Optimalisasi Klasifikasi *Sleep Disorder* Menggunakan Metode *Decision Tree*, *Random Forest*, *K-Nearest Neighbor* (KNN) dan *Support Vector Machine*

Salma Zaura Baraza[1*] and Yola Darma Dhaifulla [2]
[1]Salma Zaura Baraza: Statistika, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Yola Darma Dhaifulla: Statistika, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
*Penulis Korespondensi: 5003211151@student.its.ac.id

ABSTRAK – Gangguan tidur merupakan masalah kesehatan yang banyak dialami diberbagai kalangan yang dapat berdampak buruk pada kualitas hidup dan kesehatan seseorang. Jenis gangguan tidur seperti insominia dan sleep apnea dapat terjadi karena beberapa faktor seperti durasi tidur, kualitas tidur, tingkat stress dan variabel lain yang diduga berpengaruh terhadap gangguan tidur pada penelitian ini. Metode klasifikasi *Decision Tree, Random Forest, K-Nearest Neighbor* (KNN), dan *Support Vector Machine* (SVM) diterapkan untuk mencari klasifikasi terbaik dalam mengidentifikasi ganguan tidur. Hasil penelitian mengindikasikan bahwa keempat metode klasifikasi mampu mengklasifikasikan gangguan tidur seseorang dengan kategori *excellent classification*. *Support Vector Machine* (SVM) dengan hasil akurasi 0,90 dan AUC 0,92 menjadikan algoritma ini paling baik digunakan untuk mengklasifikasikan ganguan tidur.

Kata Kunci– Gangguan tidur, *Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine*

I. PENDAHULUAN

Gangguan tidur adalah masalah kesehatan yang banyak dialami dan cukup populer di berbagai kalangan. Gangguan tidur dapat berdampak buruk pada kualitas hidup dan kesehatan seseorang. Beberapa jenis gangguan tidur, seperti insomnia dan sleep apnea, dapat mempengaruhi pola tidur dan fungsi harian. Insomnia, misalnya, menyebabkan kesulitan untuk tidur atau mempertahankan tidur, sementara sleep apnea ditandai dengan berhentinya napas secara periodik selama tidur.

Untuk menganalisis dan mengklasifikasikan gangguan tidur, metode klasifikasi data menggunakan machine learning dapat diterapkan. Metode ini memperkenalkan kemampuan deteksi otomatis yang efisien dan efektif. Dalam studi ini, kami menggunakan dataset "The Sleep Health and Lifestyle" yang mencakup berbagai variabel, seperti durasi tidur, kualitas tidur, tingkat stres, dan aktivitas fisik, untuk membangun model prediktif yang akurat.

Tujuan utama dari penelitian ini adalah untuk mencari metode klasifikasi terbaik dalam mengidentifikasi gangguan tidur. Beberapa algoritma yang digunakan dalam klasifikasi ini adalah Decision Tree, Random Forest, K-Nearest Neighbor (KNN), dan Support Vector Machine (SVM). Algoritma-algoritma ini memungkinkan kami untuk menentukan faktor-faktor yang paling berpengaruh terhadap gangguan tidur. Penelitian ini adalah sebagai sarana untuk menerapkan ilmu statistika yang telah dipelajari selama menempuh pendidikan di Statistika ITS bagi peneliti, sebagai bahan informasi untuk masyarakat mengenai indikasi gangguan tidur yang mungkin dialami, dan sebagai referensi untuk peneliti selanjutnya yang menggunakan metode Decision Tree, Random Forest, K-Nearest Neighbor (KNN), dan SVM.

II. TINJAUAN PUSTAKA

A. Sleep Disorder

Gangguan tidur merupakan sekelompok kondisi yang ditandai dengan adanya gangguan dalam kuantitas, kualitas, atau durasi tidur. Gangguan tidur merupakan masalah kesehatan yang signifikan karena dapat berdampak negatif pada kualitas hidup dan kesejahteraan individu. Faktor-faktor yang dapat mempengaruhi gangguan tidur meliputi kualitas tidur, durasi tidur, tingkat stres, pekerjaan, tingkat aktivitas, dan faktor lainnya. Dalam data ini, gangguan tidur terdiri dari insomnia, sleep apnea, dan tidak ada gangguan tidur. Insomnia adalah kondisi di mana seseorang mengalami kesulitan untuk tidur. Sleep apnea adalah kondisi klinis yang ditandai dengan kesulitan bernafas saat tidur, yang dapat menurunkan kualitas tidur seseorang [5].

B. Decision Tree

Decision Tree merupakan metode klasifikasi yang dibangun untuk mendapatkan sebuah kesimpulan dari sejumlah data. Metode ini merupakan salah satu algoritma *supervised learning* yang melakukan prediksi menggunakan struktur pohon. *Decision tree* terbuat dari tiga *node* yaitu leaf node, lalu *root node* yang merupakan titik awal dari suatu *decision tree*, dan yang terakhir adalah simpul perantara atau *internal node* yang berhubungan dengan suatu pengujian [16]. *Decision tree* memiliki beberapa jenis seperti *Classification and Regression Tree* (CART), C4.5, C5.0, ID.3, dan lainnya [11]. *Decision Tree* dalam proses prediksinya melakukan perhitungan dengan mencari ukuran ketidakmurnian atau impurity measure. Perhitungan matematika dari ketidakmurnian ini dapat dilihat pada persamaan (1) dan (2).

Gini Impurity

$$Gini = 1 - \sum_i^n Pi^2 \quad (1)$$

Keterangan:
n : jumlah dari masing-masing atribut
Pi : jumlah atribut masing-masing kelas atau labelnya

Average Gini Impurity

$$AG = \sum \frac{data\ point\ i}{jumlah\ total\ data\ point} x\ Gi \quad (2)$$

Gini Impurity digunakan untuk menentukan pemisahan optimal pada simpul akar dan simpul-simpul berikutnya dalam Decision Tree. Ini mengukur seberapa sering elemen yang dipilih secara acak dari suatu kumpulan data salah diklasifikasikan. Perhitungan dalam pemilihan atribut sebagai akar dilakukan dengan menghitung selisih antara Gini Impurity dan Average Gini Impurity, yang dapat dilihat pada persamaan (3).

$$IG = Gi - AG \quad (3)$$

C. Random Forest

Metode *random forest* merupakan pendekatan yang dapat meningkatkan akurasi dengan melakukan pemilihan simpul anak secara acak untuk setiap node. Metode ini digunakan untuk membangun pohon keputusan yang terdiri dari *root node, internal node,* dan *leaf node* dengan memilih atribut dan data secara acak sesuai aturan yang ditetapkan. *Root node* adalah simpul yang terletak paling atas, sering disebut sebagai akar dari pohon Keputusan. *Internal node* adalah simpul percabangan yang memiliki minimal dua output dan hanya satu input. Sementara *leaf node* atau *terminal node* adalah simpul terakhir yang hanya memiliki satu input dan tidak memiliki output. Pohon keputusan dimulai dengan menghitung nilai *entropy* untuk menentukan tingkat ketidakmurnian atribut dan nilai *information gain*. Perhitungan nilai *entropy* menggunakan rumus yang terdapat pada persamaan (4), sedangkan nilai *information gain* dihitung menggunakan persamaan (5) [12].

$$Entropy(Y) = -\sum_i p(c|Y) \log_2 p(c|Y) \quad (4)$$

Dimana Y merupakan himpunan kasus dan $p(c|Y)$ adalah proporsi nilai Y terhadap kelas c.

$$Information\ gain(Y, a) = Entropy(Y) - \sum_{veValues} \frac{|Yv|}{|Ya|} Entropy(Yv) \quad (5)$$

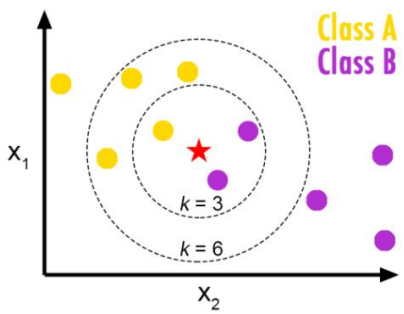
Dimana *values* (a) adalah semua nilai yang mungkin dalam himpunan kasus a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a. Ya adalah semua nilai yang sesuai dengan a.

D. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) merupakan metode klasifikasi terhadap objek berdasarkan data pembelajaran (*neighbor*) yang jaraknya paling dekat dengan objek tersebut [14]. KNN termasuk kelompok *instance-based learning*. Algoritma ini juga merupakan salah satu teknik *lazy learning.* Dekat atau jauhnya *neighbor* biasanya dihitung menggunakan jarak *Euclidean*. Diperlukan suatu sistem klasifikasi yang mampu mencari informasi [8]. Metode KNN terbagi menjadi dua fase: pembelajaran (*training*) dan klasifikasi atau pengujian (*testing*). Pada fase pembelajaran, algoritma ini hanya menyimpan vektor-vektor fitur dan klasifikasi dari data pelatihan. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk data yang akan diuji (dengan klasifikasi yang tidak diketahui). Jarak antara vektor baru ini dan seluruh vektor data pelatihan dihitung, kemudian diambil sejumlah k *neighbor* terdekat. Perhitungan jarak tetangga menggunakan algoritma *Euclidean* sebagaimana ditunjukkan pada persamaan 6.

$$euc = \sqrt{((a_1 - b_1)^2 + \cdots + (a_n - b_n)^2)} \quad (6)$$

Dimana $a = a_1, a_2, \ldots, a_n$ dan $b = b_1, b_2, \ldots, b_n$ mewakili n nilai atribut dari dua record. Untuk atribut dengan nilai kategori. Sebuah titik akan diprediksi jenisnya berdasarkan pada klasifikasi terbanyak dari *neighbor* di sekitarnya, ilustrasi dapat dilihat pada Gambar 1.

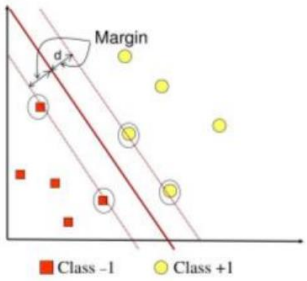


Gambar 1 Ilustrasi penggunaan nilai *k* pada metode KNN

Nilai k yang optimal untuk KNN tergantung pada data yang digunakan. Secara umum, nilai k yang lebih tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batas antara setiap klasifikasi menjadi lebih kabur. Nilai k yang baik dapat dipilih melalui optimasi parameter, seperti menggunakan *cross-validation*. Kasus khusus di mana klasifikasi diprediksi berdasarkan data pelatihan terdekat (dengan kata lain, k = 1) disebut algoritma *nearest neighbor* [9].

E. Support Vector Machine

Support Vector Machine (SVM) adalah metode pembelajaran supervised yang pertama kali diperkenalkan oleh Vapnik pada tahun 1995 dan telah terbukti sukses dalam melakukan prediksi, baik untuk kasus regresi maupun klasifikasi [19]. Tujuan utama dari metode ini adalah membangun *Optimal Separating Hyperplane* yang berfungsi sebagai pemisah optimal antara dua kelas pada input space. Gambar 2 menunjukkan sebuah data set yang terdiri dari dua kelas.



Gambar 2 *Hyperplane Optimum*

Hyperplane terbaik adalah *hyperplane* yang memiliki margin maksimal, yang ditentukan dari berbagai alternatif garis pemisah. Margin adalah jarak antara *hyperplane* dengan titik terdekat dari setiap kelas, sementara titik-titik terdekat tersebut disebut support vector [18]. Bidang pembatas pertama membatasi kelas pertama, sedangkan bidang pembatas kedua membatasi kelas kedua, sehingga didapat:

$$x_i w + b \geq +1, y_i = +1$$
$$x_i w + b \leq -1, y_i = -1 \quad (7)$$

w merupakan normal bidang dan b merupakan posisi bidang alternatif terhadap pusat koordinat. Pencarian *hyperplane* optimum dengan nilai margin maksimum dapat dirumuskan menjadi masalah optimasi konstrain ke dalam formula lagrange, yaitu:

$$L_{pri}(x, b, \alpha) = \frac{1}{2}\left|\left|w\right|\right|^2 - \sum_{i=1}^{n} \alpha_i [y_i(x_i^T w + b) - 1] \quad (8)$$

Dengan kendala $\alpha \geq 0$ (nilai koefisien lagrange)

Apabila data *training* tidak dapat dipisahkan secara linier, maka *classifier* yang diperoleh belum memiliki kemampuan generalisasi maksimal. Permasalahan tersebut bersifat *nonlinear separable* sehingga memerlukan solusi yaitu memetakan ruang input ke ruang yang berdimensi lebih tinggi. Fungsi kernel mampu menyelesaikan permasalahan SVM nonlinear, beberapa fugsi kernel yang dapat digunakan yaitu:

a. Linier
$$K(x_i, x_j) = x_i^T x_j \quad (9)$$
b. Polinomial
$$K(x_i, x_j) = (x_i^T x_j + 1)^d \quad (10)$$
c. Gaussian/Radial Basis Function
$$K(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) \quad (11)$$

Pemilihan fungsi kernel yang tepat merupakan hal yang sangat penting karena akan menentukan *feature space* dimana fungsi *classifier* yang akan dicari [10].

F. Confusion Matrix

Confusion matrix diartikan sebagai pengukuran performa pada *machine learning* dengan *output* berupa dua kelas atau lebih [1].

Tabel 1 *Confusion matrix*

Confusion matrix	*Classification*	
	***Positive* (+)**	***Negative* (-)**
Positive (+)	*True Positive*	*False Negative*
Negative (-)	*False Positive*	*True Negative*

Tabel 1 menunjukkan empat parameter berbeda yang dikombinasikan dari nilai prediksi dan nilai asli. Performa *machine learning* yang bagus atau tidak dari *confusion matrix* dengan melakukan perhitungan *accuracy precision, recall,* dan *f1-score* [4]. Berikut adalah beberapa persamaan untuk menghitung performa dari tabel *confusion matrix*.

$$Accuracy = \frac{TP+TN}{(TP+FP+FN+TN)} \quad (12)$$

Persamaan (12) menunjukkan akurasi, yaitu rasio prediksi yang benar terhadap total data. Hasil ini menggambarkan

seberapa tepat model dalam mengklasifikasikan data dengan benar.

$$Precision = \frac{TP}{(TP+FP)} \times 100\% \quad (13)$$

Precision adalah tingkat ketepatan data dari perbandingan prediksi yang benar (positif) dengan semua hasil prediksi yang benar (positif) tetapi bukan data yang benar, ditulis pada persamaan (13).

$$Recall = \frac{TP}{(TP+FN)} \times 100\% \quad (14)$$

Persamaan (14), *recall* merupakan perbandingan antara prediksi benar (positif) dengan seluruh data yang beanr (positif) tetapi prediksinya salah.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (15)$$

F1-score merupakan hasil yang didapat unutk melihat apakah hasil *precision* dan *recall* baik atau tidak dengan membandingkan di antara keduanya seperti pada persamaan (15). Parameter performa yang dilakukan dalam penelitian ini berupa *accuracy, precision, recall, f1-score,* serta waktu komputasi yaitu lama waktu proses *machine learning* bekerja.

G. ROC dan AUC

Receiver Operating Characteristic (ROC) memberikan visualisasi hubungan antara *False Positive* pada sumbu x dan *True Positive* pada sumbu y. Oleh karena itu, ROC pada dasarnya adalah visualisasi dari *confusion matrix*, yang mencakup semua *confusion matrix* yang mungkin dengan menggunakan threshold dari 0 hingga 1 untuk semua nilai *False Positive dan True Positive*. *Area Under Curve* (AUC) adalah nilai luas di bawah kurva ROC. Nilai AUC biasanya digunakan untuk membandingkan berbagai model, dengan model terbaik adalah yang memiliki nilai AUC tertinggi [7].

III. METODOLOGI

A. Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder berjudul "*Sleep Health and Lifestyle Dataset*" yang bersumber dari platform Kaggle. Data tersebut terdiri dari 375 observasi dan 13 variabel yang mencakup berbagai variabel terkait tidur dan kebiasaan sehari-hari. Unit penelitian dalam kasus ini adalah seseorang yang mengalami gangguan tidur dengan 3 kategori yang disebutkan, yaitu sleep disorder, insomnia, dan sleep apnea.

B. Variabel Penelitian

Penelitian ini menggunakan 10 variabel yang terdiri atas 1 variabel respon dan 9 lainnya merupakan variabel prediktor. Berikut disajikan lebih detail terkait variabel tersebut.

Tabel 2 Variabel Penelitian

Variabel	Keterangan	Deskripsi	Skala
Y	Sleep Disorder	0 : Insomnia (seseorang yang mengalami kesulitan untuk tertidur atau tetap tertidur). 1 : None (seseorang yang tidak mengalami gangguan tidur). 2 : Sleep Apnea (seseorang yang mengalami jeda pernapasan saat tidur).	Nominal
X_1	Age	Usia dalam tahun	Rasio
X_2	Occupation	Pekerjaan atau profesi	Nominal
X_3	Sleep Duration	Jumlah jam tidur per hari	Rasio
X_4	Quality of Sleep	Kualitas tidur (1-10)	Ordinal
X_5	Physical Activity Level	Jumlah menit seseorang melakukan aktivitas fisik	Rasio
X_6	BMI Category	Kategori BMI seseorang	Nominal
X_7	Heart Rate	Detak jantung per menit	Rasio
X_8	Daily Steps	Jumlah langkah per hari	Rasio
X_9	BP Low	Pengukuran tekanan diastolik	Rasio

C. Struktur Data

Struktur data yang digunakan dalam penelitian ini dijelaskan pada Tabel 3 berikut.

Tabel 3 Struktur Data

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
2	29	1	6,0	6	30	0	70	8000	80
2	30	4	6,4	5	35	1	78	4100	86

0	30	4	6,4	5	35	1	78	4100	86
2	31	1	7,7	7	75	0	70	8000	80
0	33	1	6,0	6	30	0	72	5000	80
…	…	...	...	...	...	...	...	...	...
1	31	4	7,9	8	75	1	69	6800	76
1	37	0	7,2	8	60	0	68	7000	75
1	53	2	8,5	9	30	0	65	5000	80
1	32	1	6.2	6	30	0	72	5000	80

D. Langkah Analisis

Adapun Langkah-langkah analisis pada penelitian ini adalah sebagai berikut.

1. Merumuskan masalah dan tujuan penelitian
2. Mengumpulkan data penelitian
3. Melakukan eksplorasi dan visualisasi data
4. Melakukan *preprocessing* data yang mencakup pengecekan data duplikat, *missing value*, dan *outliers.*
5. Melakukan feature selection menggunakan metode yang sesuai dengan model yang akan dibentuk
6. Melakukan klasifikasi dengan menggunakan motode decision tree
7. Melakukan klasifikasi dengan menggunakan metode random forest
8. Melakukan klasifikasi dengan menggunakan metode K-Nearest Neighbors (KNN)
9. Melakukan klasifikasi dengan menggunakan metode support vector machine (SVM)
10. Melakukan analisis hasil dari seluruh model
11. Membandingkan nilai akurasi dari seluruh model
12. Memilih model terbaik

IV. HASIL DAN PEMBAHASAN

A. Statistika Deskriptif

Tabel 4 Statistika Deskriptif

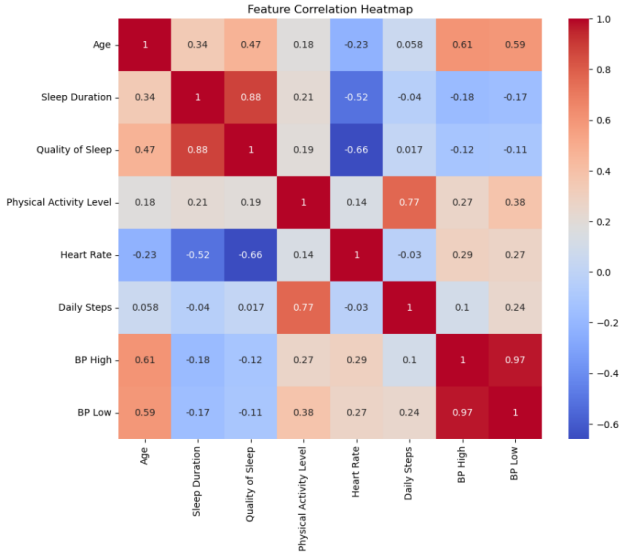
	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate	Daily Steps
Count	374	374	374	374	374	374	374
Mean	42,18	7,13	7,31	59,17	5,39	70,17	6816,84
Std	8,67	0,80	1,20	20,83	1,77	4,14	1617,92
Min	27,0	5,8	4,0	30,0	3,0	65,0	3000
25%	35,25	6,4	6,0	45,0	4,0	68,0	5600
50%	43,0	7,2	7,0	60,0	5,0	70,0	7000
75%	50,0	7,8	8,0	75,0	7,0	72,0	8000
Max	59,0	8,5	9,0	90,0	8,0	86,0	10000

Pada Tabel 4 di atas dapat dilihat bahwa rata-rata variabel Age seseorang yaitu 42 tahun dengan usia termuda adalah 27 tahun dan tertua adalah 59 tahun. Variabel sleep duration yang paling singkat pada seseorang yaitu 5,8 jam dan terlama yaitu 8,5 jam dengan rata-rata durasi tidur selama 7,1 jam. Variabel quality of sleep memiliki nilai dengan rating terendah sebesar 4 dan tertinggi sebesar 9 dengan rata-rata 7,3. Variabel physical activity level memiliki jumlah menit terendah pada seseorang yaitu minimal 30 menit dan tertinggi 90 menit dengan rata-rata tingkat aktivitas fisik seseorang 59 menit. Variabel stress level seseorang dengan rating nilai terendah sebesar 3 dan tertinggi 8 dengan rata-rata Tingkat stress 5,4. Variabel heart rate terendah seseorang yaitu 65 bpm dan tertinggi 86 bpm dengan rata-rata denyut jantung seseorang dalam detak per menit yaitu 70,2 bpm. Kemudian untuk variabel daily steps memiliki jumlah terendah sebesar 3000 langkah dan tertinggi sebesar 10000 langkah dengan rata-rata jumlah langkah yang dilakukan seseorang per hari sebesar 6816 langkah.

B. Visualisasi dan Eksplorasi Data

1. *Correlation Matrix*

Matriks korelasi adalah alat statistik untuk mengukur kekuatan dan arah hubungan antar variabel-variabel. Dalam dataset kesehatan tidur dan gaya hidup, matriks korelasi digunakan untuk mempelajari hubungan antara faktor-faktor seperti usia, durasi tidur, kualitas tidur, tingkat aktivitas fisik, detak jantung, jumlah langkah harian, tekanan darah tinggi, dan tekanan darah rendah. Nilai korelasi dapat berkisar dari -1 hingga 1, di mana nilai positif menunjukkan hubungan positif antara dua variabel, nilai negatif menunjukkan hubungan negatif, dan nilai 0 menunjukkan tidak adanya hubungan linier antara dua variabel tersebut.

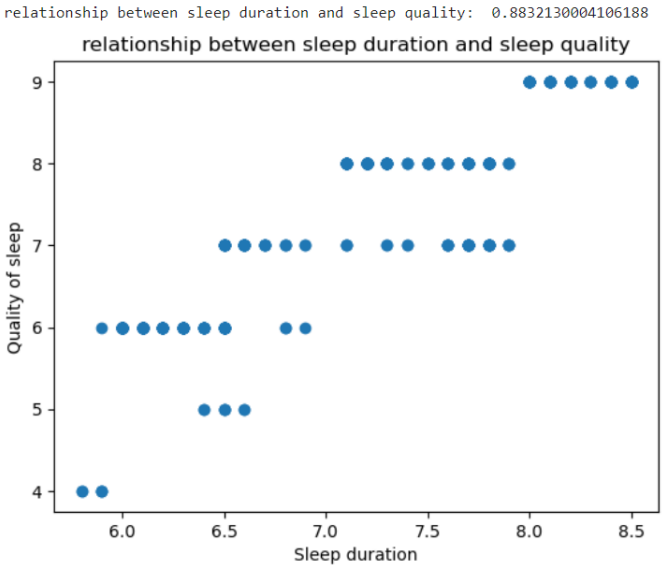


Gambar 3 Hasil *correlation matrix*

Berdasakan output *correlation matrix* pada Gambar 3, dapat diketahui bahwa terdapat beberapa hubungan positif yang signifikan. Durasi tidur berkorelasi positif tinggi dengan kualitas tidur (0.88), menunjukkan bahwa orang yang tidur lebih lama cenderung memiliki kualitas tidur yang lebih baik. Tingkat aktivitas fisik berkorelasi positif dengan detak jantung (0.77), yang menunjukkan bahwa aktivitas fisik yang lebih tinggi berhubungan dengan detak jantung yang lebih tinggi. Langkah harian berkorelasi positif dengan tekanan darah tinggi (0.27), menandakan bahwa jumlah langkah harian yang lebih banyak berkorelasi dengan tekanan darah tinggi yang lebih tinggi. Selain itu, tekanan darah tinggi berkorelasi tinggi dengan tekanan darah rendah (0.97), menunjukkan bahwa orang dengan tekanan darah tinggi juga cenderung memiliki tekanan darah rendah.

Di sisi lain, terdapat beberapa hubungan negatif yang signifikan dalam matriks korelasi ini. Durasi tidur berkorelasi negatif dengan detak jantung (-0.52), menandakan bahwa orang yang tidur lebih lama cenderung memiliki detak jantung yang lebih rendah. Kualitas tidur berkorelasi negatif dengan detak jantung (-0.66), yang menunjukkan bahwa orang dengan kualitas tidur yang lebih baik cenderung memiliki detak jantung yang lebih rendah. Usia berkorelasi negatif dengan langkah harian (-0.23), menandakan bahwa orang yang lebih tua cenderung memiliki jumlah langkah harian yang lebih sedikit.

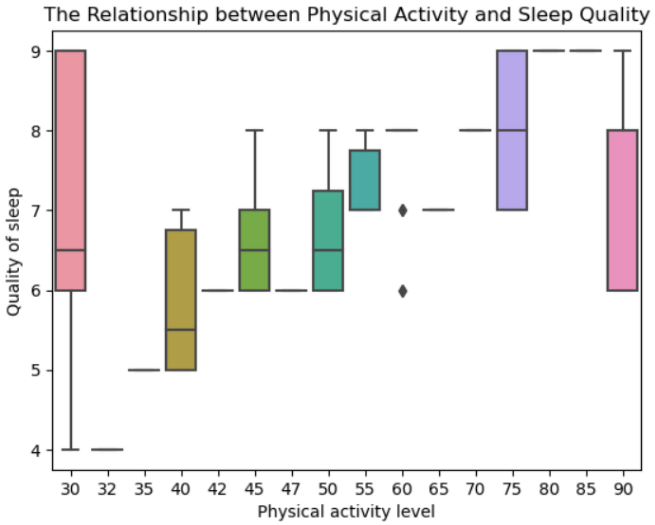
2. Korelasi antara variabel *sleep duration* dan *sleep quality*



Gambar 4 Hasil korelasi variabel *sleep duration* dan *sleep quality*

Berdasarkan output korelasi pada Gambar 4. terdapat scatter plot yang menunjukkan hubungan positif yang kuat antara *sleep duration* dan *sleep quality* yang ditunjukkan oleh nilai korelasi sebesar 0,88, yang artinya bahwa orang yang tidur lebih lama cenderung memiliki kualitas tidur yang lebih baik. Durasi tidur cukup penting untuk kesehatan dan kesejahteraan secara keseluruhan.

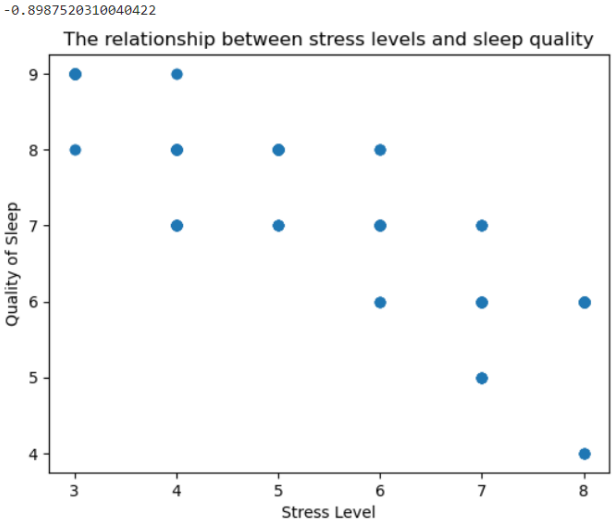
3. Korelasi antara variabel *physical activity* dan *sleep quality*



Gambar 5 Hasil korelasi variabel *physical activity* dan *sleep quality*

Hasil korelasi menunjukkan bahwa terdapat hubungan positif yang kuat antara aktivitas fisik dan kualitas tidur. Hal ini berarti bahwa orang yang lebih aktif secara fisik cenderung memiliki kualitas tidur yang lebih baik. Oleh karena itu, penting untuk meningkatkan aktivitas fisik Anda untuk mendapatkan tidur yang lebih nyenyak dan berkualitas.

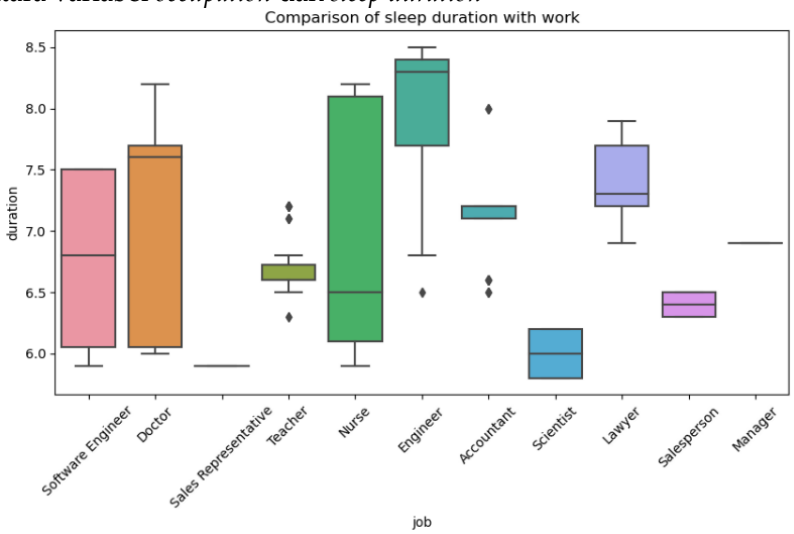
4. Korelasi antara variabel *stress level* dan *sleep quality*



Gambar 6 Hasil korelasi variabel *stress level* dan *sleep quality*

Berdasarkan output korelasi pada Gambar 6. terdapat scatter plot yang menunjukkan hubungan negatif yang kuat antara *stress level* dan *sleep quality* yang ditunjukkan oleh nilai korelasi sebesar -0,90, yang artinya bahwa seseorang yang mengalami tingkat stres yang lebih tinggi kemungkinan besar akan mengalami kualitas tidur yang lebih rendah.

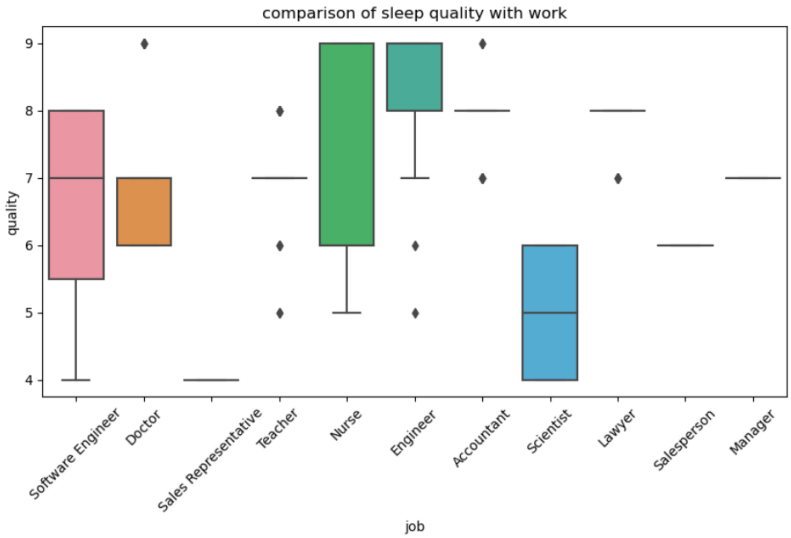
5. Perbandingan antara variabel *occupation* dan *sleep duration*



Gambar 7 Hasil perbandingan variabel *occupation* dan *sleep duration*

Berdasarkan output perbandingan pada Gambar 7 terdapat boxplot yang menunjukkan bahwa terdapat variasi yang signifikan dalam distribusi durasi tidur berdasarkan jenis pekerjaan. Secara umum, durasi tidur rata-rata untuk setiap pekerjaan cenderung lebih rendah daripada durasi tidur rata-rata secara keseluruhan. Boxplot menunjukkan bahwa median durasi tidur untuk setiap pekerjaan berada di bawah median durasi tidur total, yang ditandai dengan garis putus-putus di boxplot. Variabilitas durasi tidur antar pekerjaan juga cukup besar, seperti yang tercermin dari interquartile range yang bervariasi di setiap boxplot. Beberapa pekerjaan memiliki nilai outlier, menunjukkan adanya kasus ekstrem dengan durasi tidur yang jauh lebih rendah atau lebih tinggi dibandingkan dengan pekerjaan lainnya. Misalnya, pekerjaan Salesperson dan Manager terlihat memiliki nilai outlier. Secara keseluruhan, boxplot juga menunjukkan bahwa beberapa pekerjaan seperti Engineer, Accountant, Scientist, Lawyer, dan Salesperson memiliki durasi tidur rata-rata yang lebih tinggi dibandingkan dengan pekerjaan lainnya seperti Software Engineer, Doctor, Sales Representative, Teacher, dan Nurse.

6. Perbandingan antara variabel *occupation* dan *sleep quality*



Gambar 8 Hasil perbandingan variabel *occupation* dan *sleep quality*

Berdasarkan output perbandingan pada Gambar 8 terdapat boxplot yang menunjukkan bahwa kualitas tidur rata-rata untuk setiap jenis pekerjaan umumnya lebih rendah dibandingkan dengan kualitas tidur rata-rata secara keseluruhan. Hal ini terlihat dari posisi median kualitas tidur pada setiap boxplot yang berada di bawah garis median total (garis putus-putus). Variabilitas kualitas tidur antar pekerjaan juga bervariasi, seperti yang tercermin dari ukuran interquartile range yang berbeda-beda di setiap boxplot. Beberapa pekerjaan menunjukkan adanya nilai outlier, menandakan terdapat kasus ekstrem dengan kualitas tidur yang jauh lebih rendah atau lebih tinggi dibandingkan dengan pekerjaan lainnya, contohnya pada pekerjaan Salesperson dan Manager. Secara keseluruhan, boxplot juga menggambarkan bahwa beberapa pekerjaan seperti Engineer, Accountant, Scientist, Lawyer, dan Salesperson cenderung memiliki kualitas tidur rata-rata yang lebih tinggi dibandingkan dengan pekerjaan seperti Software Engineer, Doctor, Sales Representative, Teacher, dan Nurse.

C. *Pre-Processing Data*

Pada tahap pre-processing data, tim peneliti memeriksa kesesuaian tipe data setiap variabel. Semua variabel sudah memiliki tipe data yang sesuai. Selanjutnya, proses pembersihan data dilakukan dengan memeriksa duplikasi, *missing value*, data noise, outlier, dan keseimbangan data. Dalam dataset yang digunakan, format data semua variabel tidak terdapat duplikasi dan *missing value*. Namun, terdapat outlier pada variabel *heart rate*. Sehubungan dengan temuan yang menyatakan bahwa *heart rate* berhubungan dengan gangguan tidur. Maka, tim peneliti melakukan removal pada data outlier.

Selain itu, terdapat ketidakseimbangan pada data atau *imbalance*. Sehingga akan dilakukan pengecekan keseimbangan data dengan melihat proporsi pada masing-masing kategori variabel respon dengan hasil sebagai berikut.

Tabel 5 Pemeriksaan Keseimbangan Data

Kategori	Proporsi Data
0	20%
1	61%
2	19%

Dapat dilihat bahwa terjadi imbalance pada data. Sehingga tim peneliti akan menggunakan metode oversampling untuk mengatasi hal tersebut dan didapati bahwa data telah balance dengan proporsi 33,8%, 33,3%, dan 32,8% secara

berurutan.

Setelah melakukan tahapan di atas, akan dilakukan pengujian multikolinearitas menggunakan Variance Inflating Factor (VIF). Hasil pemeriksaan multikolinearitas yaitu sebagai berikut.

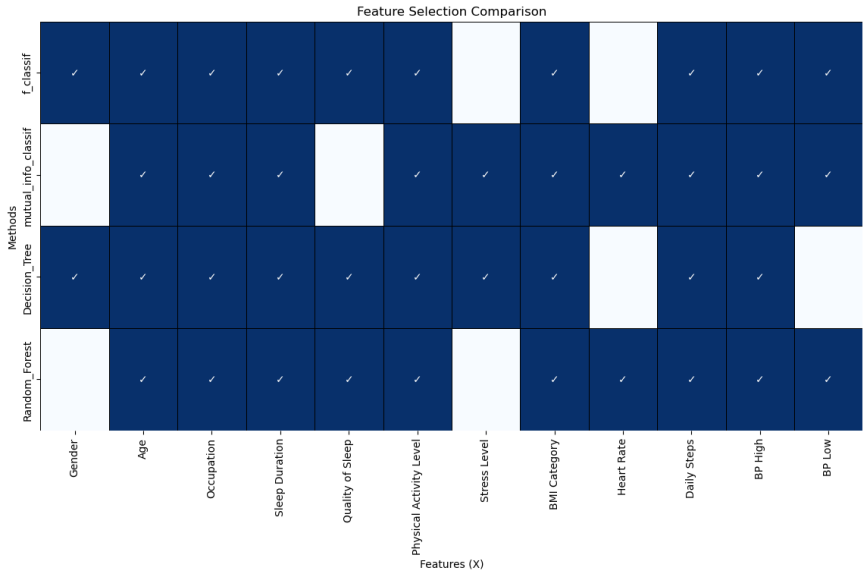
Tabel 6 Pemeriksaan Multikolinearitas

Variabel	Nilai VIF
X_1	68,35
X_4	11,17
X_5	8,18
X_7	6,23
X_8	8,15
X_9	68,62

Terlihat bahwa terdapat 3 variabel yang memiliki nilai VIF yang lebih dari 10. Hal ini menunjukkan bahwa variabel/fitur terindikasi kasus multikolinearitas. Untuk mengatasi hal tersebut, tim peneliti akan melakukan pemilihan fitur menggunakan beberapa metode dan didapati bahwa multikolinearitas telah teratasi dengan baik.

D. Pemilihan Fitur

Klasifikasi seseorang dengan gangguan tidur dilakukan menggunakan algoritma metode Decision Tree, KNN, Random Forest, dan SVM. Pada masing-masing metode tersebut dilakukan pemilihan fitur terbaik agar algoritma pemodelan berjalan lebih efisien. Berikut ini hasil pemilihan fitur terbaik.



Gambar 9 Hasil Seleksi Fitur

Dari hasil pemilihan fitur menggunakan beberapa metode pada Gambar 9 di atas, variabel yang akan digunakan dalam penelitian ini yaitu *age*, *occupation*, *sleep duration*, *quality of sleep*, *physical activity level*, *BMI category*, *heart rate*, *daily steps*, dan *BP Low*.

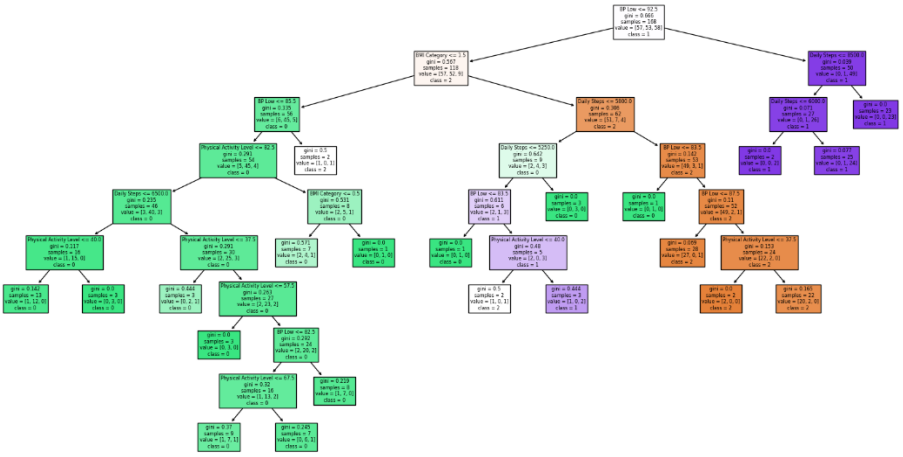
E. Decision Tree

Pada proses klasifikasi metode *Decision Tree* akan dilakukan beberapa tahapan yaitu inisialisasi model, menampilkan fitur penting dalam model, melakukan *hyperparameter tunning* dengan *Grid Search*, melatih ulang model dengan hyperparameter terbaik, hyperparameter tuning dengan *Randomized Search*, melatih ulang model dengan hyperparameter terbaik dari *Randomized Search*, evaluasi model pada dataset baru, *hyperparameter tunning* dengan *Randomized search* pada dataset baru, dan menampilkan kurva ROC pada masing-masing kategori data variabel respon. Dimana pada tunning parameter didapatkan parameter terbaik model sebagai berikut.

Tabel 7 Parameter Model Decision Tree

Parameter	Hasil
Criterion	Gini
Maximum depth	40
Minimum samples leaf	3
Minimum samples split	4

Melalui tahapan tersebut, didapatkan hasil *decision tree* pada Gambar 10 berikut.



Gambar 10 Decision Tree

Nilai Macro-Averaged AUC untuk metode *Decision Tree* adalah 0,91, yang menandakan bahwa model ini sangat baik dalam membedakan antara berbagai kelas secara keseluruhan. AUC yang mendekati 1 menunjukkan kemampuan diskriminatif model yang sangat tinggi, memberikan keyakinan lebih pada penggunaan model ini untuk klasifikasi gangguan tidur pada seseorang dalam dataset yang dianalisis. Berikut merupakan nilai AUC pada masing-masing kategori metode *decision tree.*

Tabel 8 Evaluasi AUC Metode *Decision Tree*

Kelas	Nilai AUC
0	0,87
1	0,92
2	0,93

F. Random Forest

Metode *random forest* merupakan algoritma *ensemble learning* yang menggunakan dan membangun struktur *Tree* dalam tahapannya. Untuk melakukan klasifikasi gangguan tidur seseorang dengan menggunakan metode *random forest*, tahap pertama yang dilakukan adalah melakukan *train test split* dengan perbandingan data *training* dan *testing* sebesar 80:20 persen dari keseluruhan data.

Pada proses klasifikasi metode *Random Forest* dilakukan uji coba kombinasi beberapa hyperparameter untuk memperoleh model dengan hyperparameter paling optimal. Kombinasi hyperparameter yang diuji coba yaitu jumlah pohon (n_estimator) sebanyak 50,100, dan 200. Jumlah sampel minimum yang dibutuhkan untuk memecah node internal(min_samples_split) terdiri dari 2, 5, dan 10. Serta, jumlah sampel minimum yang dibutuhkan untuk menjadi daun pada pohon (min_samples_leaf) terdiri dari 1, 2, dan 4. Pemodelan *random forest* dengan menggunakan hasil *feature importance* pada *random forest* didapatkan parameter terbaik model sebagai pada Tabel 9.

Tabel 9 Parameter Model Random Forest

Parameter	Hasil
Maximum depth	40
Maximum features	Log2
Minimum samples leaf	1
Minimum samples split	5
n estimators	192

Tahapan selanjutnya adalah mengevaluasi kebaikan model *random forest* dengan 4 fitur sederhana berdasarkan skor MDI dan parameter yang telah ditentukan berdasarkan hasil *GridSearch* menggunakan *classification report* sebagaimana Tabel 10 berikut.

Tabel 10 Classification Report

Parameter	Precision	Recall	F1-Score	Support
0	0,98	0,79	0,88	14
1	0,85	0,98	0,92	17
2	0,91	0,91	0,91	11
Accuracy			0,90	42
Macro avg	0,92	0,90	0,90	42
Weighted avg	0,92	0,90	0,90	42

Nilai Macro-Averaged AUC untuk metode *Random Forest* adalah 0,90, yang menandakan bahwa model ini sangat baik dalam membedakan antara berbagai kelas secara keseluruhan. AUC yang mendekati 1 menunjukkan kemampuan diskriminatif model yang sangat tinggi, memberikan keyakinan lebih pada penggunaan model ini untuk klasifikasi gangguan tidur pada seseorang dalam dataset yang dianalisis.

Tabel 11 Evaluasi AUC Metode *Random Forest*

Kelas	Nilai AUC
0	0,88
1	0,92
2	0,90

G. K-Nearest Neighbors (KNN)

Pemodelan dengan metode K-Nearest Neighbors dilakukan dengan inisialisasi dan pelatihan model, prediksi dan evaluasi, *confusion matrix* dan *classification* report, serta hyperparameter tuning dengan *Grid Search*. Dengan tahapan tersebut, telah didapatkan parameter terbaik dengan k sebesar 5. Dengan nilai k tersebut, didapatkan nilai akurasi pada data train dan test sebagai berikut.

Tabel 12 Nilai Akurasi Model KNN

Data	Akurasi
Train	89,29%
Test	90,48%

Nilai Macro-Averaged AUC untuk metode KNN adalah 0,91, yang menandakan bahwa model ini sangat baik dalam membedakan antara berbagai kelas secara keseluruhan. AUC yang mendekati 1 menunjukkan kemampuan diskriminatif model yang sangat tinggi, memberikan keyakinan lebih pada penggunaan model ini untuk klasifikasi gangguan tidur pada seseorang dalam dataset yang dianalisis.

Tabel 13 Evaluasi AUC Metode KNN

Kelas	Nilai AUC
0	0,88
1	0,93
2	0,93

H. Support Vector Machine

Pemodelan dengan metode *Support Vector Mechine* (SVM) dilakukan dengan memeriksa pada dua jenis kernel yaitu kernel linear dan RBF. Kemudian dilakukan tuning parameter menggunakan GridSearchCV untuk mencari kombinasi hyperparameter terbaik (C, gamma, kernel) untuk model SVM dengan kernel 'rbf'. Setelah dilakukan uji coba, paramater terbaik yang didapatkan adalah sebagai berikut.

Tabel 14 Parameter Model SVM

Parameter	Hasil
C	1
Gamma	auto
Kernel	rbf

Selain itu, juga dilakukan pemodelan menggunakan kernel linear. Dari kedua kernel tersebut, berikut merupakan hasil perbandingan akurasi antar keduanya.

Tabel 15 Perbandingan Akurasi

Kernel	Akurasi
Linear	0,7619
rbf	0,9048

Dari tabel tersebut dapat dilihat bahwa akurasi terbesar ada pada kernel rbf, sehingga dalam metode SVM ini akan digunakan kernel tersebut. Nilai Macro-Averaged AUC untuk metode SVM adalah 0,92, yang menandakan bahwa model ini sangat baik dalam membedakan antara berbagai kelas secara keseluruhan. AUC yang mendekati 1 menunjukkan kemampuan diskriminatif model yang sangat tinggi, memberikan keyakinan lebih pada penggunaan model ini untuk klasifikasi gangguan tidur pada seseorang dalam dataset yang dianalisis.

Tabel 16 Evaluasi AUC Metode SVM

Kelas	Nilai AUC
0	0,90
1	0,93
2	0,90

I. Hasil Perbandingan

Setelah dilakukan pengujian satu per satu dari tiap algoritma kemudian hasil yang didapatkan dibandingkan untuk menentukan algoritma yang cocok untuk mengklasifikasikan gangguan tidur seseorang. Berdasarkan hasil analisis yang telah dilakukan, hasil perbandingan dari keempat metode algoritma yaitu sebagaimana pada Tabel 17.

Tabel 17 Hasil Perbandingan

Metode	Akurasi	AUC
Decision Tree	0,9048	0,91
Random Forest	0,9048	0,90
K-Nearest Neighbors	0,8929	0,91
Support Vector Machine	0,9048	0,92

Terlihat bahwa seluruh metode memiliki nilai akurasi dan AUC yang mirip dengan kategori sangat baik. Namun, klasifikasi metode SVM lebih tinggi dibandingkan dengan hasil klasifikasi lainnya. Berdasarkan Tabel diatas, metode SVM dengan parameter C sebesar 1, gamma adalah auto, dan dengan kernel rbf mampu mengklasifikasikan gangguan tidur seseorang dengan skor AUC sebesar 0,92 yang masuk dalam kategori *excellent classification*.

V. KESIMPULAN DAN SARAN

Berdasarkan hasil analisis dengan membandingkan empat algoritma *machine learning* yaitu *Decision Tree*, *Random Forest*, *K-Nearest Neighbors (KNN)*, dan *Support Vector Machine* (SVM) yang dibangun dari parameter optimal dan pemilihan fitur terbaik mampu mengklasifikasikan gangguan tidur seseorang dengan kategori *excellent classification*. Diperoleh hasil *Decision Tree* dengan nilai akurasi 0,905 dan nilai AUC 0,91. Untuk *Random Forest* mendapatkan hasil akurasi 0,905 dan nilai AUC sebesar 0,90. Kemudian *K-Nearest Neighbors* (KNN) mendapatkan hasil akurasi 0,893 dan nilai AUC sebesar 0,91 dan *Support Vector Machine* (SVM) mendapatkan hasil akurasi sebesar 0,905 dan nilai AUC sebesar 0,92. Performa metode *Support Vector Machine* (SVM) untuk mengklasifikasikan gangguan tidur seseorang melalui tahap latih dan tuning parameter menghasilkan skor AUC tertinggi sebesar 0,92 . Nilai AUC tersebut lebih baik dari metode *Decision Tree*, *Random Forest*, dan *K-Nearest Neighbors* (KNN).

Berdasarkan temuan penelitian ini, disarankan agar peneliti lain mempertimbangkan penggunaan algoritma *machine learning* tambahan seperti *AdaBoost* atau *LightGBM* untuk mengevaluasi apakah mereka dapat memberikan hasil yang lebih baik atau lebih efisien. Selain itu, penggunaan teknik penyeimbangan data lainnya seperti *Random Under Sampling* atau *ADASYN* dapat membantu meningkatkan kinerja model pada dataset yang tidak seimbang. Peneliti juga dianjurkan untuk memperluas pencarian *hyperparameter* dan menggunakan teknik optimasi seperti *Grid Search* atau *Random* Search guna menemukan parameter yang lebih optimal. Uji validasi dengan dataset yang berbeda atau lebih besar juga penting untuk memastikan bahwa model dapat digeneralisasi. Akhirnya, penerapan teknik visualisasi hasil seperti t-SNE atau PCA dapat memberikan wawasan lebih jelas tentang distribusi data dan kinerja model, yang sangat berguna untuk analisis lanjutan dan aplikasi praktis.

REFERENSI

[1] A. Andriani, "Sistem Pendukung Keputusan Berbasis Decision Tree Dalam Pemberian Beasiswa (Studi Kasus: AMIK 'BSI Yogyakarta')," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2013, no. Sentika, pp. 163–168, 2013.

[2] A. C. Handoko and H. Hendry, "Perbandingan Metode Supervised Learning Untuk Prediksi Diabetes Gestasional Dengan Software Orange," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, no. 4, pp. 1238–1247, 2023, doi: 10.29100/jipi.v8i4.4166.

[3] A. Fitria, Muslim, and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," vol. 3, no. 2, pp. 102–106, 2018.

[4] B. Santoso, "An Analysis of Spam Email Detection Performance Assessment Using Machine Learning," *J. Online Inform.*, vol. 4, no. 1, p. 53, 2019, doi: 10.15575/join.v4i1.298.

[5] D. Sari, "Prediksi Gangguan Tidur pada Sleep Health and Lifestyle Menggunakan Support Vector Machine dan Neural Network," *Jav. J. Vokasi Inform.*, pp. 36–42, 2024, doi: 10.24036/javit.v4i1.168.

[6] F. Andiarna, L. P. Widayanti, I. Hidayati, E. Agustina, and K. Kunci, "Analisis Penggunaan Media Sosial Terhadap Kejadian Insomnia Pada Mahasiswa Analysis Of Social Media Usage With Insomnia Incidence Among Students," 2020.

[7] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874

[8] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, "Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[9] I. A. A. Angreni, S. A. Adisasmita, and M. I. Ramli, "Pengaruh Nilai K Pada Metode KNEAREST NEIGHBOR (KNN) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan," vol. 7, no. 2, pp. 63–70, 2018.

[10] Kusumaningrum, A. P. (2017) 'Optimasi Parameter Support Vector Machine Menggunakan Genetic Algorithm Untuk Klasifikasi Microarray Data', ITS Repository.

[11] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *J. Appl. Informatics Comput.*, vol. 5, no. 2, pp. 103–108, 2021, doi: 10.30871/jaic.v5i2.3200.
[12] M. B. Arya Darmawan, F. Dewanta, and S. Astuti, "Analisis Perbandingan Algoritma Decision Tree, Random Forest, dan Naïve Bayes untuk Prediksi Banjir di Desa Dayeuhkolot," *TELKA - Telekomun. Elektron. Komputasi dan Kontrol*, vol. 9, no. 1, pp. 52–61, 2023, doi: 10.15575/telka.v9n1.52-61.
[13] M. Anastasia, V. S. Maulivia, and S. Suharjito, "Metode Pembelajaran Mesin Untuk Menilai Data Depresi Dan Kesehatan Mental," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 7, no. 3, pp. 606–612, 2024, doi: 10.31539/intecoms.v7i3.9584.
[14] M. M. Baharuddin, H. Azis, and T. Hasanuddin, "Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 3, pp. 269–274, 2019, doi: 10.33096/ilkom.v11i3.489.269-274.
[15] N. Hadianto, H. B. Novitasari, and A. Rahmawati, "Klasifikasi Peminjaman Nasabah Bank Menggunakan Metode Neural Network," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 163–170, 2019, doi: 10.33480/pilar.v15i2.658.
[16] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," J. Inform. Univ. Pamulang, vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.
[17] S. Styawati, N. Hendrastuty, and A. R. Isnain, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," *J. Inform. J. Pengemb. IT*, vol. 6, no. 3, pp. 150–155, 2021, doi: 10.30591/jpit.v6i3.2870.
[18] Suryanto, E. and Purnami, S. W. (2015) 'Perbandingan Reduced Support Vector Machine dan Smooth Support Vector Machine untuk Klasifikasi Large Data', Jurnal Sains dan Seni ITS, 4(1), pp. D25–D30.
[19] W. E. Radityo, "Depresi dan Gangguan Tidur," *E-Jurnal Med. Udayana* , vol. 1 (1), pp. 1–16, 2012, [Online]. Available: https://ojs.unud.ac.id/index.php/eum/article/view/4267
[20] Z. Annisa and B. S. S. Ulama, "Analisis Sentimen Data Ulasan Pengguna Aplikasi 'PeduliLindungi' pada Google Play Store Menggunakan Metode Naïve Bayes Classifier Model Multinomial," *J. Sains dan Seni ITS*, vol. 11, no. 6, 2023, doi: 10.12962/j23373520.v11i6.94064.