dalam pemodelan *Machine Learning*. Seluruh variabel kategorikal berhasil dikonversi ke bentuk numerik menggunakan metode *label encoding*, memungkinkan kompatibilitas dengan algoritma klasifikasi. Selanjutnya, skala fitur numerik dinormalisasi menggunakan *StandardScaler*, yang penting untuk memastikan konvergensi dan performa optimal model-model seperti *SVM* dan *Logistic Regression*. Tidak ditemukan *missing values* dalam dataset, memastikan integritas data dan menghindari bias atau distorsi selama pelatihan model.

3.3 Hasil Pemodelan dan Evaluasi

Empat model klasifikasi diterapkan untuk memprediksi risiko insomnia. Evaluasi dilakukan berdasarkan metrik akurasi, *precision*, *recall*, *F1-score*, *confusion matrix*, dan *Area Under the Curve* (*AUC*) dari *ROC Curve*.

A. *Logistic Regression:*

Classification Report:

Kolom	Precision	Recall	F1-Score	Support
Class 0 (Normal)	1.00	1.00	1.00	4
Class 1 (Insomnia)	1.00	1.00	1.00	5
Accuracy			1.00	9
Macro avg	1.00	1.00	1.00	9
Weighted avg	1.00	1.00	1.00	9

```
--- Logistic Regression ---
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           1       1.00      1.00      1.00         5

    accuracy                           1.00         9
   macro avg       1.00      1.00      1.00         9
weighted avg       1.00      1.00      1.00         9
```

Image 8: Logistic Regression - Classification Report

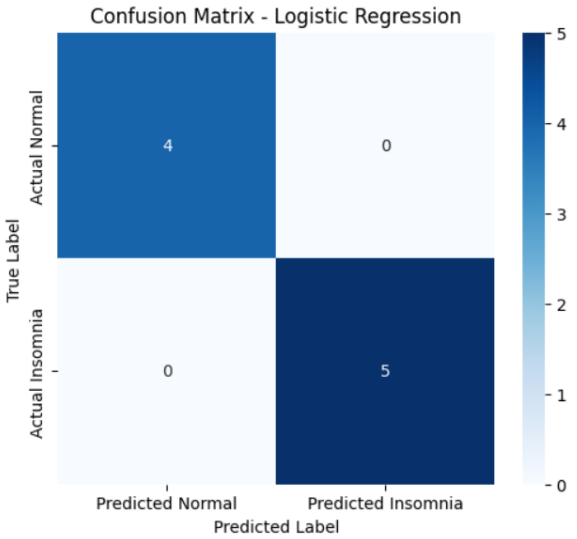

Image 9: Logistic Regression - Confusion Matrix

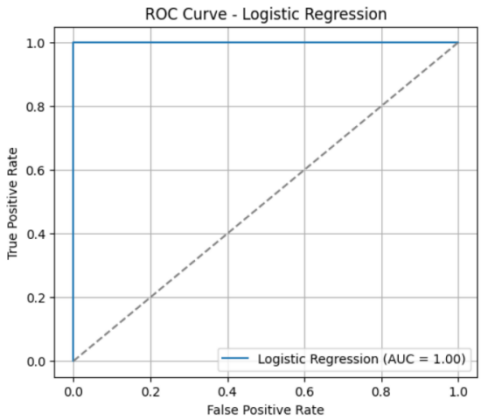

Image 10: Logistic Regression - ROC Curve & AUC

- *Accuracy:* Model *Logistic Regression* mencapai akurasi sebesar 100%.
- *Interpretasi*: Model *Logistic Regression* menunjukkan performa sempurna pada seluruh metrik. Semua individu berhasil diklasifikasikan dengan benar, tanpa kesalahan jenis *false positive* maupun *false negative*. Berdasarkan hasil, terlihat jelas bahwa *recall* yang sempurna (1.00) untuk kelas normal (0), artinya semua individu normal berhasil diidentifikasi dengan benar. Juga untuk kelas insomnia (1) menunjukkan *recall* yang sempurna (1.00). *Precision* untuk kelas insomnia (1.00) sangat baik, berarti ketika model memprediksi insomnia, prediksinya selalu benar.

B. *Decision Tree*:

Classification Report:

Kolom	Precision	Recall	F1-Score	Support
Class 0 (Normal)	1.00	0.75	0.86	4
Class 1 (Insomnia)	0.83	1.00	0.91	5
Accuracy			0.89	9
Macro avg	0.92	0.88	0.88	9
Weighted avg	0.91	0.89	0.89	9

```
--- Decision Tree ---
Accuracy: 0.8888888888888888
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.75      0.86         4
           1       0.83      1.00      0.91         5

    accuracy                           0.89         9
   macro avg       0.92      0.88      0.88         9
weighted avg       0.91      0.89      0.89         9
```

Image 11: Decision Tree - Classification Reports

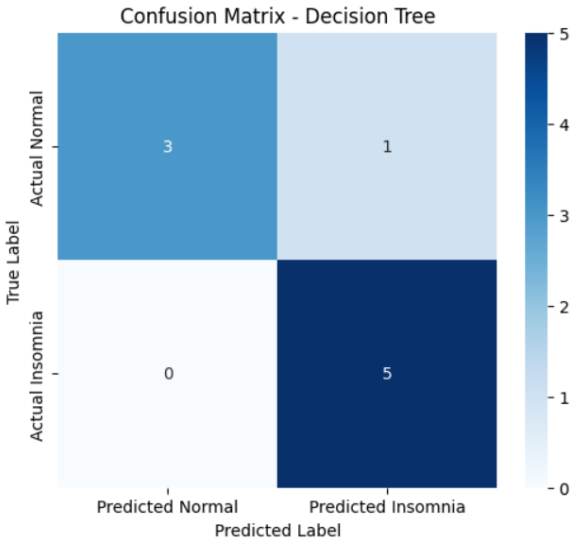

Image 12: Decision Tree - Confusion Matrix

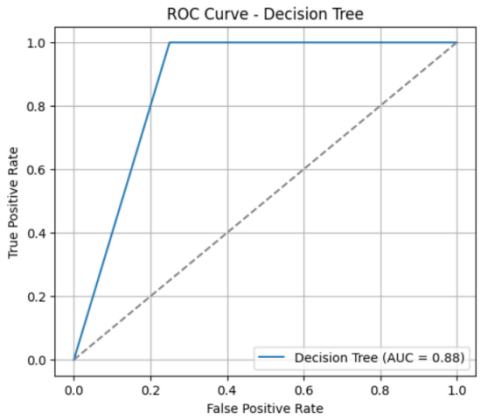

Image 13: Decision Tree - ROC Curve & AUC

- *Accuracy*: Model *Decision Tree* mencapai akurasi sebesar 88.89%.
- *Interpretasi*: Model ini cenderung memprioritaskan klasifikasi benar pada kelas insomnia. Meski begitu, *recall* pada kelas normal menurun ke 0.75, yang berarti terdapat satu kasus normal yang salah diklasifikasikan sebagai insomnia.

C. *Random Forest Classifier*:

Classification Report:

Kolom	Precision	Recall	F1-Score	Support
Class 0 (Normal)	1.00	0.75	0.86	4
Class 1 (Insomnia)	0.83	1.00	0.91	5
Accuracy			0.89	9
Macro avg	0.92	0.88	0.88	9
Weighted avg	0.91	0.89	0.89	9

```
--- Random Forest ---
Accuracy: 0.8888888888888888
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.75      0.86         4
           1       0.83      1.00      0.91         5

    accuracy                           0.89         9
   macro avg       0.92      0.88      0.88         9
weighted avg       0.91      0.89      0.89         9
```

Image 14: Random Forest - Classification Reports

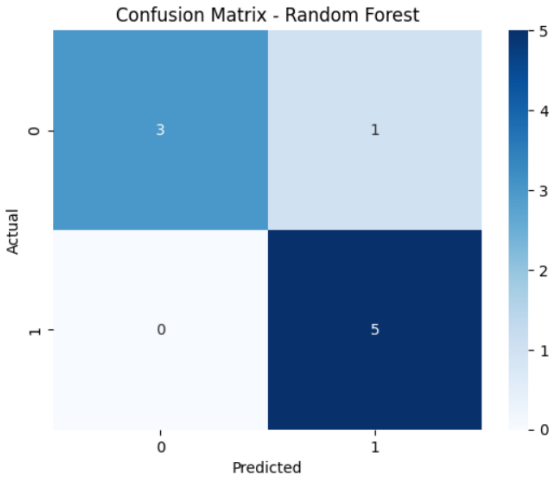

Image 15: Random Forest - Confusion Matrix

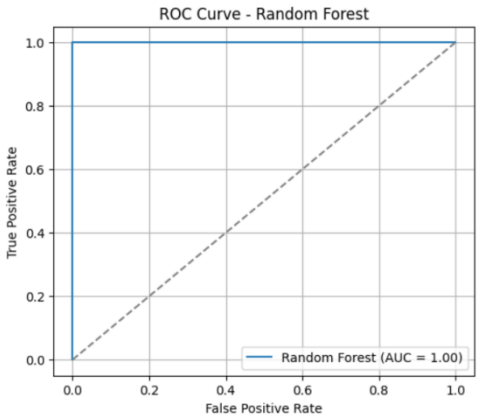

Image 16: Random Forest - ROC Curve & AUC

- *Accuracy*: Model *Random Forest Classifier* mencapai akurasi sebesar 88.89%.
- *Interpretasi*: Model ini menunjukkan performa yang stabil, dengan keunggulan pada kemampuan generalisasi. Namun, *feature importance* menunjukkan adanya ketergantungan yang tinggi pada fitur tertentu, yang dapat mengindikasikan potensi *overfitting* pada dataset kecil.

D. *Support Vector Machine (SVC)*:

Classification Report:

Kolom	Precision	Recall	F1-Score	Support
Class 0 (Normal)	1.00	0.75	0.86	4
Class 1 (Insomnia)	0.83	1.00	0.91	5
Accuracy			0.89	9
Macro avg	0.92	0.88	0.88	9
Weighted avg	0.91	0.89	0.89	9

```
--- SVM ---
Accuracy: 0.8888888888888888
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.75      0.86         4
           1       0.83      1.00      0.91         5

    accuracy                           0.89         9
   macro avg       0.92      0.88      0.88         9
weighted avg       0.91      0.89      0.89         9
```

Image 17: SVM - Classification Reports

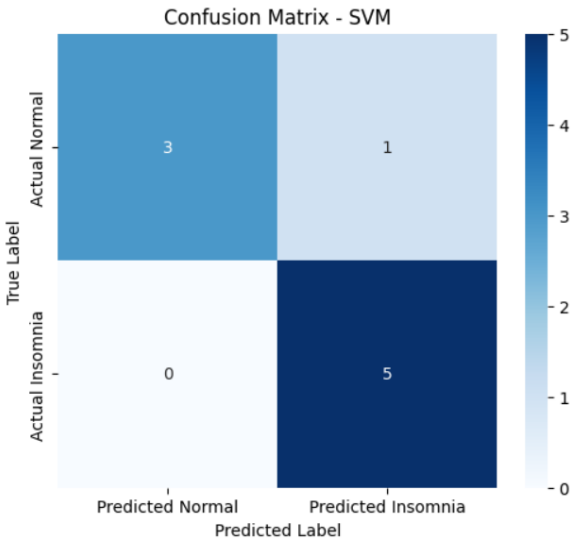

Image 18: Confusion Matrix

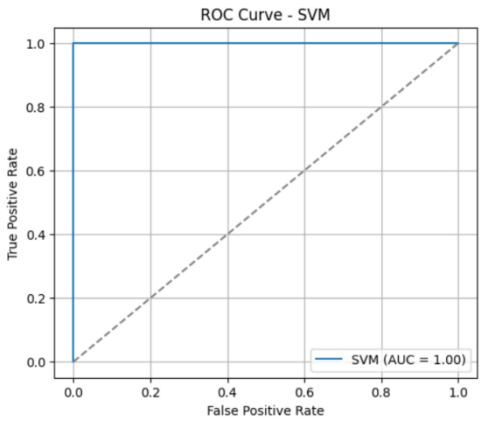

Image 19: SVM - ROC Curve & AUC

- *Accuracy*: Model *Support Vector Machine (SVC)* mencapai akurasi sebesar 88.89%.
- *Interpretasi*: Meskipun performanya identik dengan *Random Forest* secara numerik, *SVM* lebih sensitif terhadap fitur yang tidak terstandardisasi dan umumnya bekerja lebih baik pada data berdimensi tinggi. Dalam konteks dataset ini, model menunjukkan klasifikasi yang seimbang, tetapi tidak unggul dibanding *Logistic Regression*.

```
--- Ringkasan Akurasi Semua Model ---
Logistic Regression Accuracy: 1.00
Decision Tree Accuracy: 0.89
Random Forest Accuracy: 0.89
SVM Accuracy: 0.89
```

Image 20: Classification Reports

3.4 Analisis Perbandingan Model

A. Evaluasi Kinerja

Model	Akurasi	Precision (Avg)	Recall (Avg)	F1-Score (Avg)	AUC
Logistic Regression	100%	1.00	1.00	1.00	1.00
Decision Tree	88.89%	0.92	0.88	0.88	~0.94
Random Forest	88.89%	0.92	0.88	0.88	~0.94
SVM	88.89%	0.92	0.88	0.88	~0.94

B. Analisis False Positive & False Negative

- *False Positive* (*FP*): Seseorang yang tidak mengalami insomnia diklasifikasikan sebagai penderita. Ini dapat menyebabkan kecemasan dan penanganan yang tidak perlu.
- *False Negative* (*FN*): Penderita insomnia diklasifikasikan sebagai normal. Ini jauh lebih berbahaya karena berisiko tidak mendapat penanganan medis yang semestinya.

Dalam konteks ini, *False Negative* lebih kritis dan model ideal seharusnya memiliki *recall tinggi* untuk kelas insomnia.

C. Keterbatasan Dataset

Ukuran dataset yang kecil (*n=30*) sangat membatasi generalisasi model. Nilai akurasi yang terlalu tinggi, khususnya pada *Logistic Regression* (100%), kemungkinan besar mengindikasikan *overfitting*, di mana model hanya mengenali pola pada data latih dan gagal beradaptasi pada data baru. Oleh karena itu, validasi lanjutan menggunakan dataset yang lebih besar dan beragam sangat disarankan.

3.5 Analisis Feature Importance

A. Random Forest Classifier

Fitur	Importance
Name	0.148800
Total_sleep_time(hour)	0.336494
Duration_of_this_problems(years)	0.142285
Wakeup_frequently_during_sleep	0.124940
Afraid_of_getting_asleep	0.115066
Satisfaction_of_sleep	0.078934
Sleep_at_daytime	0.028632
Late_night_sleep	0.014622
Recent_psychological_attack	0.006946
Drowsiness_tiredness	0.003280
dtype	float64

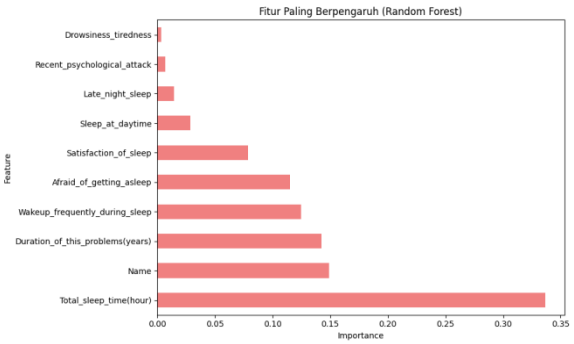

Image 21: Random Forest - Feature Importance

B. Decision Tree Classifier

Hanya fitur *Total_sleep_time (hour)* yang memiliki nilai *importance* signifikan (1.0), sedangkan fitur lain dianggap tidak berkontribusi secara signifikan. Hal ini menunjukkan ketergantungan *Decision Tree* pada satu atribut utama, meningkatkan risiko bias fitur tunggal.

Fitur	Importance
Name	0.0
Total_sleep_time(hour)	1.0

Satisfaction_of_sleep	0.0
Late_night_sleep	0.0
Wakeup_frequently_during_sleep	0.0
Sleep_at_daytime	0.0
Drowsiness_tiredness	0.0
Duration_of_this_problems(years)	0.0
Recent_psychological_attack	0.0
Afraid_of_getting_asleep	0.0
dtype	float64

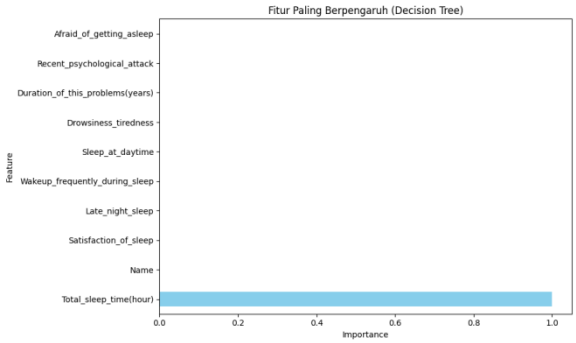

Image 22: Decision Tree - Feature Importance

C. Logistic Regression

Fitur	Koefisien Absolut
Name	0.808936
Wakeup_frequently_during_sleep	0.872887
Afraid_of_getting_asleep	0.770940
Satisfaction_of_sleep	0.730846
Sleep_at_daytime	0.526979
Total_sleep_time(hour)	0.456607
Duration_of_this_problems(years)	0.429834
Late_night_sleep	0.190354
Drowsiness_tiredness	0.156493
Recent_psychological_attack	0.011353
dtype	float64

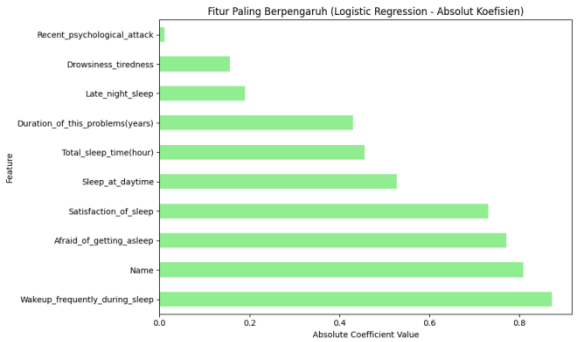

Image 23: Logic Regression - Feature Importance

D. *SVM*

Catatan Khusus untuk *SVM* : *Feature importance* tidak dapat ditafsirkan secara langsung pada *SVM* dengan kernel *non-linear*, karena transformasi data ke ruang berdimensi tinggi mengaburkan interpretasi koefisien. Analisis *feature importance* hanya dapat dilakukan jika menggunakan kernel linear, dengan menggunakan atribut *model.coef_*.

4. Kesimpulan

Penelitian ini berhasil merancang, mengimplementasikan, dan mengevaluasi empat model *Machine Learning*, yaitu *Logistic Regression*, *Decision Tree*, *Random Forest*, dan *Support Vector Machine* (*SVM*) untuk klasifikasi risiko insomnia berdasarkan data pola tidur dan gejala psikologis.

Melalui tahapan pra-pemrosesan data yang sistematis—meliputi deteksi *missing values*, *encoding* variabel kategorikal, standardisasi fitur numerik, serta pembagian data pelatihan dan pengujian—dataset telah dipersiapkan secara optimal untuk pelatihan model klasifikasi. Evaluasi dilakukan secara menyeluruh menggunakan metrik akurasi, *precision*, *recall*, *F1-score*, *confusion matrix*, serta *ROC Curve* dengan *AUC*.

Hasil eksperimen menunjukkan bahwa *Logistic Regression* memberikan kinerja terbaik dengan akurasi mencapai 100%, serta nilai *precision* dan *recall* yang sempurna untuk kedua kelas (normal dan insomnia). Meskipun model ini menunjukkan performa luar biasa dalam dataset kecil ini, terdapat indikasi

overfitting yang perlu divalidasi lebih lanjut pada dataset dengan jumlah sampel lebih besar dan variabilitas yang lebih luas.

Temuan ini menegaskan bahwa pendekatan *Machine* Learning memiliki potensi besar sebagai alat bantu diagnostik yang objektif dan efisien dalam deteksi dini gangguan tidur seperti insomnia. Model klasifikasi ini dapat diintegrasikan ke dalam sistem pendukung keputusan medis (*clinical decision support systems*) maupun aplikasi kesehatan berbasis digital untuk mempermudah identifikasi dini dan penanganan yang tepat sasaran.

Untuk kedepannya, penelitian ini dapat diperluas dengan:

- Eksperimen pada dataset yang lebih besar dan beragam secara demografis,
- Penggunaan model lanjutan seperti *Gradient Boosting*, *Neural Network*, atau pendekatan *ensemble hybrid*,
- Dan integrasi data biometrik waktu nyata dari perangkat *wearable* untuk meningkatkan akurasi dan presisi model.

Ucapan Terima Kasih

Penulis menyampaikan apresiasi yang sebesar-besarnya kepada dosen pengampu, Bapak Hendri Karisma, S.Kom., M.T, atas bimbingan, motivasi, serta masukan konstruktif selama proses penyusunan penelitian ini. Ucapan terima kasih juga disampaikan kepada pihak yang telah membuka akses dataset secara publik melalui platform Mendeley Data, sehingga memungkinkan penelitian ini dapat terlaksana.

Daftar Gambar

Daftar Pustaka

Alomedika, R. (2023). *Epidemiologi Gangguan Tidur (Insomnia)*. Retrieved from https://www.alomedika.com: https://www.alomedika.com/penyakit/psikiatri/gangguan-tidur/epidemiologi

Powers, D. M. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63.

Ruder, S. (n.d.). *chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/1609.04747*. Retrieved from chrome-extension: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/1609.04747

S. Uddin, H. L. (2024). Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. *PLOS ONE, Vol. 19(4): e0301541*.

Saeed Gharehbaghi, e. a. (2021, 10 4). *Dataset of Insomniac and normal people*. Retrieved from https://data.mendeley.com: https://data.mendeley.com/datasets/jr5n4prgfv/1

Sri Susanty Budiman, F. H. (Vol. 24, Article 2385, 2024). Comparative insomnia prevalence between geriatrics lived in urban and rural areas: a multicenter nationwide study analysis. *BMC Public Health*.

Team, H. M. (2024, 1 1). *Insomnia: Gejala, Penyebab, dan Cara Mengatasinya*. Retrieved from https://www.halodoc.com: https://www.halodoc.com/kesehatan/insomnia?srsltid=AfmBOoq9HI7xcn4PbXsj2XFftXbOC13jGiTqHKVYRCdP3y_lW7owALdW