PENAMBANGAN DATA

LAPORAN TUGAS BESAR (UAS)

Penerapakan Algoritma K-NN Pada Data Permasalahan Tidur

Disusun Oleh:

10121167 - Ahmad Jaenal Aripin

IF-5

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNIK & ILMU KOMPUTER

UNIVERSITAS KOMPUTER INDONESIA

2023

PENDAHULUAN

Latar Belakang

Tidur merupakan bagian penting dalam menjaga kesehatan dan kesejahteraan manusia. Namun, banyak individu mengalami kesulitan dalam mencapai tidur yang berkualitas atau menjaga pola tidur yang konsisten. Masalah tidur, atau yang dikenal sebagai gangguan tidur atau sleep disorder, menjadi salah satu masalah kesehatan yang signifikan di seluruh dunia.

Berbagai faktor dapat menyebabkan gangguan tidur, mulai dari stres dan kecemasan hingga kondisi kesehatan fisik yang mendasarinya. Dampak dari kurang tidur atau tidur yang tidak berkualitas dapat meliputi penurunan konsentrasi, peningkatan risiko kecelakaan, gangguan suasana hati, dan penurunan produktivitas. Jangka panjangnya, gangguan tidur juga dapat meningkatkan risiko terkena penyakit jantung, diabetes, obesitas, serta gangguan kesehatan mental.

Gangguan tidur tidak hanya memengaruhi individu secara personal, tetapi juga berdampak pada tingkat keberhasilan dan kesejahteraan sosial secara keseluruhan. Dalam konteks ini, pemahaman mendalam tentang penyebab dan penanganan masalah tidur menjadi krusial untuk meningkatkan kualitas hidup dan kesehatan masyarakat secara keseluruhan.

Dalam laporan ini, akan dikaji berbagai aspek yang berkaitan dengan permasalahan tidur, termasuk faktor penyebabnya. Dengan memahami permasalahan tidur secara komprehensif, diharapkan dapat diidentifikasi langkah-langkah yang efektif dalam meminimalkan dampak negatifnya serta meningkatkan kualitas tidur dan kesejahteraan secara keseluruhan.

Analisis Masalah

Sleep disorder atau gangguan tidur merupakan permasalahan kesehatan yang kompleks yang dapat memengaruhi kualitas hidup seseorang secara signifikan. Dalam konteks analisis ini, saya akan mengeksplorasi tiga kategori target yang menjadi fokus: none (tidak memiliki gangguan tidur), insomnia dan sleep apnea. Tujuan analisis ini adalah untuk memahami keterkaitan penyebab yang mungkin terkait dengan masing-masing kategori gangguan tidur tersebut.

1. None (Tidak Ada Gangguan Tidur):

- Kategori ini adalah kategori paling bagus diantara insomnia atau sleep apnea. Oleh karena itu akan sangat berguna untuk menganalisis apa yang menjadi penyebab dari seseorang tidak memiliki permasalahan tidur berdasarkan data yang diperoleh
- Penyebabnya bisa berkisar dari pola hidup yang sehat, pengelolaan stres yang efektif, hingga lingkungan tidur yang nyaman dan tenang.

2. Insomnia:

- Insomnia adalah gangguan tidur yang ditandai dengan kesulitan memulai tidur, mempertahankan tidur, atau tidur yang tidak cukup berkualitas.
- Faktor psikologis seperti stres, kecemasan, dan depresi seringkali berperan dalam terjadinya insomnia.
- Selain itu, gaya hidup yang tidak sehat seperti pola tidur yang tidak teratur, konsumsi kafein, serta penggunaan gadget sebelum tidur juga dapat menjadi penyebab insomnia.

3. Sleep Apnea:

- Sleep apnea merupakan gangguan tidur serius yang ditandai dengan gangguan pernapasan yang terjadi selama tidur.
- Faktor-faktor risiko sleep apnea meliputi obesitas, konsumsi alkohol dan obat-obatan tertentu.

PEMBAHASAN DAN HASIL

Algoritma yang akan digunakan pada kasus ini adalah K-NN. K-Nearest Neighbor (KNN) adalah suatu metode yang menggunakan algoritma supervised dimana hasil dari query instance yang baru diklasifikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma KNN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples. Dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN.

Tahapan-tahapan data mining sebagai berikut:

Ket: Semua tabel data yang ada di laporan ini adalah table excel jadi dapat dilihat untuk detail dari semua datanya.

1. Selection
   Pada tahap awal ini saya memiliki data 14 fitur dimana dari ke 14 fitur ini saya akan melakukan selection fitur untuk melihat fitur-fitur mana saja yang relevan terhadap kasusnya (targe/class). Untuk data awal sebagai berikut:

Person ID	Gender	Age	Occupatio	Sleep Dura	Quality of	Physical Ac	Stress Leve	BMI Categ
1	Male	27	Software E	6.1	6	42	6	Overweigh
2	Male	28	Doctor	6.2	6	60	8	Normal
3	Male	28	Sales Repr	5.9	4	30	8	Obese
4	Male	28	Software E	5.9	4	30	8	Obese
5	Male	29	Teacher	6.3	6	40	7	Obese
6	Male	29	Doctor	7.8	7	75	6	Normal
7	Male	29	Doctor	6.1	6	30	8	Normal
8	Male	29	Doctor	6.0	6	30	8	Normal
9	Female	29	Nurse	6.5	5	40	7	Normal W

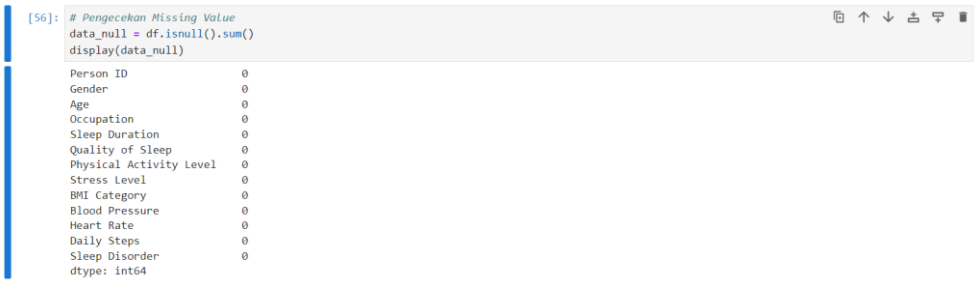
   Dari ke 14 fitur saya akan menghapus fitur Person ID karena fitur ini tidak relevan terhadap target (class) yang akan di proses di tahap selanjutnya. Jadi data yang akan digunakan pada tahap selanjutnya sebagai berikut:

Gender	Gender	Age	Occupatio	Sleep Dura	Quality of	Physical Ac	Stress Leve	BMI Categ
Male	Male	27	Software E	6.1	6	42	6	Overweigh
Male	Male	28	Doctor	6.2	6	60	8	Normal
Male	Male	28	Sales Repr	5.9	4	30	8	Obese
Male	Male	28	Software E	5.9	4	30	8	Obese
Male	Male	29	Teacher	6.3	6	40	7	Obese
Male	Male	29	Doctor	7.8	7	75	6	Normal
Male	Male	29	Doctor	6.1	6	30	8	Normal
Male	Male	29	Doctor	6.0	6	30	8	Normal
Female	Female	29	Nurse	6.5	5	40	7	Normal W

   Jadi hasil akhir pada tahap selection total fitur yang akan masuk ke tahap pre-processing adalah 13 fitur.

2. Pre-processing
   Pada tahap ini saya akan mengolah data dari tahap sebelumnya. Disini saya akan melakukan pengecekan nilai missing value dan merubah format data yang masih belum sesuai.

Untuk missing value dapat dibantu menggunakan python. Hasil dari pengecekan missing value dapat dilihat pada gambar 1 berikut:



```
[56]: # Pengecekan Missing Value
data_null = df.isnull().sum()
display(data_null)
```

```
Person ID                  0
Gender                     0
Age                        0
Occupation                 0
Sleep Duration             0
Quality of Sleep           0
Physical Activity Level    0
Stress Level               0
BMI Category               0
Blood Pressure             0
Heart Rate                 0
Daily Steps                0
Sleep Disorder             0
dtype: int64
```

Gambar 1 Hasil pengecekan missing value

Selanjutkan saya akan merubah format untuk fitur dari Blood Pressure. Blood Pressure itu sendiri memiliki dua nilai seperti 126/83, karena itu saya akan membagi menjadi dua fitur yang berbeda yaitu Systolic dan Diastolic. Selain itu saya juga akan membuat kategori pada fitur BMI Category yang awalnya terdapat Normal, Normal Weight, Overwieght dan Obese, menjadi Normal, Overwieght dan Obese, karena pada dasarnya Normal dan Normal Weight saya asumsikan adalah kategori yang sama.

Berikut adalah data hasil pengecekan missing value dan perubahan fitur dan data sebagai berikut:

Gender	Age	Occupatio	Sleep Dura	Quality of	Physical Ac	Stress Leve	BMI Categ
Male	27	Software E	6,1	6	42	6	Overweigh
Male	28	Doctor	6,2	6	60	8	Normal
Male	28	Sales Repr	5,9	4	30	8	Obese
Male	28	Software E	5,9	4	30	8	Obese
Male	29	Teacher	6,3	6	40	7	Obese
Male	29	Doctor	7,8	7	75	6	Normal
Male	29	Doctor	6,1	6	30	8	Normal
Male	29	Doctor	6	6	30	8	Normal
Female	29	Nurse	6,5	5	40	7	Normal W

3. Transformation

Pada tahap ini dilakukan binerisasi terhadap beberapa fitur yang bersifat categorical menjadi numeric diantaranya fitur Gender, Occupation, BMI Category. Selain binerisasi saya juga melakukan normalisasi untuk fitur Daily Steps agar membuat data nya lebih konsisten dengan menggunakan rumus min-max [0,1].

Binerisasi terhadap data categorical:

Gender	
Female	0
Male	1

Occupation	
Doctor	1
Nurse	2
Sales Representative	3
Software Engineer	4
Teacher	5
Engineer	6
Accountant	7

BMI Category	
Normal / Normal Weight	1
Overweight	2
Obese	3

Sedangkan untuk normalisasi fitur Daily Steps dengan min-max [0,1] sebagai berikut :
min(x1) = 3000, max(x1) = 10000

$$\hat{x}_{ik} = \frac{x_{ik} - \min(x_k)}{\max(x_k) - \min(x_k)}$$

Gambar 2 Rumus min-max

Pada gambar 2 adalah rumus yang akan digunakan pada tahap normalisasi min-max[0,1].

Berikut adalah data hasil binerisasi dan normalisasi:

Gender	Age	Occupatio	Sleep Dura	Quality of	Physical Ac	Stress Leve	BMI Categ	Systolic
1	27	4	6,1	6	42	6	2	126
1	28	1	6,2	6	60	8	1	125
1	28	3	5,9	4	30	8	3	140
1	28	4	5,9	4	30	8	3	140
1	29	5	6,3	6	40	7	3	140
1	29	1	7,8	7	75	6	1	120
1	29	1	6,1	6	30	8	1	120
1	29	1	6	6	30	8	1	120
0	29	2	6,5	5	40	7	1	132

Data tersebut adalah data yang akan masuk ke proses utama yaitu penerapan algoritma K-NN.

4. Data Mining
Bagian ini adalah proses penerapan algoritma K-Nearest Neighbor (K-NN) pada dataset hasil pengolahan dari proses sebelumnya. Berikut adalah langkah-langkahnya:

Langkah 1:
Menentukan Nilai parameter K = 5

Langkah 2:
Menghitung kuadrak jarak dengan rumus Euclidean Distance.

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{r}(x_i - y_i)^2}$$

Gambar 3

Rumus Euclidean Distance dapat dilihat pada gambar 3 dan hasil dari perhitungannya adalah sebagai berikut:

Physical Ac	Stress Leve	BMI Categ	Systolic	Diastolic	Heart Rate	Daily Steps	Sleep Diso	Euclidean I
42	6	2	126	83	77	0,171429	None	33,36622
60	8	1	125	80	75	1	None	15,30103
30	8	3	140	90	85	0	None	49,31045
30	8	3	140	90	85	0	Sleep Apne	49,36112
40	7	3	140	90	82	0,071429	Sleep Apne	39,95627
75	6	1	120	80	70	0,714286	Insomnia	8,845338
30	8	1	120	80	70	0,714286	Insomnia	45,85859
30	8	1	120	80	70	0,714286	None	45,85848
40	7	1	132	87	80	0,142857	None	36,55922

Berikut juga adalah data uji:

Quality of	Physical Ac	Stress Leve	BMI Categ	Systolic	Diastolic	Heart Rate	Daily Steps	Sleep Diso
6	75	6	1	125	80	77	0,714286	?

Langkah 3:

Selanjutnya adalah saat melakukan pengurutan berdasarkan data Euclidean Distance secara ascending.

Physical Ac	Stress Leve	BMI Categ	Systolic	Diastolic	Heart Rate	Daily Steps	Sleep Diso	Euclidean I
75	6	1	120	80	70	0,714286	Insomnia	8,845338
75	6	1	120	80	70	0,714286	None	8,863408
75	6	1	120	80	70	0,714286	None	8,882004
75	6	1	120	80	70	0,714286	None	8,901685
75	6	1	120	80	70	0,714286	None	8,922444
75	6	1	120	80	70	0,714286	None	9,031058
75	6	1	120	80	70	0,714286	None	9,049309
75	6	1	120	80	70	0,714286	None	9,049309
75	6	1	120	80	70	0,714286	None	9,068627

Langkah 4:

Mengelompokan baris data latih yang termasuk tetangga berdasarkan nilai k yang sudah ditentukan diawal dan hasilnya sebagai berikut:

Physical Ac	Stress Leve	BMI Categ	Systolic	Diastolic	Heart Rate	Daily Steps	Sleep Diso	idean Dista
75	6	1	120	80	70	0,714286	Insomnia	8,845338
75	6	1	120	80	70	0,714286	None	8,863408
75	6	1	120	80	70	0,714286	None	8,882004
75	6	1	120	80	70	0,714286	None	8,901685
75	6	1	120	80	70	0,714286	None	8,922444

Langkah 5:

Terakhir adalah melihat klasifikasi nearest neighbor yang paling mayoritas.

Physical A	Stress Leve	BMI Categ	Systolic	Diastolic	Heart Rate	Daily Steps	Sleep Diso	idean Dista
75	6	1	120	80	70	0,714286	Insomnia	8,845338
75	6	1	120	80	70	0,714286	None	8,863408
75	6	1	120	80	70	0,714286	None	8,882004
75	6	1	120	80	70	0,714286	None	8,901685
75	6	1	120	80	70	0,714286	None	8,922444
75	6	1	125	80	77	0,714286	None	-

Jumlah label “None” > Jumlah “Insomnia” dan “Sleep Apnea”.

Jadi kesimpulannya data yang akan diprediksi adalah “None”

5. Evaluasi dan Hasil
   Dari proses yang telah dilalui maka untuk hasil dari data yang telah diolah maka bisa disimpulkan bahwa kategori nilai untuk data test adalah “None”.

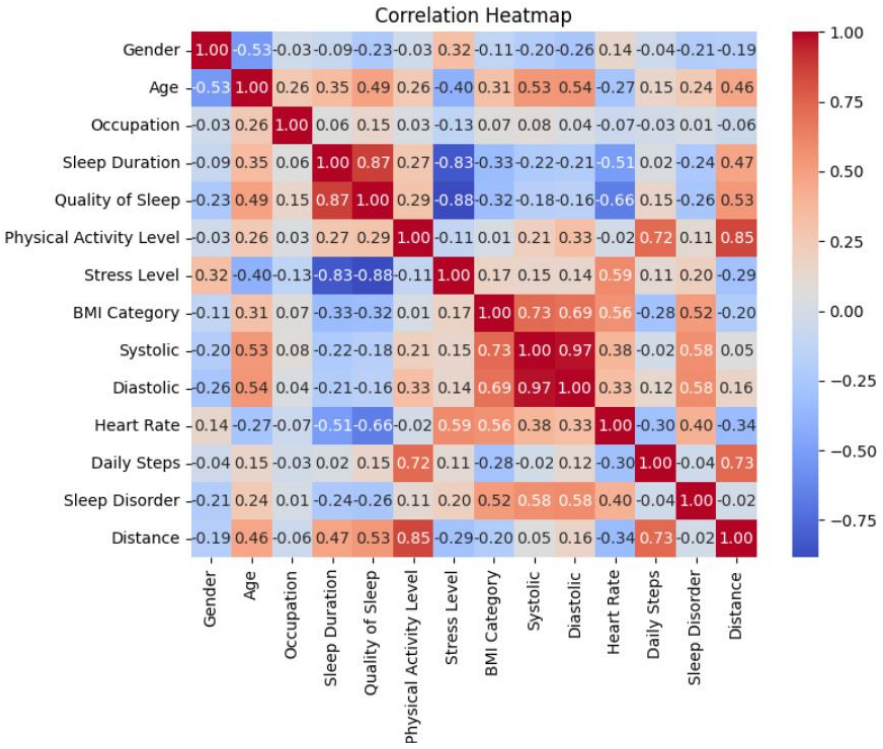
Selain melakukan klasifikasi K-NN secara manual, saya juga melakukan penerapan klasifikasi K-NN dengan python menggunakan library sklearn. Dengan menggunakan data sebanyak 374 data didapat hasil accuracy 0.93 atau 93%.

	precision	recall	f1-score	support
0	0.95	0.95	0.95	21
1	0.83	1.00	0.91	5
2	1.00	0.50	0.67	2
accuracy			0.93	28
macro avg	0.93	0.82	0.84	28
weighted avg	0.93	0.93	0.92	28

Gambar 4 Classification Report

Dengan menggunakan python untuk melakukan penerapan K-NN, semua tahapan sesuai dengan proses manual yang sebelumnya mulai dari selection sampai evaluasi hanya saja menggunakan python sebagai tools.

Supaya mempermudah dalam proses pengambilan pengetahuan saya melakukan proses visualisasi data sebagai berikut:

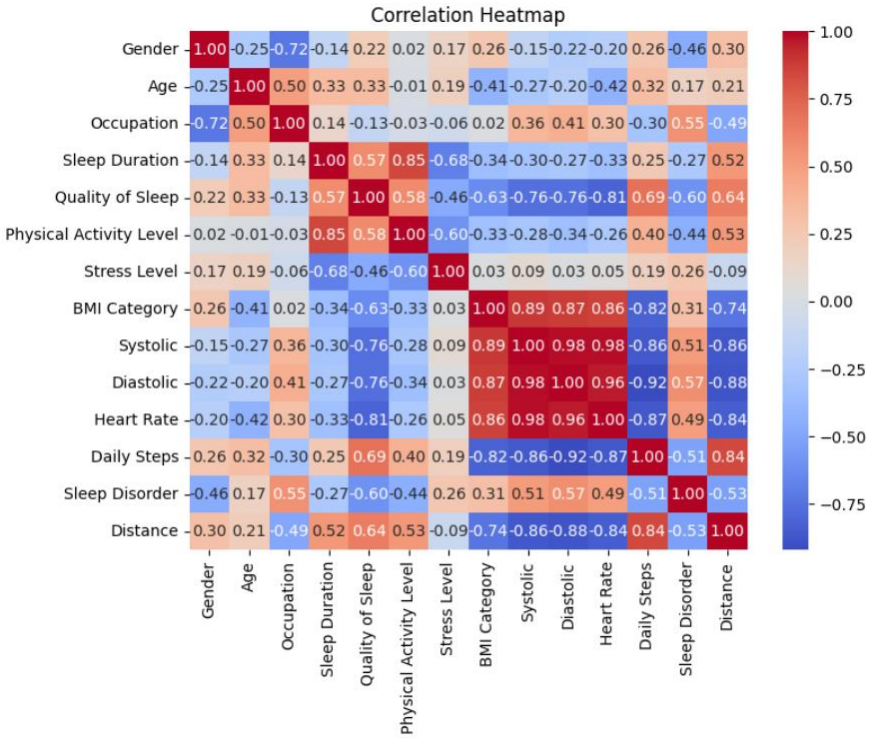


Gambar 5 Heatmap pada dataset

Dari tampilan visualisasi saya menemukan sebuah pola fitur yang saling berkolerasi terhadap class/target yaitu sebagai berikut:

- Jika fitur Stress Level, BMI Category, Systolic, Diastolic dan Heart Rate nilainya semakin tinggi maka akan mengakibatkan kemungkinan seseorang terkena sleep disorder baik itu insomnia ataupun sleep apnea.
- Sebalikanya jika fitur Sleep Duration, Quality of Sleep dan Physical Acitivity Levelnya semakin tinggi maka seseorang akan semakin kecil resikonya untuk terkena sleep disorder.

Pola akan semakin terlihat jelas pada gambar 6 berikut ini:



Gambar 6 Heatmap pada dataset