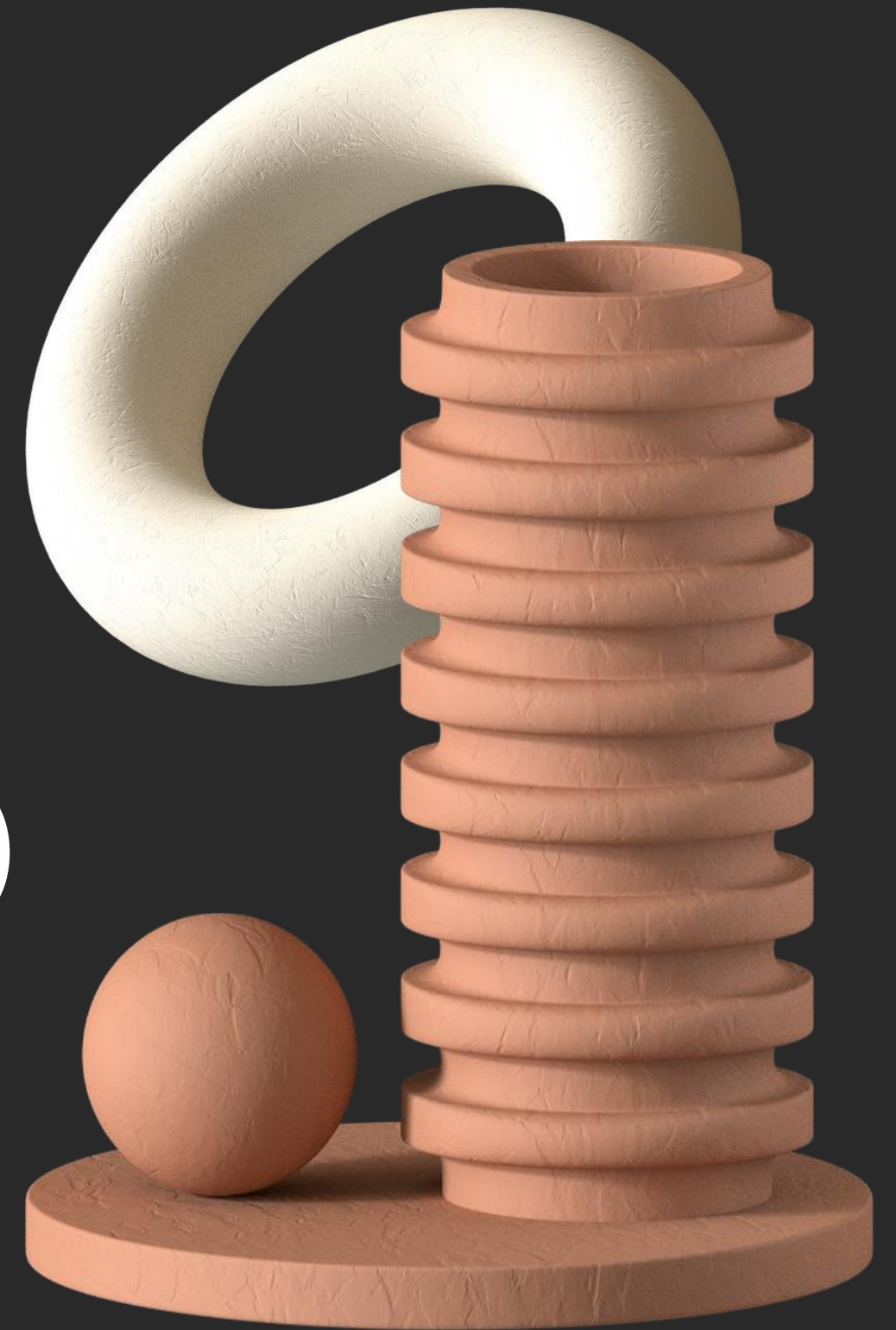


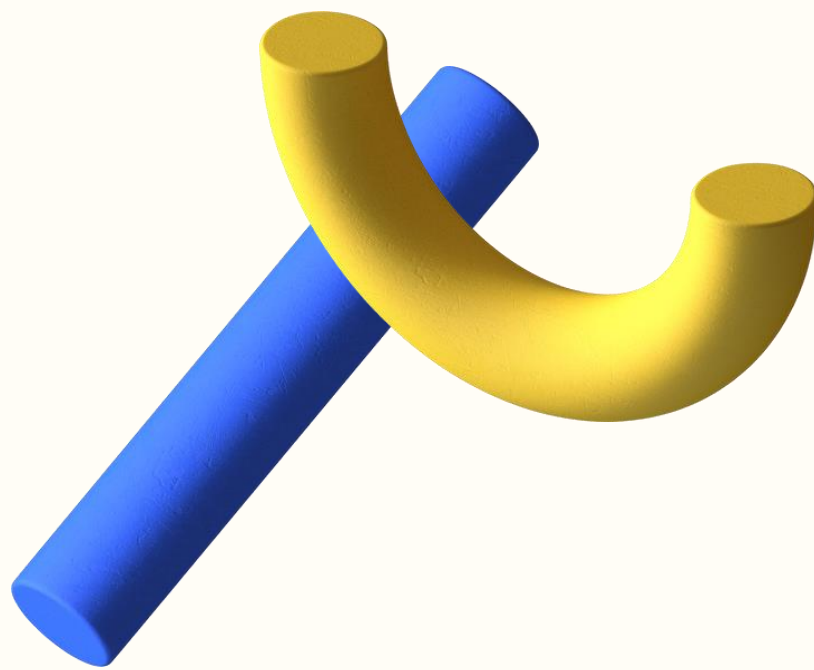
HEART ATTACK

K-NEAREST NEIGHBOUR (KNN)

&

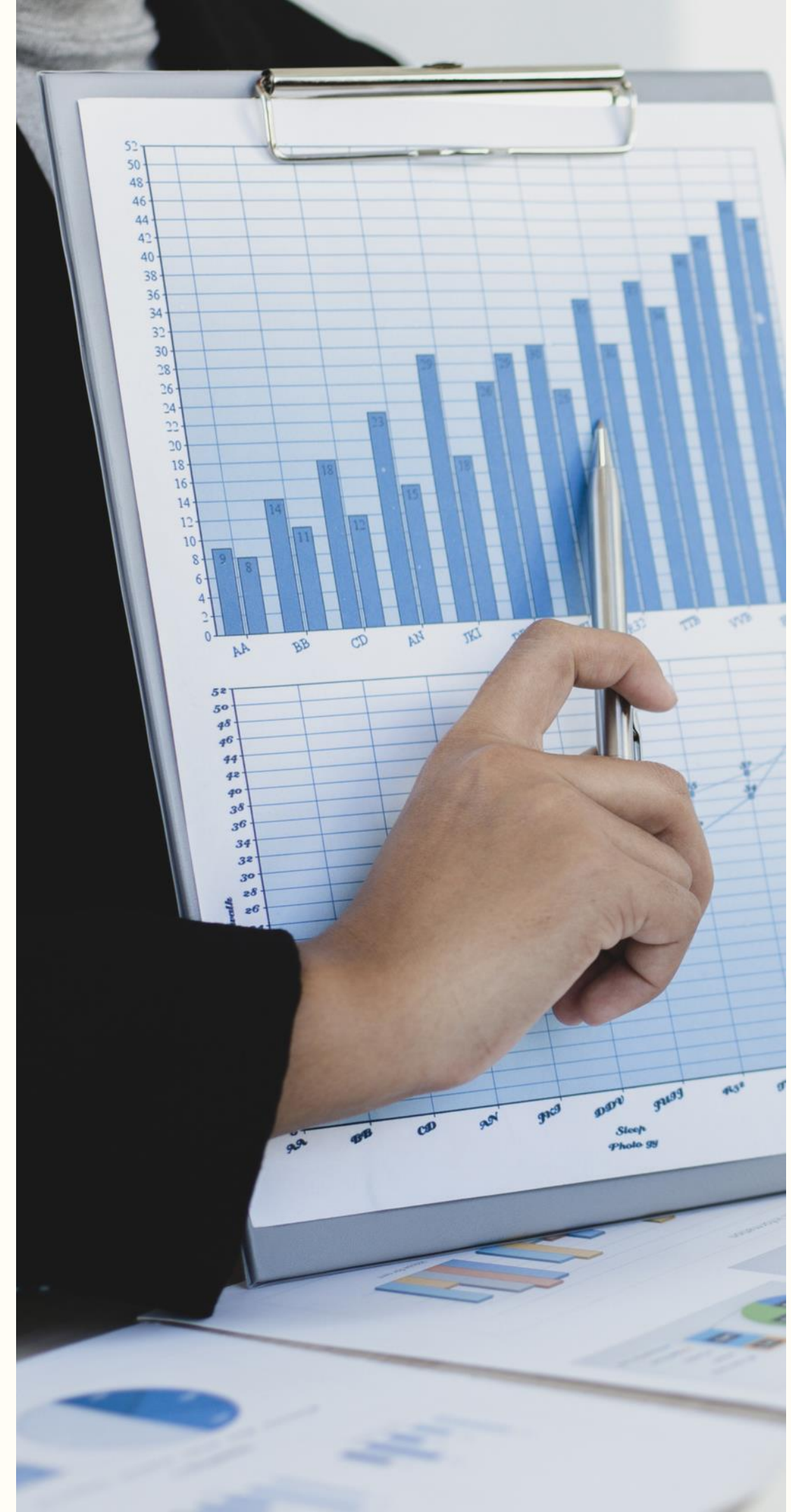
FUZZY K-NEAREST NEIGHBOUR (FKNN)





Klasifikasi

Klasifikasi adalah teknik dalam pembelajaran mesin yang digunakan untuk mengelompokkan atau mengkategorikan data ke dalam kelas atau kategori tertentu berdasarkan atribut-atribut atau fitur-fitur yang dimiliki oleh data tersebut. Tujuan utama dari klasifikasi adalah untuk memprediksi label kelas dari data yang tidak memiliki label, berdasarkan pola-pola atau hubungan yang telah dipelajari dari data pelatihan yang sudah memiliki label.



Processing Data Awal

Missing Value

Hair (1995) mengemukakan bahwa apabila persentase data missing value melebihi 30%, maka data boleh dihapus sedangkan jika persentase data missing value kurang 30%, maka data missing diimputasi

- Mean jika data kuantitatif
- Modus jika data kualitatif

Multikolinieritas

Multikolinearitas adalah adanya hubungan linear atau korelasi yang tinggi antar variabel. Pengujian multikolinearitas dilihat dari besaran VIF (Variance Inflation Factor). Nilai cut off yang umum dipakai untuk menunjukkan adanya multikolinearitas adalah nilai $VIF \geq 10$.

KNN (K-Nearest Neighbour)

K-Nearest Neighbors (KNN) adalah algoritma klasifikasi yang berfokus pada kedekatan jarak antara data. Ini adalah teknik sederhana namun sangat tergantung pada pemilihan nilai K, yang menentukan jumlah tetangga terdekat yang digunakan dalam klasifikasi. Nilai K yang terlalu besar dapat mengakibatkan inklusi tetangga yang tidak relevan, sementara K yang terlalu kecil dapat menyebabkan sensitivitas terhadap outlier. Oleh karena itu, diperlukan metode klasifikasi yang lebih canggih untuk mengatasi masalah ini dan meningkatkan akurasi klasifikasi.




Klasifikasi berbasis nearest neighbor dilakukan berdasarkan jarak antara data testing dengan data training yang dapat dihitung salah satunya dengan jarak Euclidean dengan yang didefinisikan dalam persamaan 3 (Han & Kamber, 2006).

FKNN (Fuzzy K-Nearest Neighbour)

FKNN adalah perkembangan dari KNN yang menggabungkan konsep Fuzzy. Konsep himpunan fuzzy digunakan untuk mengukur tingkat keanggotaan data pada setiap kelas. Himpunan fuzzy adalah bentuk umum dari himpunan biasa dengan tingkat keanggotaan dalam interval $[0,1]$. Semakin dekat data ke tetangganya, semakin tinggi tingkat keanggotaan data tersebut pada kelas tetangganya.

Perhitungan jarak antara data disesuaikan dengan jenis datanya.

Hasil prediksi FKNN adalah kelas dengan nilai keanggotaan terbesar pada data tersebut.



```
</div>  
</div>
```



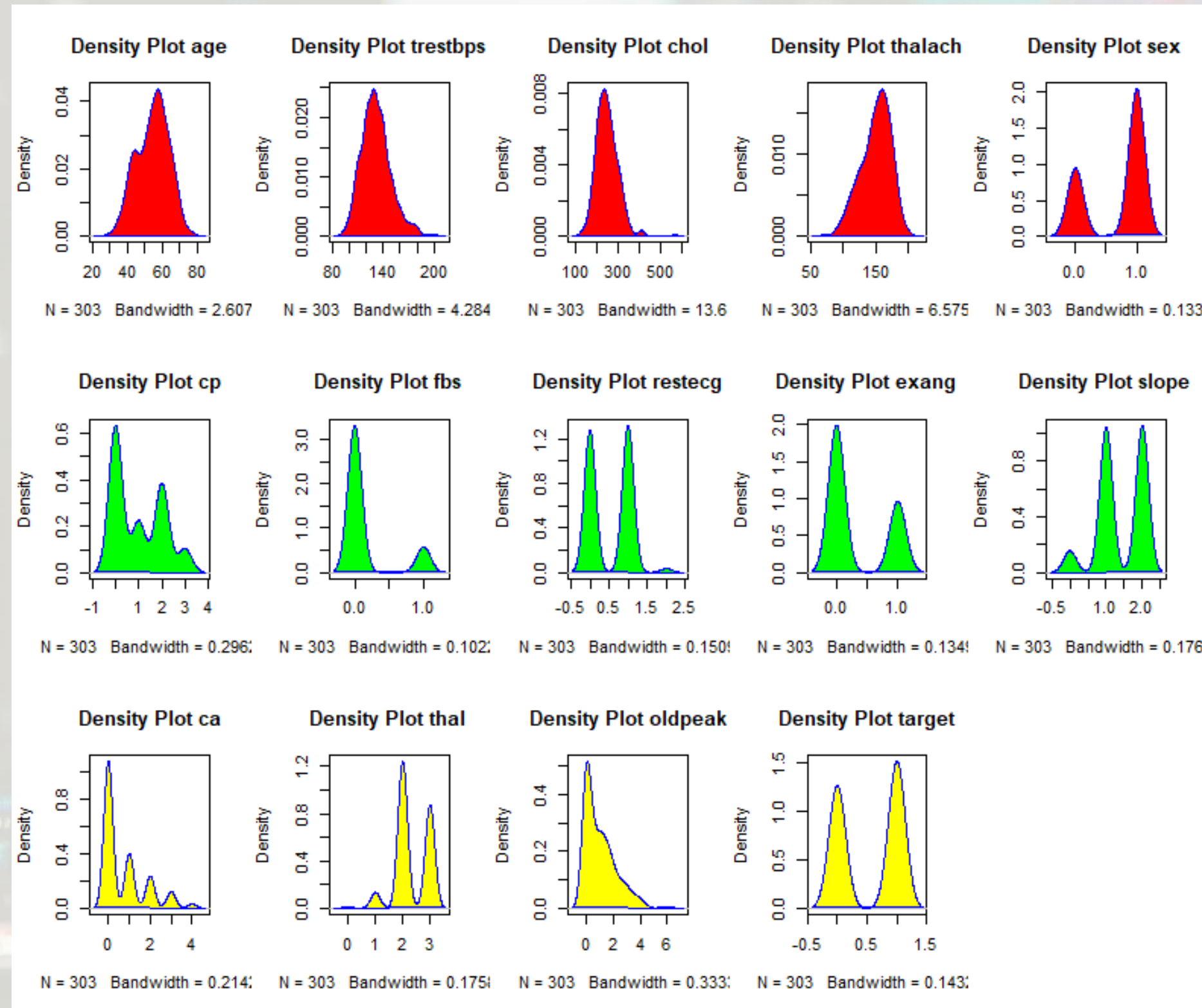
INPUT DATA

Data yang digunakan merupakan dataset serangan jantung yang diambil dari <https://www.kaggle.com/datasets/pritsheta/heart-attack>.

Set data terdiri dari 14 variabel dengan jumlah data sebanyak 303 baris.



PLOT DENSITAS SETIAP VARIABEL



Jika dilihat pada output plot densitas tersebut terbukti bahwa sebagian besar dari variabel dataset ini membentuk dan mengikuti distribusi normal.



k-Nearest Neighbor

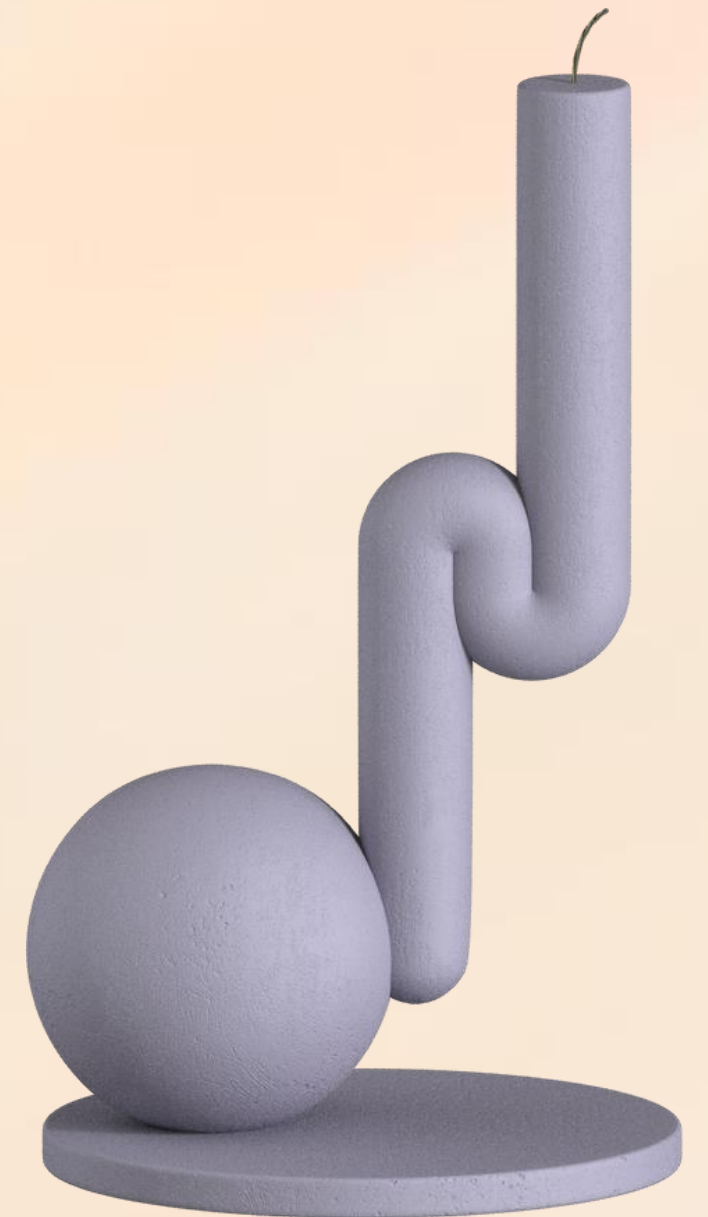
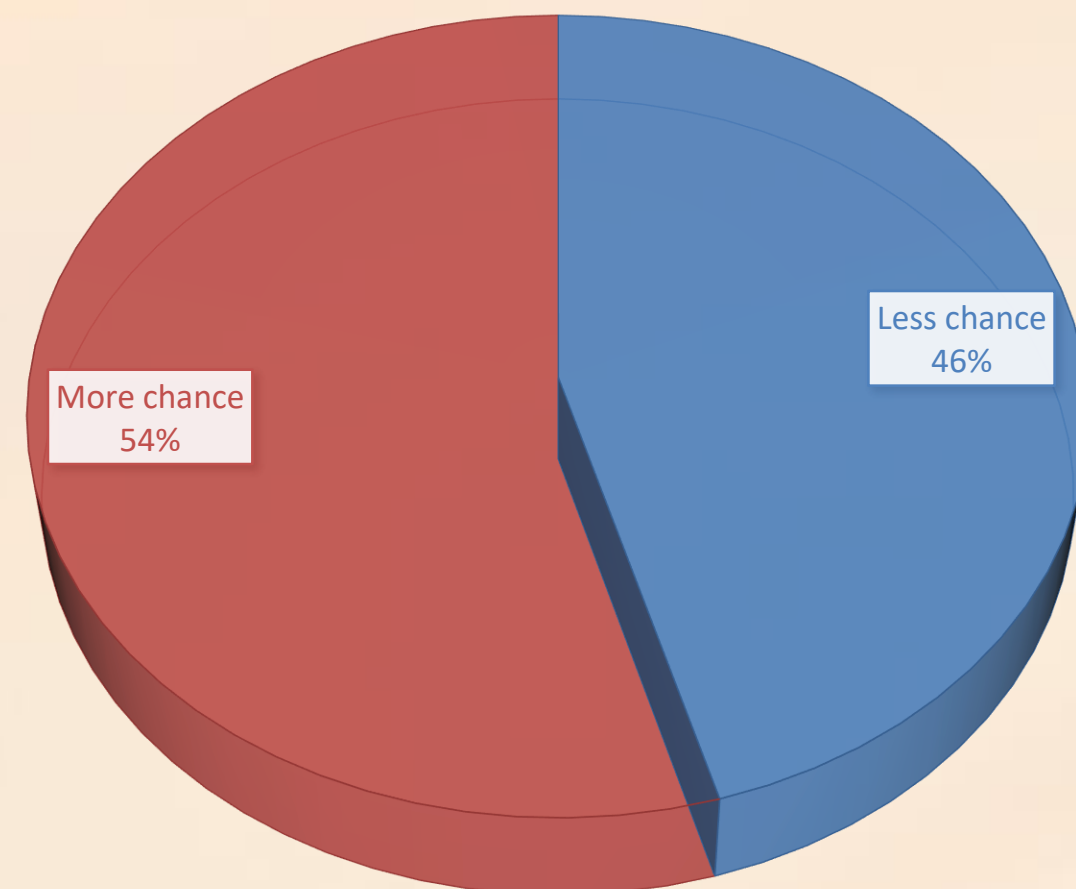


CROSS VALIDATION

Validasi dilakukan menggunakan metode *Hold out* dengan perbandingan data *training* dengan data *testing* sebesar 80 : 20

Sebelumnya dilakukan proses pengacakan data dengan menggunakan sintaks `set.seed`, agar saat dilakukan *running* ulang akan menghasilkan nilai yang sama.

PROPORSI TARGET



PENENTUAN K ATAU TETANGGA TERDEKAT TERBAIK

Pada tahap ini, penentuan K atau tetangga terdekat dengan menggunakan nilai akar dari panjang data training.

```
> round(sqrt(nrow(heart_train)),0)  
[1] 16
```

Diperoleh nilai $K = 16$, namun dalam klasifikasi KNN disarankan menggunakan K dengan jumlah ganjil agar tidak membingungkan algoritma dari mesin pengolah. Sehingga untuk tahap selanjutnya akan digunakan nilai $K = 17$ untuk memperoleh tingkat keakuratan klasifikasi.

OUTPUT

Confusion Matrix and Statistics

Reference

Prediction	less chance	more chance
less chance	16	7
more chance	13	25

Accuracy : 0.6721

95% CI : (0.54, 0.7869)

No Information Rate : 0.5246

P-Value [Acc > NIR] : 0.01394

Kappa : 0.3362

McNemar's Test P-Value : 0.26355

Sensitivity : 0.5517

Specificity : 0.7812

Pos Pred Value : 0.6957

Neg Pred Value : 0.6579

Prevalence : 0.4754

Detection Rate : 0.2623

Detection Prevalence : 0.3770

Balanced Accuracy : 0.6665

'Positive' Class : less chance

Berdasarkan output diperoleh nilai *accuracy* sebesar 67,21%, nilai *recall* sebesar 55,17%, nilai *specificity* sebesar 78,12% dan nilai *precision* sebesar 69,56%.



Fuzzy K -Nearest Neighbor



MENENTUKAN JARAK & EUCLIDIAN

```
for(i in 1:nrow(train_data)){
  train_data$jarak[i]<-sqrt((train_data$age[i]-
test_data$age[1])^2+(train_data$sex[i]-test_data$sex[1])^2+
                        (train_data$cp[i]-
test_data$cp[1])^2+(train_data$trtbps[i]-test_data$trtbps[1])^2+
                        (train_data$chol[i]-
test_data$chol[1])^2+(train_data$fbs[i]-test_data$fbs[1])^2+
                        (train_data$restecg[i]-
test_data$restecg[1])^2+(train_data$thalachh[i]-test_data$thalachh[1])^2+
                        (train_data$exng[i]-
test_data$exng[1])^2+(train_data$oldpeak[i]-test_data$oldpeak[1])^2+
                        (train_data$slp[i]-
test_data$slp[1])^2+(train_data$caa[i]-test_data$caa[1])^2+
                        (train_data$thall[i]-test_data$thall[1])^2)
  train_data$d[i]=train_data$jarak[i]^(-2)
}
View(train_data)
```

	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output	jarak	d
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1	29.33070	1.162399e-03
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1	27.96873	1.278364e-03
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1	35.88816	7.764216e-04
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1	11.02905	8.220980e-03
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1	118.95529	7.066958e-05
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1	52.12638	3.680313e-04
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1	62.51952	2.558402e-04
44	1	1	120	263	0	1	173	0	0.0	2	0	3	1	31.68596	9.960159e-04
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1	57.52608	3.021833e-04
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1	70.93349	1.987455e-04
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1	17.90084	3.120709e-03
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1	53.20752	3.532271e-04
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1	31.26276	1.023164e-03
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1	44.54481	5.039713e-04
50	0	1	150	202	0	0	155	0	1.0	2	0	2	1	53.55173	3.630055e-04

MENENTUKAN JARAK MINIMUM (TETANGGA TERDEKAT)

```
> #Kelas 0
> k1
[1] 12.85457
> k2
[1] 16.8333
> k3
[1] 20.49293
> #Kelas 1
> k1
[1] 11.02905
> k2
[1] 11.18034
> k3
[1] 12.08305
```

Jarak minimum dari tiga tetangga terdekat ini akan menentukan probabilitas klasifikasi dari kelas 0 dan kelas 1.

Probabilitas klasifikasi

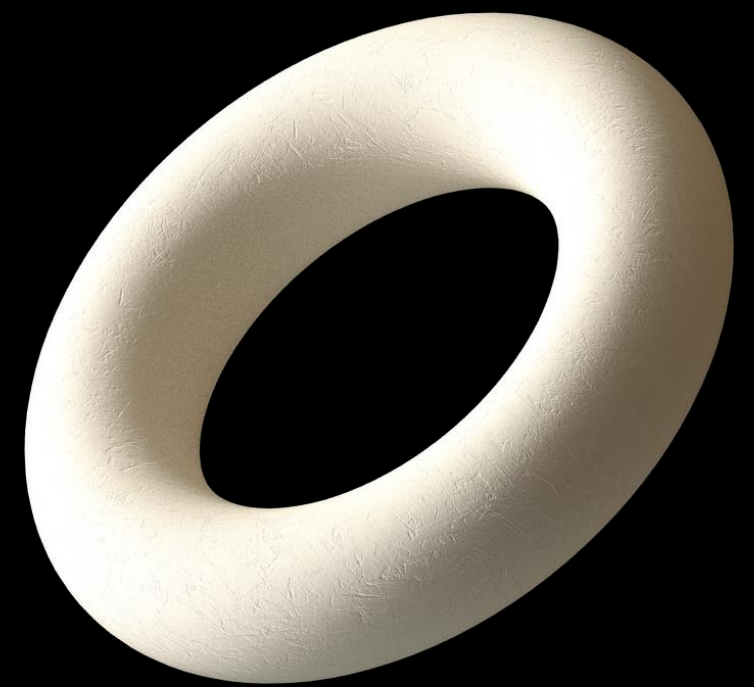
```
> cat("Probabilitas Kelas 0 adalah ",Prob0)
Probabilitas Kelas 0 adalah 0.4502571
> cat("Probabilitas Kelas 1 adalah ",Prob1)
Probabilitas Kelas 1 adalah 0.5497429
```

Berdasarkan *output* di atas diperoleh hasil klasifikasi dari sampel yang tidak terkena serangan jantung (0) sebesar 0,4502571 atau 45,03% dan sampel yang terkena serangan jantung (1) sebesar 0,5497429 atau 54,74%.



A series of seven thick, orange, hand-painted brushstrokes radiating from a central point on the left side of the image. The strokes vary in length and angle, creating a fan-like effect.

KESIMPULAN



KESIMPULAN

KNN (K-Nearest Neighbour)

- Penentuan K atau tetangga terdekat terbaik

Diperoleh nilai $K = 16$, namun dalam klasifikasi KNN disarankan menggunakan K dengan jumlah ganjil agar tidak membingungkan algoritma dari mesin pengolah. Sehingga digunakan nilai $K=17$

Nilai Accuracy atau rasio prediksi benar dengan keseluruhan data sebesar 67,21%.

Nilai Recall atau rasio prediksi benar positif dengan keseluruhan data benar positif sebesar 55,17%.

Nilai Specificity atau rasio prediksi benar negatif dengan keseluruhan data benar negatif sebesar 78,12%

Nilai Precision atau rasio prediksi benar positif dengan keseluruhan data hasil prediksi positif sebesar 69,56%.



KESIMPULAN

FKNN (Fuzzy K-Nearest Neighbour)

- Menentukan Jarak Minimum dari Tetangga Terdekat
Diperoleh 4 jarak minimum pada kelas yaitu:

Kelas 0	Kelas 1
K1=12.85457	K1=11.02905
K2=16.8333	K2=11.18034
K3=20.49293	K3=12.08305

- Menentukan Probabilitas Hasil Klasifikasi dari Masing-masing Kelas
Hasil klasifikasi dari sampel yang tidak terkena serangan jantung (0) sebesar 0,4502571 (45,03%) dan sampel yang terkena serangan jantung (1) sebesar 0,5497429 (54,74%).

