BAB 4
HASIL PENELITIAN DAN PEMBAHASAN

4.1 Perbandingan Penggunaan Pembagian Dataset

Tujuan dari data-data yang disajikan pada Tabel 4.1 ini adalah untuk melihat mana dari pembagian *dataset* yang paling baik, untuk kita gunakan dalam penelitian ini.

Tabel 4.1 – Perbedaan Kinerja yang dihasilkan dari penggunaan Pembagian Dataset. Nilai-nilai tersaji dalam persentase

		(% TRAINING) : (% TESTING)				
		50 : 50	**60 : 40**	**70 : 30**	**80 : 20**	**90 : 10**
AKURASI	**kNN**	89.71	87.40	91.96	92.63	96.84
	Random Forest	94.96	94.75	95.45	95.26	95.79
	Deep Learning	93.70	92.13	93.36	91.58	94.74
PRESISI	**kNN**	96.70	84.71	96.72	95.35	96.15
	Random Forest	92.97	91.43	95.89	89.29	92.59
	Deep Learning	94.07	83.48	88.61	83.64	89.29
RECALL	**kNN**	65.67	67.29	73.75	77.36	92.59
	Random Forest	88.81	89.72	87.50	94.34	92.59
	Deep Learning	82.84	89.72	87.50	86.79	92.59
AUC	**kNN**	0.947	0.932	0.965	0.958	0.978
	Random Forest	0.981	0.980	0.989	0.985	0.988
	Deep Learning	0.965	0.961	0.974	0.958	0.978

Dari kinerja yang dihasilkan dari ketiga algoritma pada tabel di atas, terlihat bahwa pembagian 90% *data training* dan 10% *data testing* adalah pembagian yang paling baik (angka-angka yang diwarnai merah).

4.2 Evaluasi Kinerja kNN, RF, DL

Selanjutnya, data evaluasi yang disajikan dalam Tabel 4.2 mencakup metrik kinerja tiga algoritma klasifikasi yang berbeda: kNN, Random Forest, dan Deep Learning. Mari kita lihat perbandingan kinerja dari ketiganya:

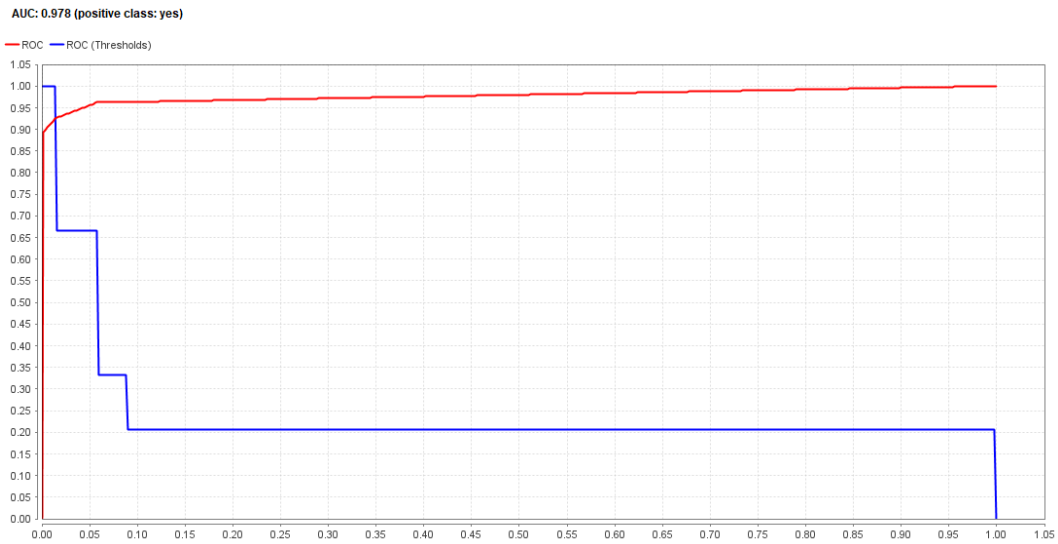
Tabel 4.2 – Hasil kinerja dari kNN, RF, dan DL sebelum dioptimasi PSO. Nilai akurasi, presisi, dan *recall* dalam persentase

	AKURASI	**PRESISI**	***RECALL***	**AUC**
kNN	96.84	96.15	92.59	0.978
Random Forest	95.79	92.59	92.59	0.988
Deep Learning	93.68	95.65	81.48	0.969

kNN memiliki akurasi tertinggi, diikuti oleh Random Forest dan Deep Learning. Akurasi mencerminkan sejauh mana algoritma kNN mampu mengklasifikasikan data secara benar secara keseluruhan.

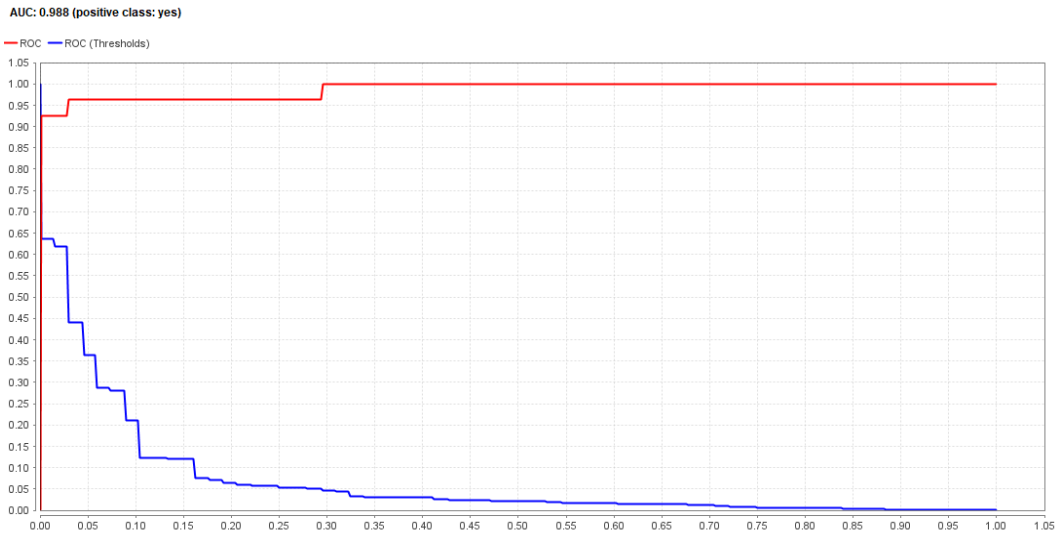
Presisi mengukur proporsi positif yang diprediksi benar-benar positif, memberikan gambaran seberapa baik algoritma dalam mengidentifikasi kelas tertentu tanpa memberikan banyak *false positive*. Meskipun presisi RF ini lebih rendah dibandingkan dua yang lainnya, nilai ini masih termasuk tinggi –ini menunjukkan kemungkinan adanya *false positive* yang lebih tinggi.

Recall (sensitivitas) mengukur sejauh mana algoritma dapat mengidentifikasi semua *instance* dari kelas positif, memberikan pandangan tentang seberapa baik algoritma dapat menangkap kasus positif. Di sini RF memiliki nilai yang sama dengan kNN. Sementara DL lebih rendah dibandingkan kNN dan Random Forest, yang menunjukkan bahwa DL mungkin melewatkan beberapa kasus positif.



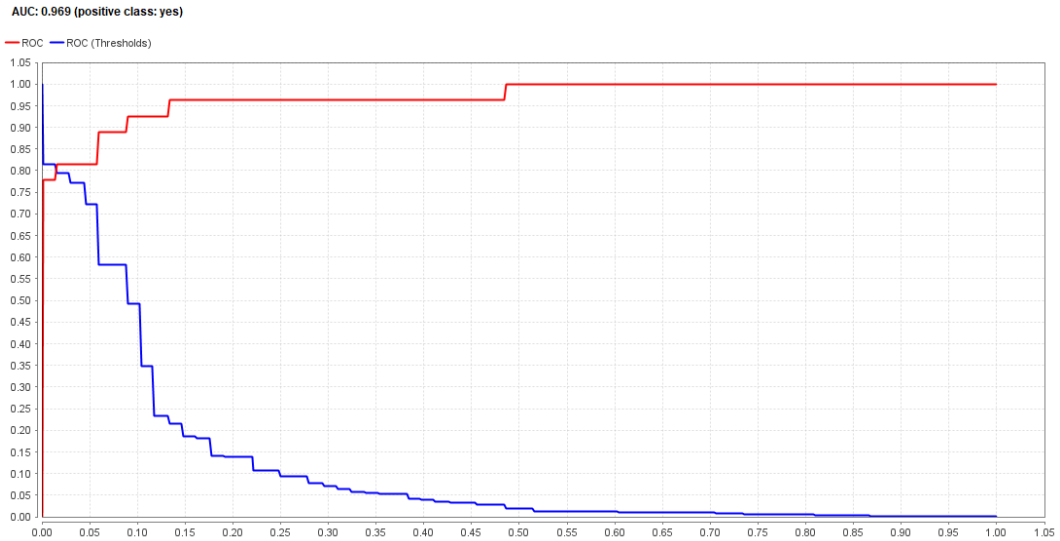
Gambar 4.1 – Grafik AUC dari kNN sebelum dioptimasi, dengan nilai 0.978

AUC mengukur performa algoritma di seluruh rentang *trade-off* antara *true positive rate* dan *false positive rate*. Semakin tinggi nilainya, maka semakin baik algoritma tersebut.



Gambar 4.2 – Grafik AUC dari RF sebelum dioptimasi, dengan nilai 0.988

Dapat kita lihat bahwa AUC dari RF sangat tinggi, menandakan bahwa RF memiliki performa yang sangat baik dalam mengklasifikasikan *instance* positif dan negatif. Sementara DL, meskipun tidak setinggi kNN dan RF, masih menunjukkan kinerja yang baik secara keseluruhan.



Gambar 4.3 – Grafik AUC dari DL sebelum dioptimasi, dengan nilai 0.969

Jika kita merujuk pada klasifikasi rentang nilai dari AUC di Tabel 2.3 pada Bab 2, maka ketiga algoritma ini tergolong bernilai Excellent.

4.3 Evaluasi Kinerja Menggunakan Optimasi PSO

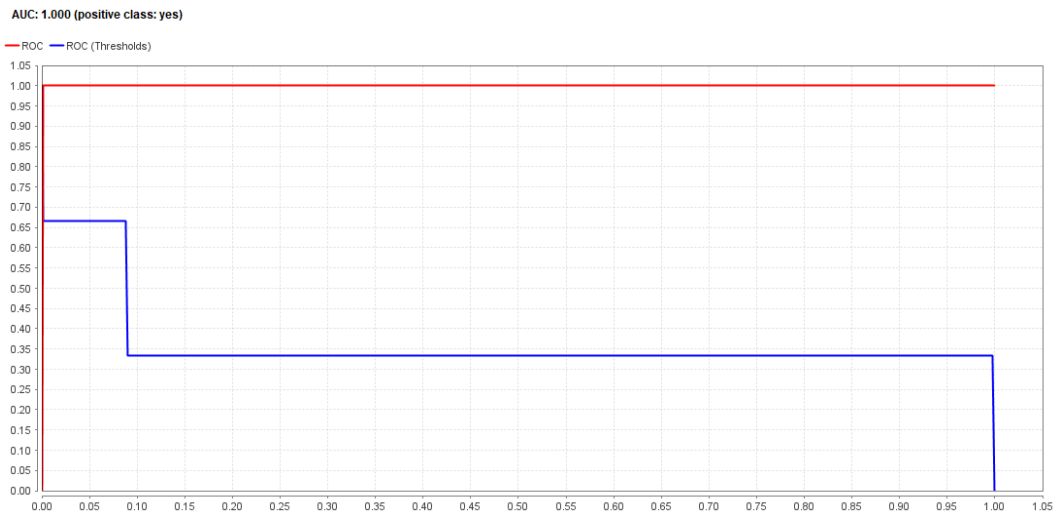
Setelah dioptimasi dengan menggunakan PSO, ketiga algoritma (kNN, RF, dan DL) mencapai akurasi, presisi, *recall*, dan ROC-AUC yang sangat tinggi atau bahkan sempurna.

Tabel 4.3 – Hasil kinerja dari kNN, RF, dan DL setelah dioptimasi PSO. Nilai akurasi, presisi, dan *recall* dalam persentase

	AKURASI	**PRESISI**	***RECALL***	**AUC**
kNN	100	100	100	1.000
Random Forest	100	100	100	1.000
Deep Learning	98.95	100	96.30	0.996

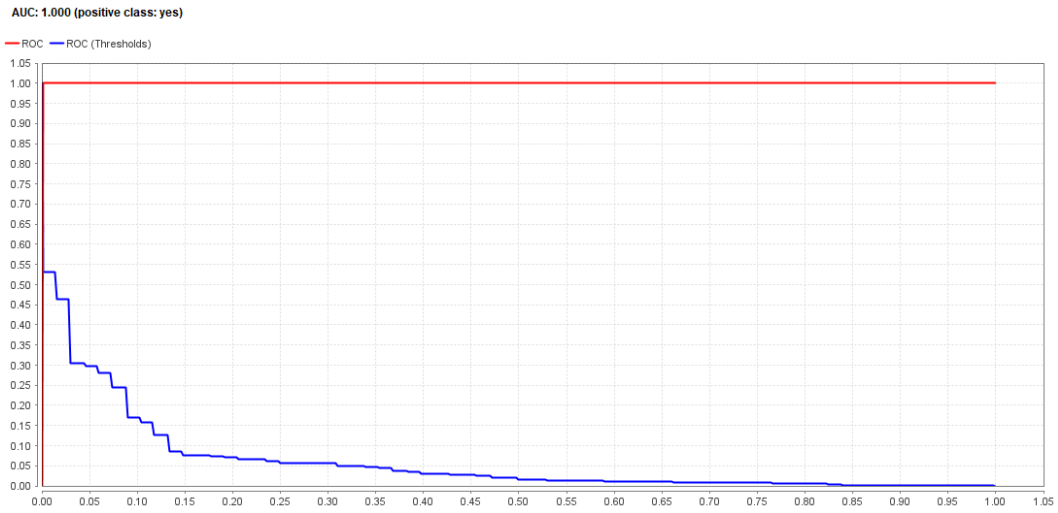
Kinerja yang luar biasa ini menunjukkan efektivitas optimalisasi menggunakan PSO dalam meningkatkan kualitas model.

Akurasi sempurna (mencapai 100%) menunjukkan bahwa setelah dioptimasi PSO, baik kNN maupun RF dapat mengklasifikasikan semua *instance* dengan benar. Sementara DL meskipun tidak sempurna, akurasinya juga tetap sangat tinggi setelah dioptimasi PSO, dengan peningkatan mencapai 5.27%.

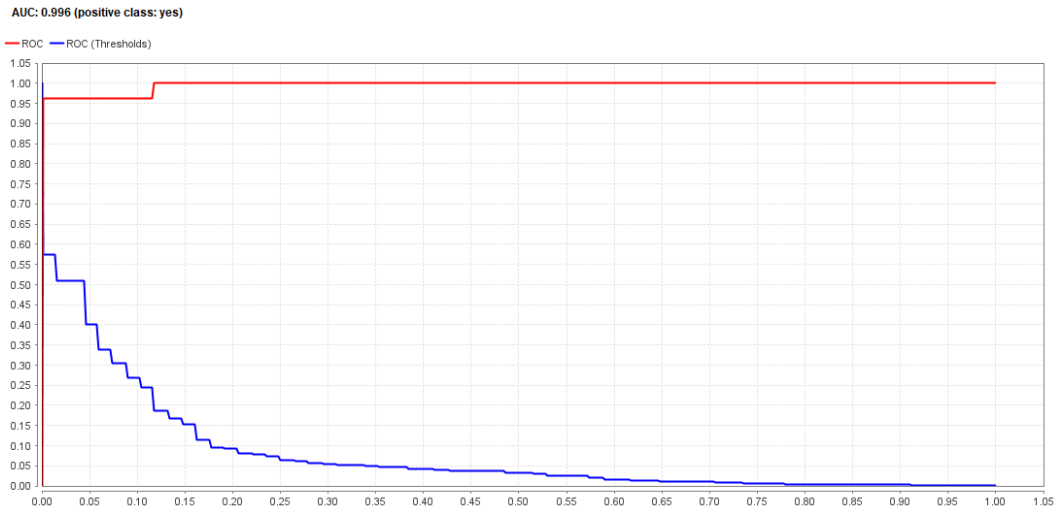


Gambar 4.4 – Grafik AUC dari kNN setelah dioptimasi PSO, dengan nilai 1.000

Persentasi presisi dari tiga algoritma adalah 100%, yang menunjukkan tidak adanya *false* positif setelah dioptimasi. Untuk *recall*: kNN dan RF mencapai 100% setelah dioptimasi, sementara DL nyaris sempurna karena masih ditemukan *false* negatif meski hanya sedikit sekali. Untuk AUC: meskipun DL tidak sesempurna kNN dan RF, namun tetap menunjukan AUC yang tinggi yang berarti memiliki performa yang sangat baik setelah dioptimasi.



Gambar 4.5 – Grafik AUC dari RF setelah dioptimasi PSO, dengan nilai 1.000



Gambar 4.6 – Grafik AUC dari DL setelah dioptimasi PSO, dengan nilai 0.996

Di sini bisa kita lihat bahwa kNN dan RF mencapai performa sempurna setelah dioptimasi oleh PSO. Sementara DL, juga tetap memiliki performa sangat tinggi meskipun tidak mencapai 100%. Lalu adakah perbedaan signifikan dari ketiganya? Istilah "beda signifikan" dalam konteks uji statistik pada algoritma *data mining* merujuk pada kebermaknaan atau signifikansi statistik dari perbedaan kinerja di antara dua atau lebih algoritma yang dibandingkan. Dalam konteks ini, uji t-Test digunakan untuk mengukur Apakah perbedaan yang diamati antara kelompok algoritma ini signifikan secara statistik atau apakah perbedaan tersebut mungkin terjadi oleh kebetulan.

Tabel 4.4 – Hasil uji signifikan (t-Test) ketiga algoritma: kNN, RF, dan DL

	kNN 1.000 +/-?	RF 1.000 +/-?	DL 0.989 +/-?
kNN 1.000 +/-?		1.000	1.000
RF 1.000 +/-?			1.000
DL 0.989 +/-?			

Dari hasil uji t-Test yang disajikan dalam Tabel 4.4 dapat kita amati bahwa tidak terdapat perbedaan yang signifikan dari ketiga algoritma ini. Karena nilai-p > 0,05 yang berarti bahwa kinerja dari masing-masing algoritma ini: baik itu k-Nearest Neighbor, Random Forest, maupun Deep Learning memang serupa. Dalam konteks *data mining* atau analisis klasifikasi, ini berarti bahwa algoritma-algoritma tersebut menghasilkan prediksi atau *output* yang tidak secara signifikan berbeda satu sama lain pada kumpulan data yang digunakan. Maka *statement* interpretasi statistik-nya dinyatakan sebagai berikut,

$$H_0: \mu_1 - \mu_2 = 0$$

yang mana,

- H_0 adalah hipotesis nol
- μ_1 adalah rata-rata kinerja dari algoritma pertama
- μ_2 Adalah rata-rata kinerja dari algoritma kedua

"Nilai p 1.000 > 0.05, maka ini berarti kita tidak akan menolak hipotesis nol pada tingkat signifikansi 0.05."

Dengan kata lain, kita tidak menemukan bukti statistik yang cukup untuk menyatakan bahwa ada perbedaan yang signifikan antara dua kinerja algoritma. Dengan demikian, hipotesis nol tetap berlaku.

Selanjutnya kita lihat *attribute weights* dari ketiga algoritma (disajikan dalam Tabel 4.5). Di sini hasil dari optimasi menggunakan PSO juga dapat memberikan informasi mengenai bobot atribut (*attribute weights*), yang mirip dengan apa yang biasanya ditemukan dalam hasil seleksi fitur. Bedanya, pada seleksi fitur, atribut yang kurang berbobot atau kurang informatif cenderung dihilangkan dari *dataset*. Ini digunakan ketika kita memiliki banyak atribut tapi tidak semua dari atribut itu relevan. Karena tujuan utama seleksi fitur ini adalah untuk memilih *subset* dari atribut yang paling relevan atau yang berkontribusi signifikan terhadap prediksi atau analisis yang ingin dilakukan, dan mengabaikan atribut yang kurang informatif atau mengandung *noise*.

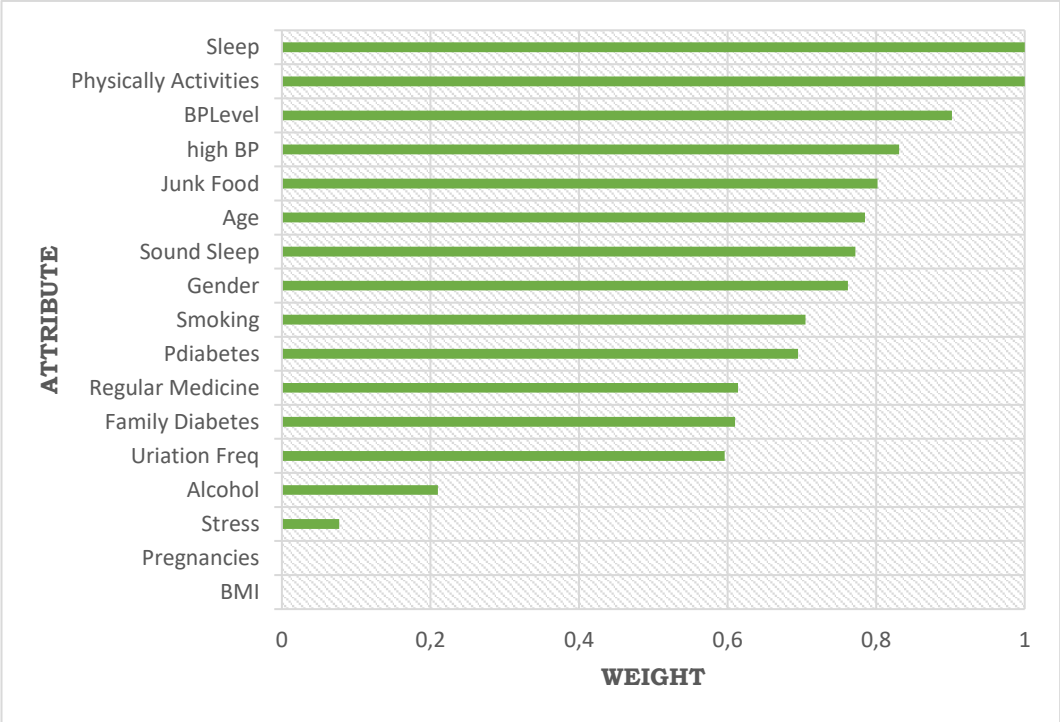
Tabel 4.5 – *Attribute Weights* dari ketiga algoritma setelah dioptimasi

NO	ATTRIBUTE	WEIGHTS		
		kNN	RF	DL
1	Age	0.785	1	0
2	Gender	0.762	0.133	1
3	high BP	0.831	1	0
4	BPLevel	0.902	0	0.728
5	Stress	0.077	1	0.055
6	BMI	0	0.713	0.341
7	Junk Food	0.802	0.979	0.034
8	Physically Activities	1	0.382	0.268
9	Sleep	1	0	0.789
10	Sound Sleep	0.772	0.885	0
11	Uriation Freq	0.596	0	0.611
12	Smoking	0.705	0.051	0.071
13	Alcohol	0.210	1	0
14	Regular Medicine	0.614	0.683	0.204
15	Family Diabetes	0.610	0.993	0.223
16	Pregnancies	0	0.515	0.556
17	Pdiabetes	0.695	0	0.057

Sementara model prediktif yang disajikan dalam penelitian ini memiliki struktur yang memungkinkan atribut-atribut memiliki bobot yang berbeda dalam mempengaruhi hasil akhir. Sehingga selain PSO secara tradisional digunakan untuk

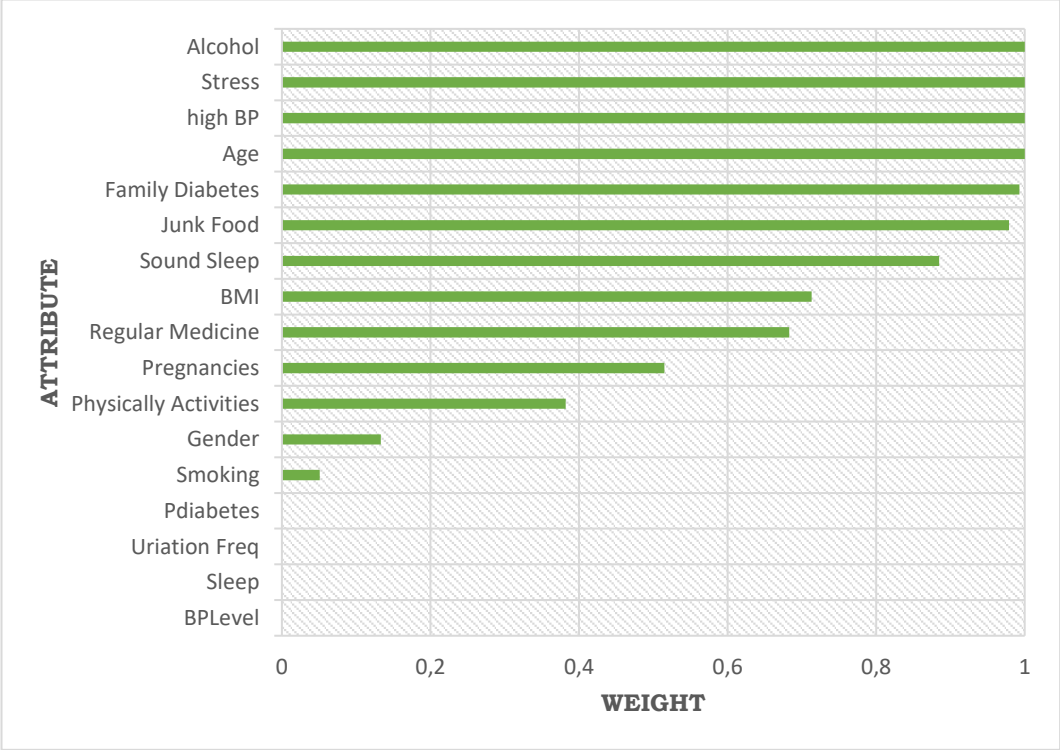
mengoptimasi set parameter algoritma, juga mengoptimalkan bobot atribut, yang kemudian memberikan hasil yang mirip dengan seleksi fitur.

Jadi selama iterasi, PSO juga akan mencoba untuk menemukan *subset* atribut yang memberikan kinerja model terbaik berdasarkan fungsi objektif-nya (yaitu, akurasi klasifikasi). Dengan demikian, atribut yang kurang berbobot atau kurang informatif cenderung tidak akan dimasukkan ke dalam *subset* atribut optimal yang ditemukan oleh PSO. Dari *atribute weights* yang disajikan di Tabel 4.5, kita dapat melihat bahwa setiap algoritma memberikan bobot yang berbeda untuk atribut yang sama, menunjukkan preferensi dan kepekaan yang berbeda terhadap fitur-fitur tertentu.

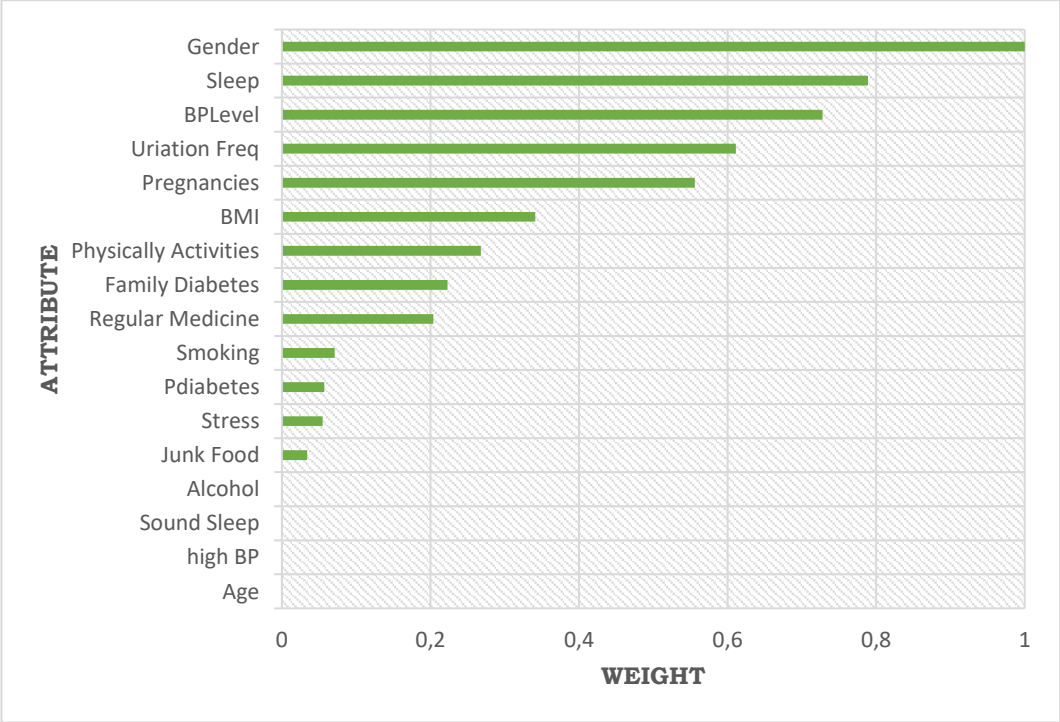


Gambar 4.7 – *Attribute Weights* dari algoritma kNN

Pada algoritma kNN (Gambar 4.7), atribut yang paling berpengaruh: Sleep (1), Physically Activities (1), BPLevel (0.902), high BP (0.831), Junk Food (0.802). Atribut dengan dampak rendah: BMI (0), Pregnancies (0), Stress (0.077), Alcohol (0.210).



Gambar 4.8 – *Attribute Weights* dari algoritma RF



Gambar 4.9 – *Attribute Weights* dari algoritma DL

Pada algoritma RF (Gambar 4.8), atribut yang paling berpengaruh: Stress (1), high BP (1), Alcohol (1), Age (1), Family Diabetes (0.993).
Atribut dengan dampak rendah: BPLevel (0), Sleep (0), Pdiabetes (0), Uriation Freq (0), Gender (0.133), Smoking (0.051), Physically Activities (0.382).

Pada algoritma DL (Gambar 4.9), atribut yang paling berpengaruh: Gender (1), BPLevel (0.728), Sleep (0.789), Uriation Freq (0.611).
Atribut dengan dampak rendah: Age (0), high BP (0), Sound Sleep (0), Alcohol (0), Junk Food (0.034), Stress (0.055), Pdiabetes (0.057), Smoking (0.071).

Dengan demikian, beberapa hal bisa kita simpulkan di sini:

- **Physically Activities:** sangat penting untuk kNN, sedangkan untuk RF dan DL memiliki dampak yang lebih rendah.
- **BMI:** tampaknya memiliki dampak yang cukup signifikan untuk RF, tapi untuk DL malah memiliki dampak yang rendah, bahkan tidak signifikan untuk kNN.
- **highBP dan Age:** sangat penting untuk RF dan cukup signifikan untuk kNN, tapi tidak signifikan untuk DL.
- **Stress:** sangat penting untuk RF, tapi untuk kNN dan DL malah memiliki dampak yang rendah.
- **Sleep:** sangat penting untuk kNN, cukup signifikan untuk DL, tapi tidak signifikan untuk RF.
- **Sound Sleep:** cukup signifikan untuk kNN dan RF, tapi tidak signifikan untuk DL.
- **Smoking:** sangat rendah untuk RF dan DL, tetapi cukup signifikan untuk kNN.
- **Alcohol:** meski tidak signifikan untuk DL dan memiliki dampak rendah untuk kNN, tapi justru sangat penting untuk RF.
- **Uriation Freq dan BPLevel:** memiliki dampak yang cukup signifikan untuk kNN dan DL, tapi tidak signifikan untuk RF.
- **Family Diabetes, Regular Medicine, dan Junk Food:** cukup signifikan untuk kNN dan RF, tapi malah rendah untuk DL.

- **Gender:** sangat penting untuk DL dan cukup signifikan untuk kNN, tapi untuk RF memiliki dampak yang rendah.
- **Pregnancies:** memiliki dampak yang cukup signifikan untuk RF dan DL, tapi tidak signifikan untuk kNN.
- **Pdiabetes:** cukup signifikan untuk kNN tapi lebih rendah untuk DL, namun samasekali tidak signifikan untuk RF.

Meski penting juga diketahui, bahwa *atribute weights* ini bisa berubah-ubah bergantung saat pembagian *dataset* di awal. Terutama karena kita menggunakan teknik pembagian *dataset* secara *stratified*. Saat membagi *dataset* secara *stratified*, kita memastikan bahwa proporsi kelas (penderita dan non-penderita diabetes) tetap seimbang, baik *data training* maupun *testing*-nya sehingga mempengaruhi *atribute weights*. Karena tiap bagian dari pembagian ini mungkin memiliki distribusi yang berbeda pada kelas-kelasnya.

Selaras dangan ini, beberapa peniliti juga meneliti bahwa faktor-foktor ini (*atribute*) *memang* memiliki pengaruh terhadap risiko seseorang terkena diabetes. Irwan, dkk [46] meneliti bahwa Family Diabetes atau keluarga yang memiliki penderita diabetes di dalamnya, memiliki potensi mewariskan penyakit ini ke anak turunannya. Priscila Evangelin Asa, dkk [47] bahkan menyebutkan bahwa Family Diabetes memliki risiko 41 kali mengidap diabetes tipe 2, disamping juga pengaruh dari kurangnya aktifitas fisik (Physically Activities) seperti olahraga dan kebiasaan merokok (Smoking).

Diabetes mudah dikenali dari gejela-gejala umum seperti intensitas buang air kecil yang cukup sering (Uriation Freq), merasa cepat lapar, luka yang sulit sembuh, cepet mengantuk, atau penglihatan yang mulai kabur [48]. Tingkat depresi (Stress) juga bisa memicu diabetes [49]. Hal ini bisa berbeda-beda berdasarkan kelompok umur (Age), Gender, dan penggunaan obat (Regular Medicine). Laki-laki memiliki risiko 2,48 kali lebih tinggi dari perempuan, bahkan untuk umur di atas 50 tahun meningkatkan risiko ini menjadi 2,16 kali lebih tinggi [50].

BMI berlebih atau berat badan tidak ideal pada seseorang (obesitas) memiliki 3,377 kali potensi terkena diabetes [51]. Ini juga berkaitan erat dengan tingkat stress pada diri seseorang, kualitas tidurnya (Sleep), pola makan yang tidak sehat seperti banyak mengonsumsi makanan cepat saji (Junk Food) dan alkohol [52][53][54], ataupun gestasional pada ibu hamil [55][56]. Selain itu penderita diabetes juga rentan terhadap hipertensi dan berpotensi mengalami stroke [2][3].