**Laporan Project UTS Mata Kuliah Analisis Visualisasi Data**

**“Application of Tableau in Visual Analysis Data of a US Supermarket Sales”**



**Kelompok 1 :**

1. **Johan Rolanda Napitupulu (081711633028)**
2. **Ferry Triwantono (081711633048)**

**Program Studi S1 – Sistem Informasi**

**Fakultas Sains dan Teknologi**

**Universitas Airlangga**

**Surabaya**

**Juni 2020**

# DAFTAR ISI

[DAFTAR ISI 2](#_Toc43838020)

[BAB I 3](#_Toc43838021)

[PENDAHULUAN 3](#_Toc43838022)

[1.1 Latar Belakang 3](#_Toc43838023)

[1.2 Tujuan 3](#_Toc43838024)

[1.3 Manfaat 4](#_Toc43838025)

[BAB II 5](#_Toc43838026)

[DASAR TEORI 5](#_Toc43838027)

[2.1 Kanker Prostat 5](#_Toc43838028)

[2.1.1 Gejala Kanker Prostat 5](#_Toc43838029)

[2.2 Tumor 6](#_Toc43838030)

[2.3 K-Nearest Neighbors 6](#_Toc43838031)

[2.3.1 Cara Kerja KNN 6](#_Toc43838032)

[2.3.2 Langkah Kerja KNN 6](#_Toc43838033)

[BAB III 8](#_Toc43838034)

[METODOLOGI PENELITIAN 8](#_Toc43838035)

[3.1 Langkah-langkah melakukan Metode KNN 8](#_Toc43838036)

[BAB IV 17](#_Toc43838037)

[UJI COBA DAN PEMBAHASAN 17](#_Toc43838038)

[4.1 Perbandingan hasil setiap percobaan 17](#_Toc43838039)

[4.2 Analisa Masalah 18](#_Toc43838040)

[4.3 Solusi Permasalahan 19](#_Toc43838041)

[BAB V 23](#_Toc43838042)

[KESIMPULAN DAN SARAN 23](#_Toc43838043)

[5.1 Kesimpulan 23](#_Toc43838044)

[5.2 Saran 23](#_Toc43838045)

[DAFTAR PUSTAKA 24](#_Toc43838046)

# BAB I

# PENDAHULUAN

## 1.1 Latar Belakang

Kanker Prostat adalah penyakit kanker yang berkembang di prostat, sebuah kelenjar dalam sistem reproduksi lelaki. Hal ini terjadi ketika sel prostat mengalami mutasi dan mulai berkembang di luar kendali. Kanker prostat dapat menimbulkan rasa sakit, kesulitan buang air kecil, dan disfungsi ereksi. Di Indonesia, pada tahun 1992 saja sudah disimpulkan bahwa kanker prostat menduduki urutan ke 9 dengan 310 kasus baru (4,07%) dari 10 kasus kanker yang diperoleh dari laporan berbagai rumah sakit. Disimpulkan pula bahwa pada laki – laki di atas usia 65 tahun, kanker prostat menempati urutan ke 2 dengan 202 kasus (12,31%) (Sarjadi,1999). Kanker prostat umumnya tidak menunjukkan gejala khas. Karena itu, sering terjadi keterlambatan diagnosa. Gejala yang ada umumnya sama dengan gejala pembesaran prostat jinak atau Benign Prostate Hyperplasia (BPH), yaitu buang air.

Penyebab kanker prostat tidak diketahui secara tepat, meskipun beberapa penelitian menunjukkan adanya hubungan antara diet tinggi lemak dengan peningkatan kadar hormon testosteron. Pada bagian lain, Rindiastuti (2007) misalnya menyimpulkan bahwa usia lanjut mengalami penurunan beberapa unsur esensial tubuh seperti kalsium (Ca) dan vitamin D. Tetapi pola makan dengan Ca tinggi secara berlebihan dapat meningkatkan risiko kanker prostat pada usia lanjut.

Banyak orang belum mengetahui secara pasti apakah tubuhnya telah diserang oleh kanker prostat atau tidak. Pada penelitian ini, penulis menggunakan bantuan algoritma *K-Nearest Neighbors*. *KNN* merupakan metode untuk klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumnya. Klasifikasi ini didasarkan pada analogi pembelajaran, yaitu dengan membandingkan data testing dengan data training yang mirip (Han, 2006). Melihat fakta tersebut, penelitian ini akan memprediksi kanker prostat menggunakan bantuan algoritma *K-Nearest Neighbors* sebagai pengklasifikasinya.

## 1.2 Tujuan

Memprediksi adanya kanker atau tumor ganas berdasarkan data pemeriksaan pasien lebih dini.

## 1.3 Manfaat

Mengetahui adanya kanker atau tumor ganas berdasarkan data pemeriksaan pasien lebih dini.

# BAB II

# DASAR TEORI

## 2.1 Kanker Prostat

Kelenjar prostat terletak di bawah kandung kencing pria. Fungsinya adalah untuk menghasilkan air mani sebagai sumber nutrisi sperma. Kanker prostat adalah bentuk kanker yang berkembang di prostat, sebuah kelenjar dalam sistem reproduksi laki-laki. kanker prostat Kebanyakan lambat berkembang, namun terdapat kasus kanker prostat agresif. Sel-sel kanker dapat metastasis (menyebar) dari prostat ke bagian tubuh lainnya, terutama tulang dan kelenjar getah bening. kanker prostat dapat menyebabkan rasa sakit, kesulitan buang air kecil, masalah selama hubungan seksual, atau disfungsi ereksi. Gejala lain yang berpotensi dapat mengembangkan selama stadium penyakit.

Harga deteksi kanker prostat sangat bervariasi di seluruh dunia, dengan Asia Selatan dan Timur deteksi lebih jarang daripada di Eropa, dan khususnya Amerika Serikat. Kanker prostat cenderung untuk mengembangkan pada pria berusia lebih dari lima puluh dan meskipun ini adalah salah satu jenis kanker yang paling umum pada laki-laki, banyak yang tidak pernah mengalami gejala, menjalani terapi tidak, dan akhirnya meninggal karena penyebab lainnya.

Hal ini karena kanker prostat adalah, dalam banyak kasus, lambat berkembang, gejala-bebas, dan karena laki-laki dengan kondisi yang lebih tua mereka sering mati karena sebab-sebab yang tidak terkait dengan kanker prostat, seperti jantung / penyakit peredaran darah, pneumonia, lainnya tidak terkait kanker, atau usia tua. Sekitar 2 / 3 dari kasus lambat tumbuh “kucing”, yang lain ketiga lebih agresif, cepat berkembang secara informal dikenal sebagai “macan” (Dewiana, 2015).

### 2.1.1 Gejala Kanker Prostat

**a)** Terlalu sering kebelet pipis, keluarnya sedikit. Mirip gejala diabetes.

**b)** Urine sulit keluar saat akan pipis.

**c)** Rasa sakit di penis saat ejakulasi.

**d)** Ada darah dalam urine maupun semen.

## 2.2 Tumor

Tumor atau barah (bahasa Inggris: tumor, tumour) adalah sebutan untuk neoplasma atau lesi padat yang terbentuk akibat pertumbuhan sel tubuh yang tidak semestinya, yang mirip dengan simtoma bengkak. Tumor berasal dari kata tumere dalam bahasa latin yang berarti "bengkak". Pertumbuhannya dapat digolongkan sebagai ganas (malignan) atau jinak (benign) (Wikipedia, 2018).

1. **Tumor Ganas (Malginan)**

Tumor ganas disebut kanker. Kanker memiliki potensi untuk menyerang dan merusak jaringan yang berdekatan dan menciptakan metastasis.

1. **Tumor Jinak (Benign)**

Tumor jinak tidak menyerang tissue berdekatan dan tidak menyebarkan benih (metastasis), tetapi dapat tumbuh secara lokal menjadi besar. Mereka biasanya tidak muncul kembali setelah penyingkiran melalui operasi

## 2.3 K-Nearest Neighbors

K-NN adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumnya. Klasifikasi ini didasarkan pada analogi pembelajaran, yaitu dengan membandingkan data testing dengan data training yang mirip (Han, 2006). Data training adalah data yang mentah / data yang dilatih untuk *algoritma machine learning*, sedangkan data testing adalah data yang sudah dievaluasi.

### 2.3.1 Cara Kerja KNN

Data training diproyeksikan ke ruang berdimensi banyak, yang mana masing-masing dimensi menjelaskan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data training.

### 2.3.2 Langkah Kerja KNN

1) Tentukan parameter K = jumlah dari persekitaran (nearest neighbors).

2) Hitung jarak antara data baru yang ditanyakan dengan seluruh sampel data pelatihan.

3) Urutkan seluruh jarak berdasarkan jarak minimum dan tetapkan persekitaran sesuai dengan nilai K.

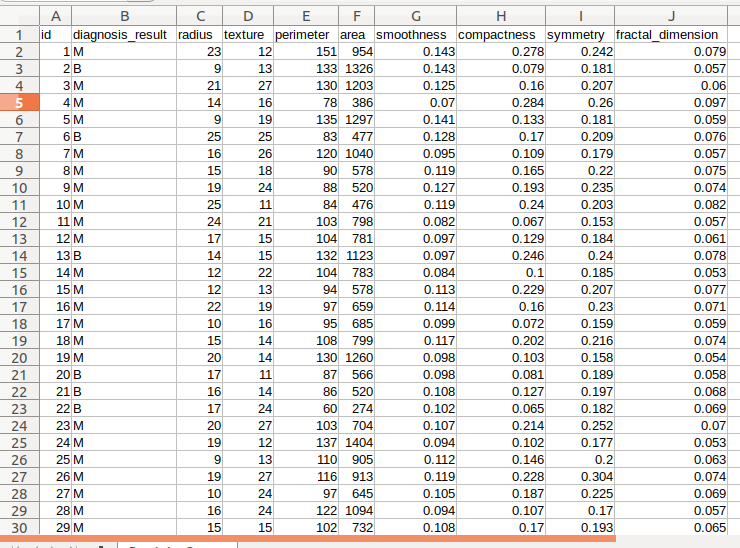
4) Sesuaikan klasifikasi dari kategori Y dengan persekitaran yang telah ditetapkan.

5) Gunakan kelas dengan jumlah terbanyak sebagai dasar menentukan kelas dari data baru yang ditanyakan.

# BAB III

# METODOLOGI PENELITIAN

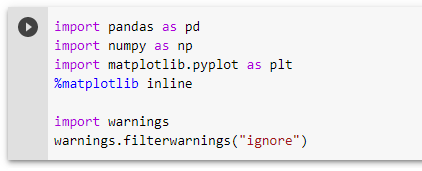
Dataset yang digunakan adalah 100 pasien dengan hasil pemeriksaan terkait tumor yang ada di prostat. Data tersebut diambil dari https://www.kaggle.com/sajidsaifi/prostate-cancer/



Gambar Dataset Pasien Tumor Prostat

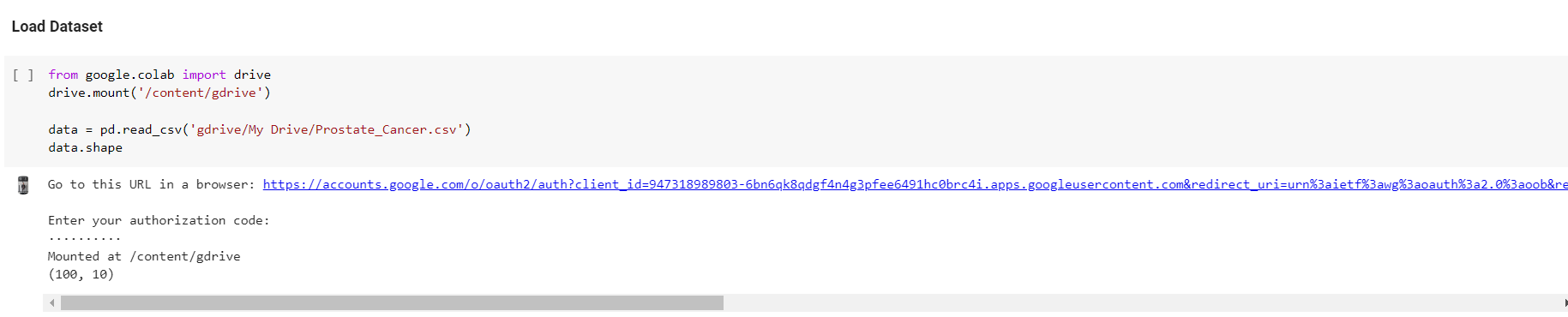
## 3.1 Langkah-langkah melakukan Metode KNN

1. Mengimport Library

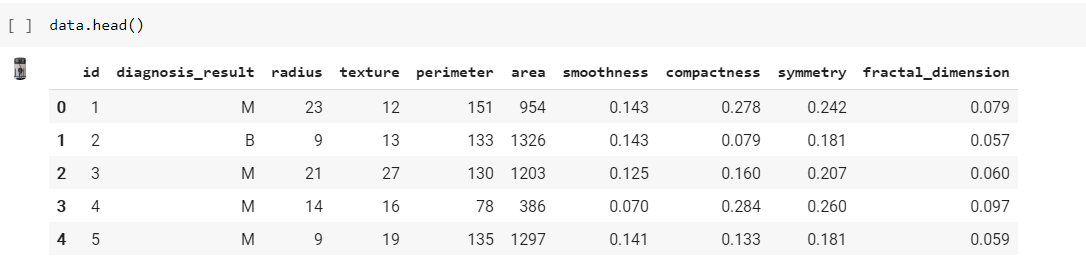


Langkah ini untuk mengimport library yang diperlukan.

1. Load Dataset

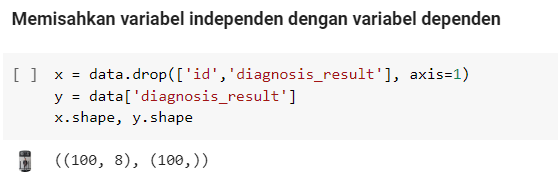


Pada step untuk memuat data yang sudah diperoleh. Di sini dataset yang diperoleh dimasukkan ke dalam google drive. Terlihat data yang dimuat terdapat 100 baris dan 10 kolom



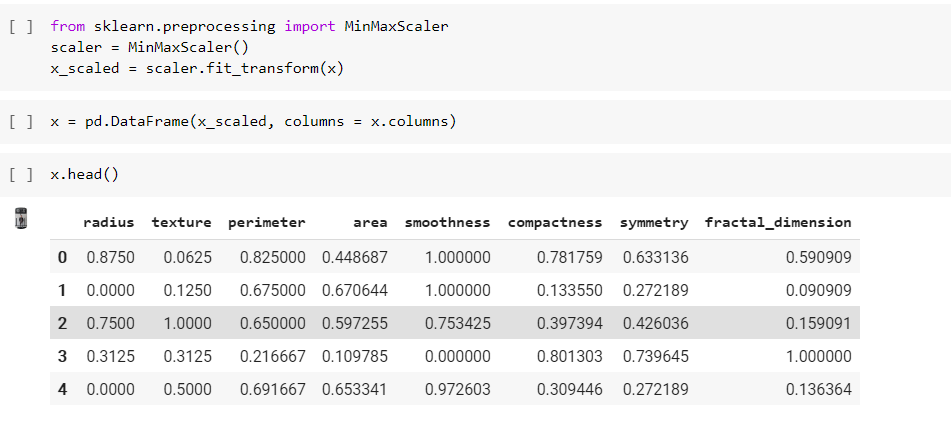
Menampilkan data yang sudah diload. Data yang ditampilkan 5 ter atas.

1. Memisahkan Variabel Independen Dengan Variabel Dependen\



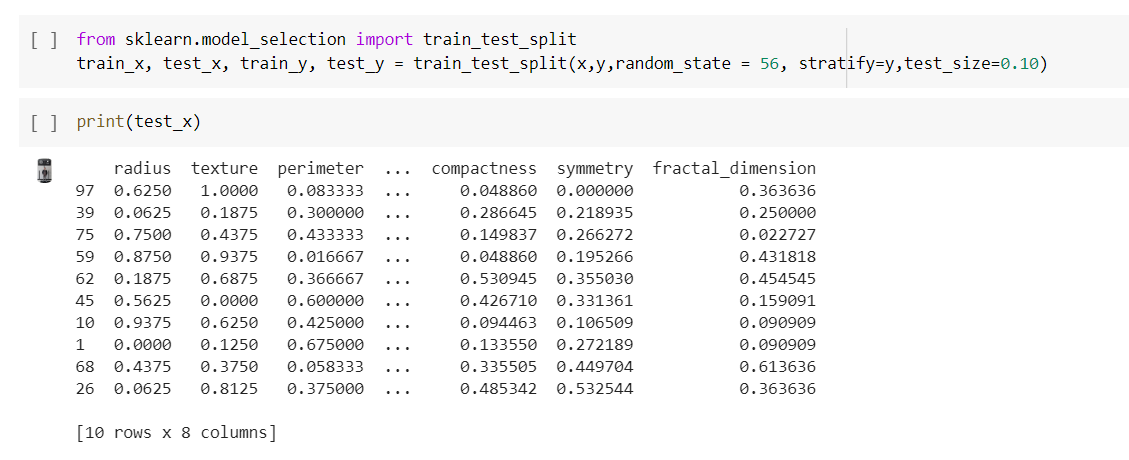
Langkah ini untuk memisahkan variabel independen dengan dependen.

1. Mengubah Skala (Menggunakan MinMax Scaler)



Langkah ini bertujuan untuk menormalisasi data dan mengubah ke skala umum.

1. Pembagian Dataset (Data Training 90% dan Data Test 10%)

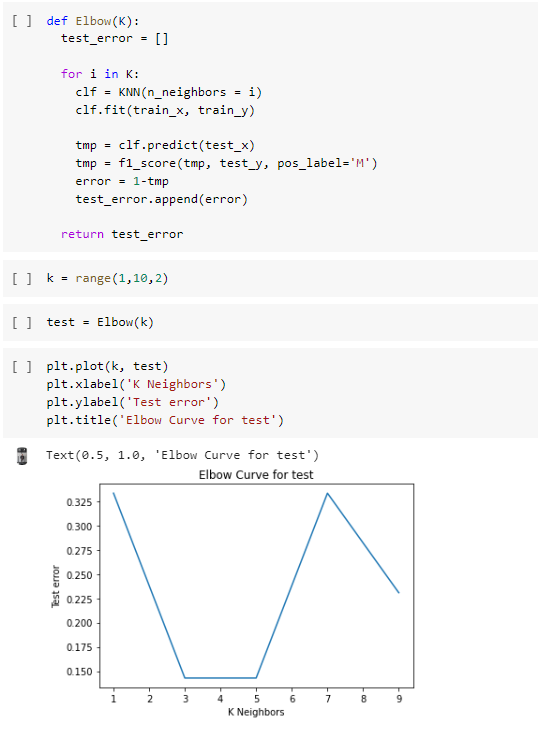


* Langkah ini dataset diatur menjadi data training sebanyak 90% dan data test 10%. Terdapat 10 baris dan 8 kolom di dataset ini.
* Menerapkan KNN Classifier



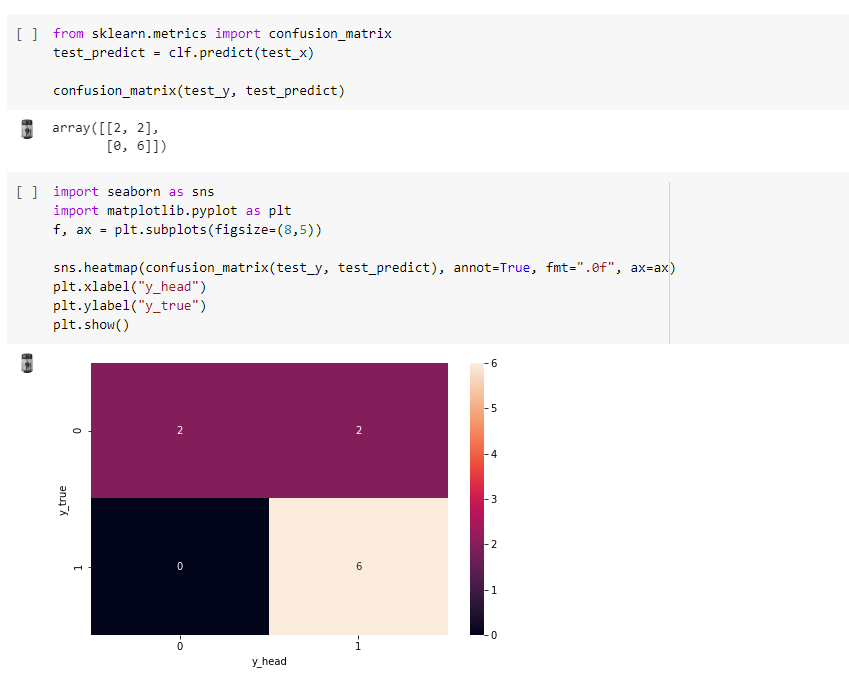
Menerapkan KNN Classifier pada dataset yang sudah diatur dengan K=5. Mendapat F1 Score 0.8571428571428571.

* Mengumpulkan Error Tiap K Yang Dicoba



Langkah ini untuk menentukan nilai k yang terbaik bedasarkan nilai error terkecil. Terlihat bahwa error terkecil dari 3 – 5.

* Evaluasi Performa



Mengevaluasi model menggunakan Confusion Matrix dengan memperhatikan accuracy, precision, sensitivity, specificity, dan f1 score,lalu menghasilkan :

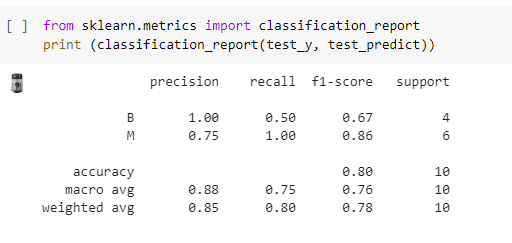
True Positive : 2

False Positive : 2

False Negative : 0

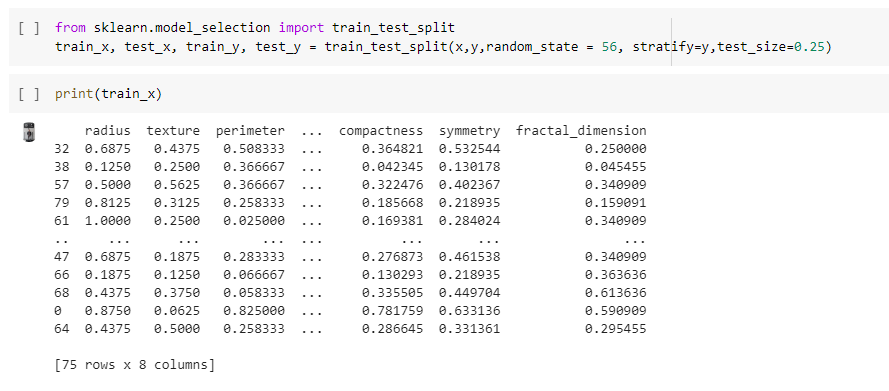
True Negative : 6

* Hasil Klasifikasi Dataset (90% dan 10%)

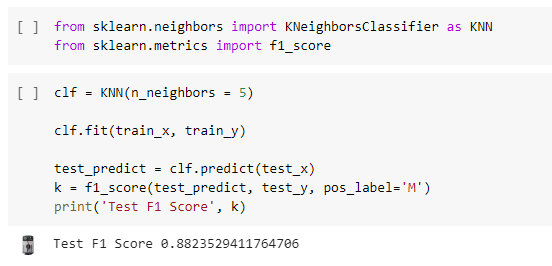


Diatas merupakan hasil klasifikasi terhadap dataset (90% dan 10%).

1. Pembagian Dataset (Data Training 75% dan Data Test 25%)

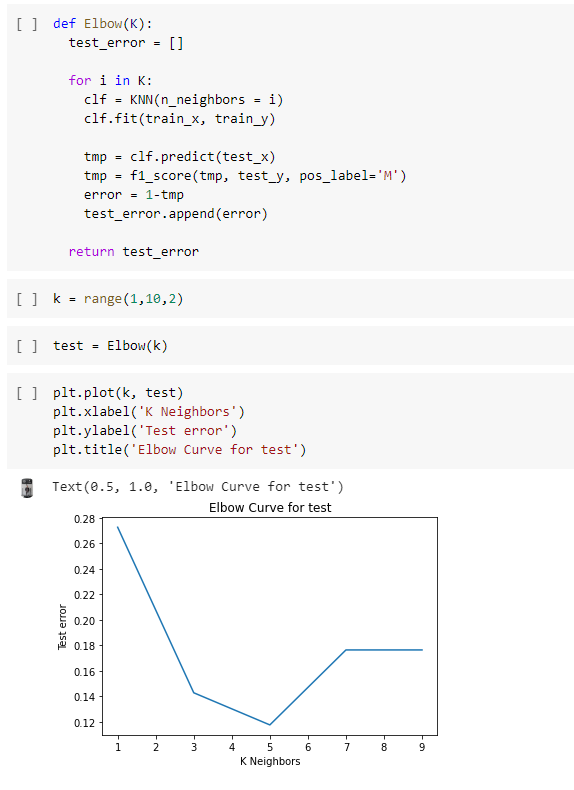


* Langkah ini dataset diatur menjadi data training sebanyak 75% dan data test 25%. Terdapat 75 baris dan 8 kolom.
* Menerapkan KNN Classifier



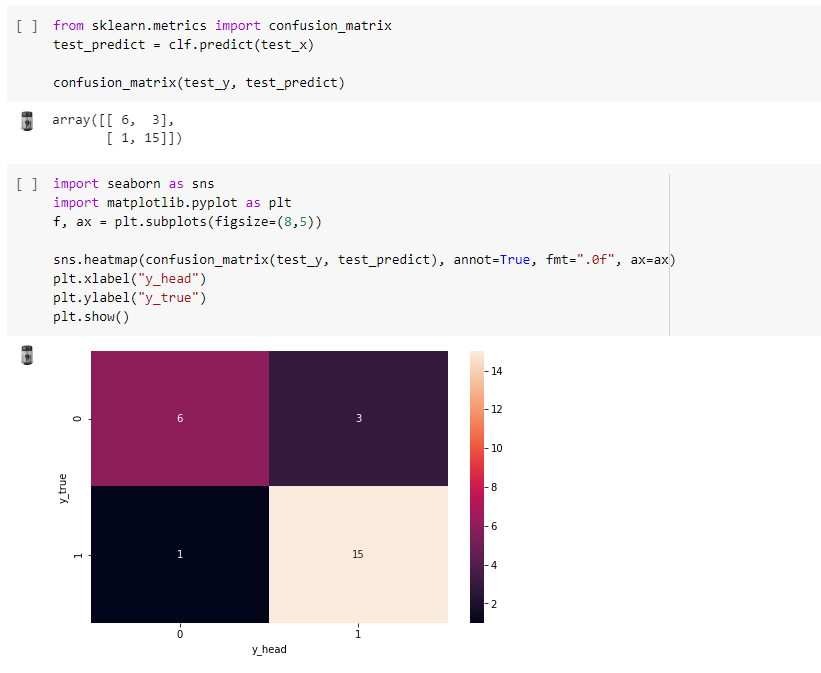
Menerapkan KNN Classifier pada dataset yang sudah diatur dengan K=5. Mendapatkan F1 Score 0.8823529411764706.

* Mengumpulkan Error Tiap K Yang Dicoba



Langkah ini untuk menentukan nilai k yang terbaik bedasarkan nilai error terkecil. Error terkecil adalah 5.

* Evaluasi Performa



Mengevaluasi model menggunakan Confusion Matrix dengan memperhatikan accuracy, precision, sensitivity, specificity, dan f1 scor menghasilkan :

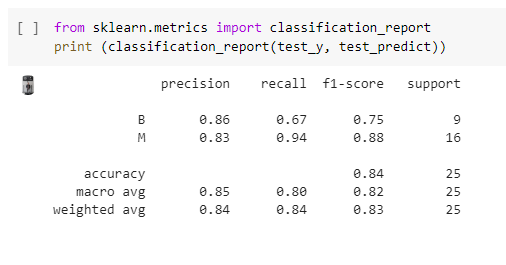
True Positive : 6

False Positive : 3

False Negative : 1

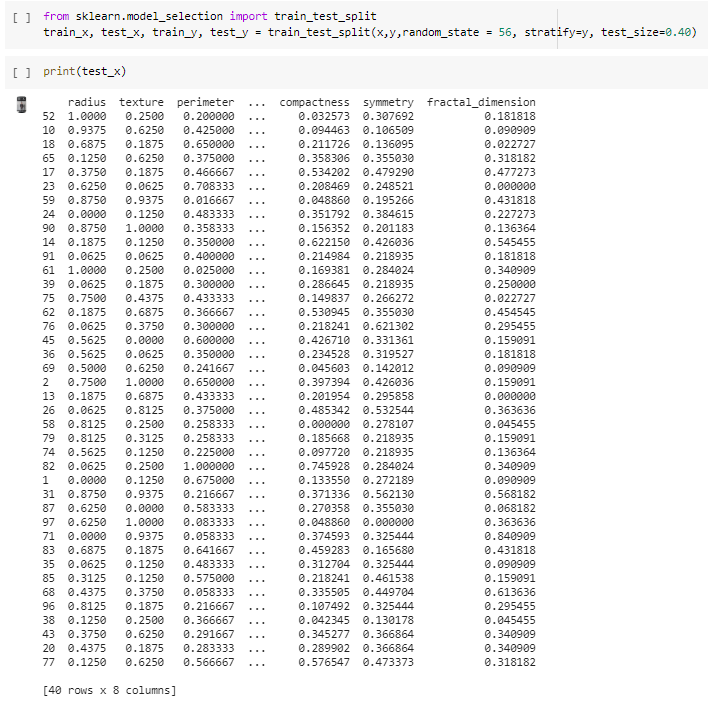
True Negative : 15

* Hasil Klasifikasi Dataset (75% dan 25%)



Diatas merupakan hasil klasifikasi terhadap dataset (90% dan 10%).

1. Pembagian Dataset (Data Training 60% dan Data Test 40%)

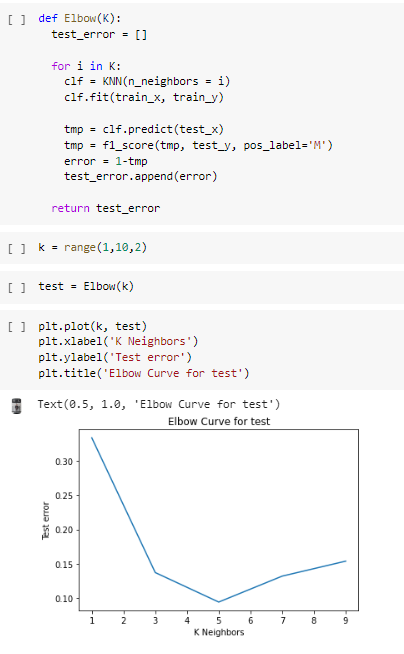


* Langkah ini dataset diatur menjadi data training sebanyak 60% dan data test 40%. Terdapat 40 baris dan 8 kolom.
* Menerapkan KNN Classifier



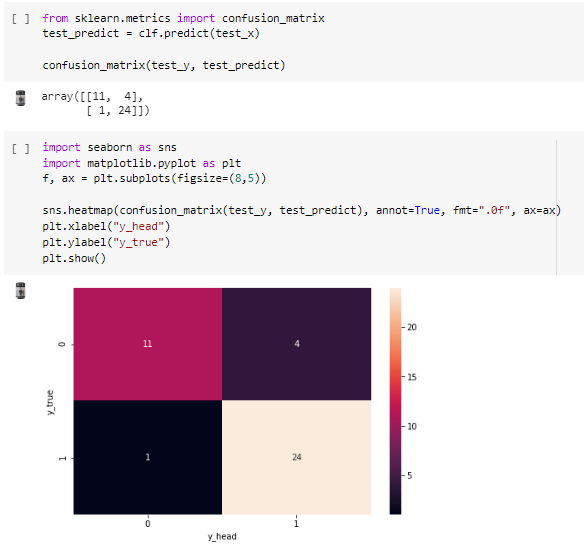
Menerapkan KNN Classifier pada dataset yang sudah diatur dengan K=5. Mendapatkan F1 Score 0.9056603773584904.

* Mengumpulkan Error Tiap K Yang Dicoba



Langkah ini untuk menentukan nilai k yang terbaik bedasarkan nilai error terkecil. Error terkecil adalah 5.

* Evaluasi Performa



Mengevaluasi model menggunakan Confusion Matrix dengan memperhatikan accuracy, precision, sensitivity, specificity, dan f1 score, lalu menghasilkan :

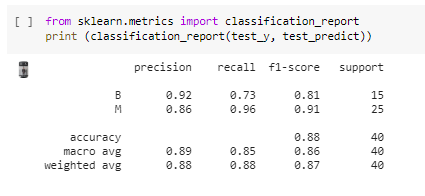
True Positive : 11

False Positive : 4

False Negative : 1

True Negative : 24

* Hasil Klasifikasi Dataset (75% dan 25%)

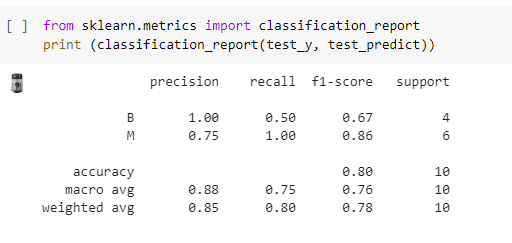
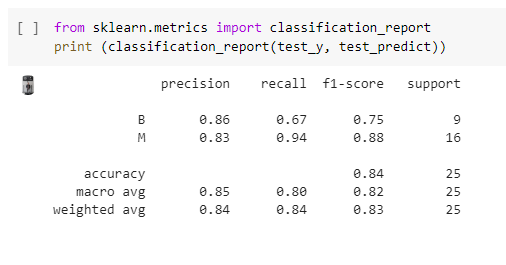
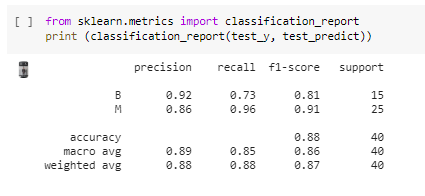


Diatas merupakan hasil klasifikasi terhadap dataset (90% dan 10%).

# BAB IV

# UJI COBA DAN PEMBAHASAN

## 4.1 Perbandingan hasil setiap percobaan

(90% dan 10%)

(75% dan 25%)

(60% dan 40%)

Untuk percobaan ke-3 dengan data training 60% dan data testing 40%, kita mendapatkan hasil yang paling tinggi tingkat akurasinya yakni sebesar 0,88. Begitu pula untuk nilai precision, recall, dan f1-scorenya menunjukkan nilai yang cukup tinggi dibandingkan dengan percobaan ke-1 dan ke-2.

Namun, percobaan di atas menghasilkan hasil yang kurang tepat. Karena, data training yang lebih sedikit (60% data training) memilki tingkat akurasi yang lebih tinggi dibandingkan dengan yang data training lebih banyak (75% dan 90% data training).

Hal ini dikarenakan :

1. Adanya parameter random\_state = 56 yang digunakan. Hal ini menyebabkan pembagian data testing dan data training tidak akan random / acak, sehingga output akurasi akan selalu sama.

Lalu, peneliti melakukan percobaan ulang random\_state = 56 dengan menggunakan data training sebagai berikut:

|  |  |
| --- | --- |
| Data Training | Hasil |
| 10% | 0.71 |
| 20% | 0.76 |
| 30% | 0.74 |
| 40% | 0.80 |
| 50% | 0.82 |
| 60% | 0.88 |
| 70% | 0.80 |
| 75% | 0.84 |
| 80% | 0.80 |
| 85% | 0.80 |
| 90% | 0.80 |
| 95% | 1 |

Dari data training yang digunakan, peneliti menemukan masalah yang ada saat melakukan percobaan.

## 4.2 Analisa Masalah

* Jumlah dan nilai dari tiap data testing berpengaruh
* random \_state menyebabkan pembagian dataset akan tetap sama
* Sehingga sebanyak apapun percobaan yang dilakukan angka akurasi akan tetap sama.
* Padahal nyatanya, setiap percobaan harus kita coba dengan data pembagian data yang berbeda.

Dengan adanya analisa masalah yang ditemukan, maka peneliti membuat solusi yang digunakan untuk permasalahan yang ditemukan.

## 4.3 Solusi Permasalahan

Menghapus parameter random\_state agar pembagian data training dan testing menjadi benar-benar random. Lalu, dilakukan percobaan 10 kali untuk tiap porsi uji.

1. 60% data training, 40 % data testing
2. 75% data training, 25% data testing
3. 90% data training, 10% data testing

Lalu menghitung rata-rata akurasi yang muncul tiap porsi uji. Nilai rata-rata inilah yang menunjukkan hasil akurasi percobaan secara lebih akurat.

Berikut merupakan table percobaan ulang sebanyak 10x tiap kasus.

* Tabel 60% Data Training



* Tabel 75% Data Training



* Tabel 90% Data Training



* Tabel Perbandingan Rata-Rata Data Training



* Diagram Batang Data Training

Dari hasil percobaan di atas, peneliti sudah menemukan hasil yang tepat. Di mana data training yang lebih banyak (90% data training) memiliki tingkat akurasi yang lebih tinggi dibandingkan data training yang lebih sedikit (75% dan 60% data training).

# BAB V

# KESIMPULAN DAN SARAN

## 5.1 Kesimpulan

Model yang saat ini dibuat berhasil memprediksi dengan tingkat akurasi mencapai 0.9 dengan menggunakan 90% data training. Semakin banyak data training yang digunakan maka, akurasi yang dihasilkan akan semakin tinggi.

## 5.2 Saran

Hal ini sudah cukup baik, namun akan jauh lebih baik lagi apabila disediakan lebih banyak data dari data yang saat ini digunakan (100 record) agar Prediksi Kanker Prostat dapat lebih akurat.

# DAFTAR PUSTAKA

Han, J.,&Kamber, M. (2006). *Data Mining Concept and Tehniques.San Fransisco : Morgan Kauffman.*

Kevin Adrian. (2019). *Fungsi Kelenjar Prostat dan Gangguan yang Bisa Menghantuinnya.* Dikutip 21 Juni 2020 dari Internet: <https://www.alodokter.com/fungsi-kelenjar-prostat-dan-gangguan-yang-bisa-menghantuinya>.

Sajid Saifi. (2018). *Prostate Cancer Dataset.* Dikutip 21 Juni 2020 dari Internet. https://www.kaggle.com/sajidsaifi/prostate-cancer/.

Wa Ode Dewiana. (2015). *Kanker Prostat.* Dikutip 21 Juni 2020 dari Internet: https://burangasitamaymo.wordpress.com/2015/06/26/makalah-kanker-prostat/.

Wikipedia. (2018). *Tumori.* Dikutip 21 Juni 2020 dari Internet: <https://id.wikipedia.org/wiki/Tumor#:~:text=Tumor%20atau%20barah%20(bahasa%20Inggris,latin%20yang%20berarti%20%22bengkak%22>.