KLASIFIKASI STADIUM KANKER TIROID DENGAN MENGGUNAKAN *MACHINE LEARNING* (STUDI KASUS: PASIEN KANKER TIROID DI RUMAH SAKIT ONKOLOGI SURABAYA)

Nanda Gita Aprilia[1] dan Jerry Dwi Trijoyo Purnomo[2]
Departemen Statistika, Fakultas Sains dan Analitika Data (FSAD),
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya, 60111, Indonesia
e-mail: nanda.gita22@gmail.com[1], jerry@statistika.its.ac.id[2]

Abstrak*—Kanker hingga saat ini masih menjadi masalah kesehatan dunia, termasuk di Indonesia. The International Agency for Research on Cancer (IARC) mengestimasi secara global, 1 dari 8 laki-laki dan 1 dari 11 perempuan meninggal karena kanker. Berdasarkan data Globocan tahun 2020, kanker tiroid merupakan kanker peringkat 5 tertinggi yang diderita oleh perempuan di seluruh dunia. Kanker tiroid menempati urutan ke-9 dari 10 kanker terbanyak di Indonesia, serta angkanya mengalami peningkatan tiap tahunnya. Pengklasifikasian stadium kanker tiroid penting dilakukan karena memengaruhi perawatan yang akan diberikan untuk kelangsungan hidup pasien. Tujuan dari penelitian ini adalah melakukan klasifikasi stadium kanker tiroid menggunakan metode Naïve Bayes dan k-Nearest Neighbor (k-NN). Data yang digunakan adalah data rekam medis pasien kanker tiroid di Rumah Sakit Onkologi Surabaya pada bulan Januari 2008 hingga Februari 2019 yang terdiri dari 141 pasien. Hasil klasifikasi dengan metode Naïve Bayes menghasilkan nilai AUC sebesar 0,7185 untuk data training dan 0,7326 untuk data testing, sedangkan metode k-NN dengan k = 7 menghasilkan nilai AUC sebesar 0,7405 untuk data training dan 0,8102 untuk data testing. Metode terbaik dipilih berdasarkan nilai AUC terbesar, sehingga metode k-NN dengan k = 7 dipilih sebagai metode terbaik untuk mengklasifikasikan stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya.

***Kata Kunci*— Kanker Tiroid, Klasifikasi, k-Nearest Neighbor, Naïve Bayes.**

I. PENDAHULUAN

Kanker merupakan salah satu penyebab kematian di dunia. Kanker terjadi karena adanya pembelahan sel yang tidak terkendali dan mampu menyerang jaringan biologis lainnya [1]. *The Union for International Cancer Control* dalam artikelnya menyebutkan bahwa berdasarkan estimasi *Globocan*, beban kanker global menunjukkan peningkatan menjadi 19,3 juta kasus dan 10 juta kematian akibat kanker pada tahun 2020 [2]. Di Indonesia, angka kejadian penyakit kanker berada pada urutan 8 di Asia Tenggara dan urutan 23 di Asia [3]. Kanker tiroid merupakan kanker pada kelenjar tiroid, yaitu kelenjar yang berada pada bagian depan leher sedikit di bawah laring berbentuk kupu-kupu. Menurut *Global Cancer Observatory* berdasarkan data *Globocan* tahun 2020, kanker tiroid merupakan kanker peringkat 5 tertinggi yang diderita oleh perempuan di seluruh dunia [4]. Berdasarkan registrasi Perhimpunan Doktor Spesialis Patologi Indonesia didapatkan bahwa kanker tiroid menempati urutan ke-9 dari 10 kanker terbanyak di Indonesia, yaitu sebesar 4,43% serta angkanya mengalami peningkatan tiap tahunnya [5].

Kanker tiroid diklasifikasikan untuk menunjukkan seberapa besar kanker berkembang dan menyebar di tubuh pasien, klasifikasi terbagi menjadi empat stadium, yaitu stadium I, II, III, dan IV. Menurut publikasi *National Comprehensive Cancer Network*, gejala indikasi adanya kanker tiroid antara lain adanya benjolan pada leher, sakit pada area leher, adanya perubahan suara, kesulitan bernapas, dan kesulitan menelan. Beberapa faktor risiko yang dapat memengaruhi kanker tiroid adalah riwayat penyakit kanker pada keluarga dan paparan radiasi [6]. Menurut *American Society of Clinical Oncology*, kanker tiroid dipengaruhi oleh faktor risiko jenis kelamin, usia, genetik, paparan radiasi, diet mengurangi iodin, ras, dan kanker payudara [7]. Shi, et al. [8] menyimpulkan bahwa status pernikahan dan usia merupakan faktor risiko yang dapat memengaruhi kanker tiroid, pasien berstatus janda dan berusia di atas 45 tahun memiliki kecenderungan untuk berada pada stadium III atau IV. Menurut Hegedus [9], faktor yang berguna untuk mencurigai keganasan tiroid adalah usia di bawah 20 tahun atau di atas 60 tahun, jenis kelamin laki-laki, riwayat radiasi di daerah leher, dan tumor dengan diameter lebih dari 4 cm. *National Cancer Institute* (NCI) memberikan rekomendasi perawatan atau *treatment* yang berbeda kepada pasien kanker tiroid berdasarkan stadium kankernya. Oleh karena itu, proses klasifikasi stadium kanker tiroid ini penting untuk dilakukan karena berpengaruh terhadap kelangsungan hidup pasien.

Mir dan Mittal [10] melakukan prediksi penyakit tiroid menggunakan *machine learning* menyimpulkan bahwa algoritma klasifikasi *Naïve Bayes* memiliki nilai akurasi yang lebih tinggi dibandingkan dengan algoritma J48 yang merupakan algoritma turunan dari C4.5 *Decision Tree*. Penelitian serupa juga dilakukan oleh Somali dan Shammari [11] dan menyimpulkan bahwa algoritma *Naïve Bayes* menghasilkan nilai akurasi yang lebih tinggi jika dibandingkan dengan algoritma *Sequential Minimal Optimization* (SMO) yang merupakan pengembangan dari *Support Vector Machine* (SVM). Shalini dan Ghalib [12] membandingkan algoritma *k-NN* dan *Neural Network* untuk memprediksi kelainan hipotiroidisme dan menyimpulkan bahwa algoritma *k-NN* menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan *Neural Network.*

Pada penelitian ini akan dilakukan klasifikasi stadium kanker tiroid menggunakan data Tugas Akhir Husna [13] tentang analisis regresi logistik ordinal pada faktor-faktor yang memengaruhi stadium kanker tiroid di Rumah Sakit Onkologi

Surabaya. Berdasarkan penelitian tersebut disimpulkan bahwa variabel yang berpengaruh signifikan terhadap stadium kanker tiroid adalah usia dan riwayat kanker pada keluarga pasien. Pasien kanker tiroid di Rumah Sakit Onkologi Surabaya cenderung memiliki risiko tidak terklasifikasi stadium I, II, dan III sebesar 1,089 kali dibandingkan dengan stadium IV pada setiap kenaikan satu tahun umur pasien dan pasien yang tidak memiliki riwayat kanker pada keluarga memiliki kecenderungan 6,593 kali lebih besar untuk terjadi peningkatan stadium kanker tiroid dibandingkan pasien yang memiliki riwayat kanker pada keluarga. Hasil penelitian tersebut menunjukkan bahwa model regresi logistik ordinal yang terbentuk menghasilkan nilai akurasi sebesar 57,45%. Pada penelitian ini diusulkan beberapa metode lain untuk mengklasifikasikan stadium kanker tiroid, yaitu metode *Naïve Bayes* dan *k-NN*. Kedua metode tersebut dibandingkan menggunakan kriteria ketepatan klasifikasi untuk mengetahui metode yang lebih baik digunakan untuk mengklasifikasikan stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya

Batasan masalah dalam penelitian ini adalah data yang digunakan merupakan data rekam medis pasien kanker tiroid bulan Januari 2008 hingga Februari 2019 di Rumah Sakit Onkologi Surabaya sebanyak 141 pasien. Pasien yang diamati merupakan pasien yang didiagnosis mengidap kanker tiroid mulai dari stadium I, II, III, dan IV berdasarkan diagnosis ketika pasien melakukan pengobatan di Rumah Sakit Onkologi Surabaya.

II. TINJAUAN PUSTAKA

A. *Statistika Deskriptif*

Statistika deskriptif adalah suatu metode statistik yang meringkas dan mendeskripsikan fitur-fitur penting dari suatu data [14]. Statistika deskriptif yang digunakan dalam penelitian ini adalah sebagai berikut.

1). Mean

Mean atau nilai rata-rata (*average*) dari suatu data adalah jumlah data dibagi dengan banyaknya data [14]. Persamaan (1) digunakan untuk menghitung *mean.*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

keterangan :

$\bar{x}$: Rata-rata
x_i : Data ke-*i*, *i* = 1, 2,…, *n*
n : Banyaknya data

2). Varians

Varians adalah suatu ukuran penyebaran data, dibangun dengan menjumlahkan simpangan (deviasi) kuadrat dan membaginya dengan total pengamatan dikurangi satu [14]. Varians merupakan kuadrat dari standar deviasi. Cara menghitung varians dijelaskan pada persamaan (2).

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2}$$

keterangan :

s^2 : Varians
$\bar{x}$: Rata-rata
x_i : Data ke-*i*, *i* = 1, 2,…, *n*
n : Banyaknya data

3). Minimum dan Maksimum

Nilai minimum merupakan nilai terendah atau terkecil dari suatu data sementara nilai maksimum merupakan nilai tertinggi atau terbesar dari suatu data [15].

4). Pie Chart

Visualisasi data digunakan untuk menampilkan atau memvisualisasikan data sehingga dapat lebih mudah dipahami. *Pie chart* atau diagram lingkaran adalah diagram di mana tiap segmen dalam lingkaran menunjukkan frekuensi relatif dari kategori [14].

5). Cross Tabulation

Cross tabulation atau tabulasi silang adalah metode statistik yang menggambarkan dua atau lebih variabel secara simultan dan hasilnya disajikan dalam bentuk tabel yang merefleksikan distribusi bersama dua atau lebih variabel dengan jumlah kategori yang terbatas [16].

B. *Klasifikasi*

Klasifikasi adalah suatu bentuk analisis data yang mengekstrak model yang mendeskripsikan kelas data. Klasifikasi terdiri dari dua tahap, tahap yang pertama adalah membangun model klasifikasi berdasarkan data sebelumnya (data *training*). Tahap yang kedua adalah menggunakan model tersebut untuk mengklasifikasikan data baru (data *testing*) untuk mengukur ketepatan klasifikasinya [17].

C. *Naïve Bayes*

Naïve Bayes merupakan salah satu metode klasifikasi sederhana menggunakan metode probabilitas yang didasarkan pada teorema *Bayes* [17]. Dimisalkan $\boldsymbol{X}$ adalah suatu data dan *H* merupakan hipotesis data $\boldsymbol{X}$ termasuk dalam suatu kelas *C* tertentu. Untuk permasalahan klasifikasi, ingin ditentukan $P(H|\boldsymbol{X})$, yaitu probabilitas hipotesis *H* terjadi atau probabilitas data $\boldsymbol{X}$ termasuk dalam suatu kelas *C*, dengan mengetahui deskripsi atribut dari $\boldsymbol{X}$. Nilai $P(H|\boldsymbol{X})$ dapat dihitung menggunakan persamaan (3).

$$P(H|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)P(H)}{P(\boldsymbol{X})} \tag{3}$$

Gorunescu [18] menuliskan persamaan (3) menjadi persamaan (4).

$$posterior = \frac{likelihood \times prior}{evidence} \tag{4}$$

Misalkan $\{x_1, x_2, \dots, x_p\}$ merupakan variabel-variabel yang digunakan untuk menentukan kelas *y*. Perhitungan *posterior probability* untuk setiap kelas y_i dapat ditulis menjadi persamaan (5).

$$P(y_i|x_1, x_2, \dots, x_p) = \frac{P(x_1, x_2, \dots, x_p|y_i)P(y_i)}{P(x_1, x_2, \dots, x_p)} \tag{5}$$

Kelas yang terpilih adalah kelas yang memaksimalkan nilai *posterior probability*. Karena $P(x_1, x_2, \dots, x_p)$ atau *evidence* bersifat konstan untuk semua kelas maka cukup memaksimalkan nilai dari $P(x_1, x_2, \dots, x_p|y_i)P(y_i)$. *Naïve Bayes* mengasumsikan bahwa efek dari atribut pada kelas tertentu tidak tergantung pada nilai atribut lainnya. Setiap variabel diasumsikan saling bebas untuk kelas *y* [17], sehingga $P(x_1, x_2, \dots, x_p|y_i)$ dapat ditulis menjadi $P(x_1|y_i)P(x_2|y_i)$ ... $P(x_p|y_i)$ seperti pada persamaan (6).

$$P(x_1, x_2, \dots, x_p|y_i)P(y_i) = P(x_1|y_i)P(x_2|y_i) \dots P(x_p|y_i).P(y_i) \quad (6)$$

Jika terdapat variabel yang bersifat kontinu atau kuantitatif, maka $P(x_k|y_i)$ dihitung menggunakan pendekatan distribusi normal seperti pada persamaan (7).

$$P(x_k|y_i) = \frac{1}{\sqrt{2\pi}\sigma_{ki}} \exp\left(-\frac{(x_k - \mu_{ki})^2}{2\sigma_{ki}^2}\right) \quad (7)$$

Estimasi peluang $P(x_k|y_i)$ dapat dihitung untuk setiap variabel x_k dengan $k = 1,2, \dots, p$ dan kelas y_i, sehingga data baru diklasifikasikan ke dalam kelas y_i apabila peluangnya paling besar dibandingkan yang lainnya.

D. *k-Nearest Neighbor (k-NN)*

Metode *k-Nearest Neighbor* atau *k-NN* adalah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pelatihan (*training*) yang jaraknya paling dekat dengan objek tersebut. Prinsip kerja dari *k-NN* adalah mencari jarak terdekat antara data yang akan dievaluasi dengan *k neighbor* terdekatnya dalam data pelatihan [19]. Peringkat *k neighbor* terdekat berdasarkan nilai kesamaan dihitung menggunakan jarak *Euclidean* (*Euclidean Distance*) yang didefinisikan pada persamaan (8) [17].

$$D(x_i, z_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - z_{jk})^2} \quad (8)$$

keterangan :

$D(x_i, z_j)$: Jarak *Euclidean*

x_i : Data *training* ke-*i*, dengan $i = 1, 2, \dots, n$

z_j : Data *testing* ke-*j*, dengan $j = 1, 2, \dots, m$

k : Banyaknya variabel independen, dengan $k = 1, 2, \dots, p$

Algoritma *k-NN* dituliskan sebagai berikut [20].

1. Menentukan parameter *k*
2. Menghitung jarak data *testing* dengan data *training* menggunakan jarak *Euclidean*
3. Mengurutkan jarak dari yang terdekat
4. Memeriksa kelas *k neighbor* terdekat
5. Kelas data *training* = kelas mayoritas *neighbor* terdekat.

E. *Stratifikasi*

Stratifikasi merupakan salah satu cara untuk pengambilan sampel atau *sampling*. Misalkan data *D* dibagi dengan bagian-bagian yang saling lepas, bagian-bagian data *D* itu disebut strata. Proses pengambilan sampel secara acak pada tiap-tiap strata itulah yang disebut stratifikasi. Stratifikasi membantu memastikan proses pengambilan sampel yang representatif untuk masing-masing strata, terutama pada data *imbalance* [17].

F. *Ketepatan Klasifiaksi*

Ketepatan klasifikasi merupakan evaluasi prosedur untuk melihat peluang kesalahan klasifikasi yang dilakukan oleh suatu fungsi klasifikasi [21]. Klasifikasi akan membentuk *confusion matrix* yang memuat ketepatan hasil klasifikasi. Pada klasifikasi empat kelas atau kategori, *confusion matrix* yang terbentuk adalah seperti Tabel 1.

Tabel 1. *Confusion Matrix* Empat Kategori

Aktual	Prediksi				Jumlah
	1	2	3	4	
1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
4	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$
Jumlah	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$

keterangan :

n_{ij} : Jumlah pengamatan dari kategori ke-*i* yang diklasifikasikan sebagai kategori ke-*j*, dengan $i = 1, 2, 3, 4$ dan $j = 1, 2, 3, 4$

Ketepatan klasifikasi dapat diukur menggunakan beberapa kriteria. Kriteria yang sering digunakan untuk mengukur ketepatan klasifikasi adalah akurasi. Akurasi merupakan perbandingan seluruh data yang diklasifikasikan dengan benar dengan banyaknya data uji dan tepat digunakan jika data *balance* pada tiap kategorinya [17]. Persamaan (9) digunakan untuk menghitung akurasi dari data dengan empat kategori.

$$Akurasi = \frac{n_{11} + n_{22} + n_{33} + n_{44}}{n_{..}} \quad (9)$$

Pada data *imbalance*, pengukuran ketepatan klasifikasi yang dapat digunakan adalah *Area Under Curve* (*AUC*). *AUC* merupakan indikator performasi kurva *Receiver Operating Characteristic* (*ROC*) yang dapat meringkas kinerja sebuah *classifier* menjadi suatu nilai [22]. Misalkan f_i adalah estimasi peluang kelas *i* pada data *testing* terklasifikasikan sebagai kelas *i* dan g_i adalah estimasi peluang kelas *j* terklasifikasikan sebagai kelas *i* maka $\{f_1, \dots, f_{ni}\}$ dan $\{g_1, \dots, g_{nj}\}$ adalah sampel-sampel dari suatu distibusi *f* dan *g*. Urutkan kombinasi nilai-nilai $\{f_1, \dots, f_{ni}, g_1, \dots, g_{nj}\}$ dengan *increasing order*. Misalkan r_i adalah urutan ke-*i* dari kelas *i* pada data *testing* maka S_i adalah jumlah seluruh r_i seperti pada persamaan (10). *AUC* untuk kelas *i* dan *j* dapat dihitung dengan persamaan (11).

$$S_i = \sum_{i=1}^{n_i} r_i \quad (10)$$

$$\hat{A}_{(i,j)} = \frac{S_i - n_i(n_i + 1)/2}{n_i n_j} \quad (11)$$

Dalam kasus jumlah kelas lebih dari dua, $\hat{A}_{(i,j)} \neq \hat{A}_{(j,i)}$ maka untuk mendapatkan $\hat{A}_{(i,j)}$ dihitung dari rata-rata kedua nilai tersebut seperti pada persamaan (12).

$$\hat{A}_{(i,j)} = \frac{\hat{A}_{(i,j)} + \hat{A}_{(j,i)}}{2} \quad (12)$$

Nilai *AUC* keseluruhan diperoleh dari rata-rata persamaan (12) dari seluruh pasangan kelas, yang dapat dituliskan menjadi persamaan (13). *C* merupakan jumlah kelas atau kategori yang digunakan [23].

$$AUC_{total} = \frac{2}{C(C-1)} \sum_{i<j} \hat{A}_{(i,j)} \quad (13)$$

Kriteria nilai *AUC* menurut Gorunescu [18] ditampilkan pada Tabel 2.

Tabel 2. Kriteria Nilai *AUC*

Nilai *AUC*	Keterangan
0,5 – 0,6	Kegagalan
0,6 – 0,7	Klasifikasi buruk
0,7 – 0,8	Klasifikasi cukup
0,8 – 0,9	Klasifikasi baik
0,9 – 1,0	Klasifikasi sangat baik

G. *Kanker Tiroid*

Kanker tiroid merupakan kanker pada kelenjar tiroid, yaitu kelenjar yang berada pada bagian depan leher sedikit di bawah laring berbentuk kupu-kupu. Terjadi 2,5% pada perempuan dan 0,85% pada laki-laki dari seluruh keganasan kanker tiroid dengan perbandingan 3:1. Umumnya, kanker tiroid paling sering muncul pada usia 20-50 tahun, namun kanker ini dapat terjadi pada semua usia [24]. *American Joint Committee on Cancer* (AJCC) mengklasifikasikan stadium kanker menggunakan TNM (*tumor*, *node*, *metastatis*) *stage*. T menggambarkan ukuran tumor dan apakah tumor tersebut telah menyerang jaringan sekitar, N menggambarkan kelenjar getah bening terdekat yang terlibat, dan M menggambarkan metastasis atau penyebaran ke bagian tubuh lainnya.

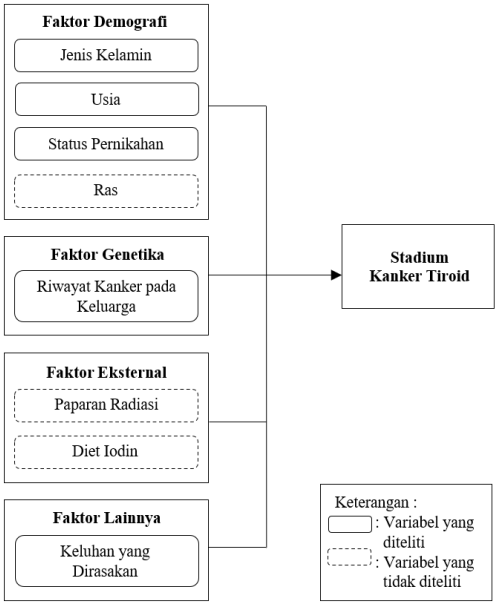
III. METODOLOGI PENELITIAN

A. *Sumber Data*

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Tugas Akhir Husna [13] tentang Analisis Regresi Logistik Ordinal pada Faktor-Faktor yang Memengaruhi Stadium Kanker Tiroid. Data merupakan data rekam medis 141 pasien kanker tiroid pada bulan Januari 2008 hingga Februari 2019 di Rumah Sakit Onkologi Surabaya.

B. *Kerangka Konsep*

Konsep yang digunakan dalam penelitian ini didasari oleh informasi yang didapat dari *National Comprehensive Cancer Network* (NCCN), *American Society of Clinical Oncology* (ASCO), serta penelitian lain yang berkaitan dengan gejala klinis dan faktor risiko yang dapat memengaruhi perkembangan kanker tiroid. Faktor lain yang dapat memengaruhi stadium kanker tiroid namun tidak digunakan dalam penelitian ini disebabkan karena keterbatasan data. Kerangka konsep pada penelitian ini ditampilkan pada Gambar 1.



Gambar 1. Kerangka Konsep

C. *Variabel Penelitian*

Variabel penelitian yang digunakan dalam penelian ini dirangkum dalam Tabel 3.

Tabel 3. Variabel Penelitian

Variabel	Keterangan	Skala
X_1	Jenis Kelamin Pasien 1 = Laki-laki 2 = Perempuan	Nominal
X_2	Usia Pasien	Rasio
X_3	Status Pernikahan Pasien 1 = Lajang 2 = Menikah 3 = Duda atau Janda	Ordinal
X_4	Keluhan yang Dirasakan Pasien 1 = Tanpa Keluhan 2 = Benjolan pada Leher/Tiroid 3 = Benjolan pada Payudara 4 = Lainnya (Pusing, Nyeri, Suara Serak, dan Lainnya)	Nominal
X_5	Riwayat Kanker dalam Keluarga Pasien 1 = Tidak Ada 2 = Ada	Nominal
Y	Stadium Kanker Tiroid 1 = Stadium 1 2 = Stadium II 3 = Stadium III 4 = Stadium IV	Ordinal

D. *Struktur Data*

Data penelitian terdiri dari lima variabel independen (X) dan satu variabel dependen (Y) serta terdapat 141 data observasi. Struktur data untuk penelitian ini dapat dilihat pada Tabel 4.

Tabel 4. Struktur Data

Pasien ke-	X_1	X_2	X_3	X_4	X_5	Y
1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	Y_1
2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$	$X_{2,5}$	Y_2
3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,4}$	$X_{3,5}$	Y_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
141	$X_{141,1}$	$X_{141,2}$	$X_{141,3}$	$X_{141,4}$	$X_{141,5}$	Y_{141}

E. *Langkah Analisis*

Langkah analisis yang dilakukan dalam penelitian ini adalah sebagai berikut.

1. Mendeskripsikan data pasien kanker tiroid di Rumah Sakit Onkologi Surabaya.
2. Melakukan stratifikasi pada data kemudian membagi data menjadi 90% data *training* dan 10% data *testing*.
3. Melakukan klasifikasi menggunakan variabel independen yang signifikan terhadap stadium kanker tiroid dengan metode *Naïve Bayes* sebagai berikut.
   a. Menghitung nilai *prior probability* variabel stadium kanker tiroid data *training.*
   b. Menghitung nilai *likelihood* dari setiap variabel independen pada masing-masing stadium kanker tiroid pada data *training*. Apabila variabel independen bersifat numerik maka menghitung nilai rata-rata dan standar deviasinya, lalu menghitung nilai *likelihood* menggunakan pendekatan distribusi normal.
   c. Menghitung *posterior probability* pada data *testing* untuk setiap kategori.
   d. Menentukan kelompok kategori berdasarkan nilai *posterior probability* tertinggi.
   e. Menghitung ketepatan klasifikasi untuk data *training* dan data *testing*.

4. Melakukan klasifikasi menggunakan variabel independen yang signifikan terhadap stadium kanker tiroid dengan metode *k-NN* sebagai berikut.
   a. Menghitung jarak antara data *training* dengan data *testing* menggunakan jarak *Euclidean*.
   b. Mengurutkan jarak mulai dari yang terdekat.
   c. Menentukan nilai *k*, mulai dari 3 hingga 10.
   d. Memerhatikan urutan jarak terdekat sampai pada nilai *k*.
   e. Mengklasifikasikan data berdasarkan jumlah kategori terbanyak.
   f. Mengitung ketepatan klasifikasi untuk data *training* dan data *testing*.
   g. Mendapatkan nilai *k* terbaik berdasarkan kesalahan klasifikasi paling rendah dan ketepatan klasifikasi tertinggi.
5. Membandingkan ketepatan klasifikasi antara metode *Naïve Bayes* dan *k-NN* untuk mendapatkan metode terbaik.
6. Membandingkan ketepatan klasifikasi metode terbaik dengan penelitian sebelumnya.
7. Menarik kesimpulan dan membuat saran untuk penelitian selanjutnya.

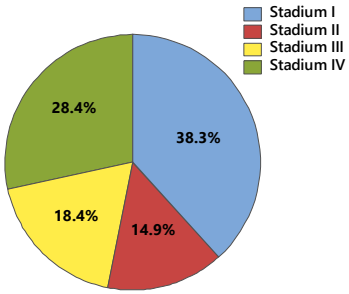
IV. ANALISIS DAN PEMBAHASAN

A. Karakteristik Data

Berikut ini karakteristik data pasien kanker tiroid di Rumah Sakit Onkologi Surabaya.

1). Karakteristik Stadium Pasien Kanker Tiroid

Pada penelitian ini, stadium pasien kanker tiroid yang diamati antara lain stadium I, stadium II, stadium III, dan stadium IV.



Gambar 2. Karakteristik Stadium Pasien Kanker Tiroid

Gambar 2 menunjukkan bahwa stadium kanker tiroid pasien yang menjalani pengobatan di Rumah Sakit Onkologi Surabaya didominasi oleh stadium dengan tingkat keparahan paling rendah dan paling tinggi, yaitu stadium I dan stadium IV.

2). Karakteristik Usia Pasien Kanker Tiroid

Usia pasien yang diteliti adalah usia pasien saat didiagnosis mengidap kanker tiroid. Berdasarkan Tabel 5, dapat diketahui bahwa dari keseluruhan pasien kanker tiroid di Rumah Sakit Onkologi Surabaya, rata-rata berusia 48 tahun. Usia pasien memiliki tingkat keragaman yang tinggi, yaitu mulai dari usia 16 hingga 86 tahun. Pasien yang didiagnosis pada stadium akhir memiliki rata-rata usia lebih tua dari pada pasien yang didiagnosis pada stadium awal. Hal ini dapat dikaitkan dengan peningkatan risiko keparahan penyakit seiring dengan bertambahnya usia.

Tabel 5. Karakteristik Usia Pasien Kanker Tiroid

Stadium	*Mean*	Varians	Min	Maks
Stadium I	38,83	10.88	16	65
Stadium II	45,52	14.97	17	76
Stadium III	56,69	14.56	18	86
Stadium IV	58,85	12.29	33	83
Keseluruhan	47,65	262,34	16	86

3). Karakteristik Jenis Kelamin Pasien Kanker Tiroid

Persentase jenis kelamin pasien kanker tiroid di Rumah Sakit Onkologi Surabaya dapat dilihat pada Gambar 4.2. Gambar tersebut menunjukkan bahwa pasien berjenis kelamin perempuan lebih banyak dijumpai daripada laki-laki dengan perbandingan 7:3. Hal ini sesuai dengan Handayani dan Purnami [24] yang menyatakan bahwa kanker tiroid lebih banyak terjadi pada perempuan dari pada laki-laki. Laki-laki paling banyak didiagnosis stadium IV sedangkan perempuan paling banya didiagnosis stadium I. Hal ini sesuai dengan Hegedus [9] yang menyimpulkan bahwa salah satu faktor yang berguna untuk mencurigai keganasan tiroid adalah jenis kelamin laki-laki.

Tabel 6. Karakteristik Jenis Kelamin Pasien Kanker Tiroid

Jenis Kelamin	Stadium				Jumlah
	I	II	III	IV	
Laki-laki	11 (7,8%)	8 (5,7%)	9 (6,4%)	14 (9,9%)	42 (29,8%)
Perempuan	43 (30,5%)	13 (9,2%)	17 (12,1%)	26 (18,4%)	99 (70,2%)
Jumlah	54 (38,3%)	21 (14,9%)	26 (18,4%)	40 (28,4%)	141 (100%)

4). Karakteristik Status Pernikahan Pasien Kanker Tiroid

Status pernikahan pasien yang diteliti adalah status pernikahan pasien saat didiagnosis mengidap kanker tiroid. Berdasarkan Tabel 7, diketahui bahwa pasien kanker tiroid di Rumah Sakit Onkologi Surabaya mayoritas berstatus menikah. Ada cukup banyak pasien berstatus duda atau janda yang didiagnosis stadium IV. Hal ini sesuai dengan penelitian yang dilakukan oleh Shi, et al. [8] yang menyimpulkan bahwa pasien berstatus janda memiliki kecenderungan untuk berada pada stadium III atau IV. Pasien berstatus menikah paling banyak didiagnosis stadium I. Hal ini dapat dikaitkan dengan peran keluarga, khususnya pasangan sebagai *support system* bagi pasien yang menderita kanker [25].

Tabel 7. Karakteristik Status Pernikahan Pasien Kanker Tiroid

Status Pernikahan	Stadium				Jumlah
	I	II	III	IV	
Lajang	9 (6,4%)	3 (2,1%)	1 (0,7%)	2 (1,4%)	15 (10,6%)
Menikah	45 (31,9%)	14 (9,9%)	22 (15,6%)	34 (24,1%)	115 (81,6%)
Duda atau Janda	0 (0,0%)	4 (2,8%)	3 (2,1%)	4 (2,8%)	11 (7,8%)
Jumlah	54 (38,3%)	21 (14,9%)	26 (18,4%)	40 (28,4%)	141 (100%)

5). Karakteristik Keluhan Pasien Kanker Tiroid

Keluhan pasien kanker tiroid dibagi dalam empat kategori dan ditampilkan pada Tabel 8. Mayoritas pasien kanker tiroid di Rumah Sakit Onkologi Surabaya memiliki keluhan adanya benjolan pada leher/tiroid. Benjolan pada leher pada kelenjar tiroid berkaitan dengan nodul tiroid, yaitu istilah dari kelenjar tiroid yang membesar. Pada umumnya, penyebab benjolan ini berkaitan dengan salah satu faktor seperti kanker tiroid [26].

Tabel 8. Karakteristik Keluhan Pasien Kanker Tiroid

Keluhan	Stadium				Jumlah
	I	II	III	IV	
Tanpa Keluhan	10 (7,1%)	3 (2,1%)	2 (1,4%)	5 (3,5%)	20 (14,2%)
Benjol pada Leher/Tiroid	35 (24,8%)	17 (12,1%)	21 (14,9%)	32 (22,7%)	105 (74,5%)
Benjol pada Payudara	1 (0,7%)	0 (0,0%)	1 (0,7%)	2 (1,4%)	4 (2,8%)
Lainnya	8 (5,7%)	1 (0,7%)	2 (1,4%)	1 (0,7%)	12 (8,5%)
Jumlah	54 (38,3%)	21 (14,9%)	26 (18,4%)	40 (28,4%)	141 (100%)

6). Karakteristik Riwayat Kanker pada Keluarga Pasien Kanker Tiroid

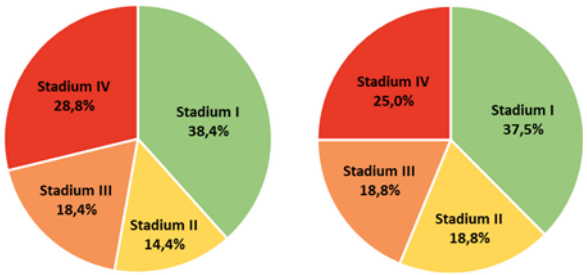
Riwayat kanker pada keluarga merupakan salah satu faktor genetika yang berperan mengatur pertumbuhan dan perkembangan sel-sel dalam tubuh. Tabel 9 menunjukkan bahwa dari total 141 pasien, hanya terdapat 10,6% pasien kanker tiroid di Rumah Sakit Onkologi Surabaya yang memiliki riwayat kanker pada keluarganya. *American Thyroid Association* (ATA) menjelaskan bahwa hampir seluruh kasus karsinoma tiroid papiler hanya menyerang satu individu dalam keluarga tanpa ada riwayat anggota keluarga lain yang juga menderita keganasan pada tiroid [27].

Tabel 9. Karakteristik Riwayat Kanker pada Keluarga Pasien Kanker Tiroid

Riwayat Kanker pada Keluarga	Stadium				Jumlah
	I	II	III	IV	
Tidak Ada	42 (29,8%)	20 (14,2%)	25 (17,7%)	39 (27,7%)	126 (89,4%)
Ada	12 (8,5%)	1 (0,7%)	1 (0,7%)	1 (0,7%)	15 (10,6%)
Jumlah	54 (38,3%)	21 (14,9%)	26 (18,4%)	40 (28,4%)	141 (100%)

B. Pembagian Data Training dan Data Testing

Sebelum melakukan klasifikasi, data penelitian terlebih dahulu dipartisi menjadi data *training* dan data *testing.* Gambar 2 menunjukkan bahwa pada data penelitian, tiap stadium kanker tiroid memiliki persentase yang berbeda-beda. Oleh karena itu, perlu dilakukan stratifikasi agar didapatkan sampel yang representatif dari tiap stadium saat pembagian data *training* sebesar 90% dan data *testing* sebesar 10% dari keseluruhan data penelitian.



(a) Persentase Stadium pada Data *Training* (b) Persentase Stadium pada Data *Testing*

Gambar 3. Persentase Stadium pada Data *Training* dan *Testing*

Berdasarkan Gambar 3, dapat diketahui bahwa stadium kanker tiroid pada data *training* dan data *testing* telah menunjukkan persentase-persentase yang kurang lebih sama seperti data penelitian.

C. Analisis Klasifikasi Stadium Kanker Tiroid dengan Metode Naïve Bayes

Analisis klasifikasi menggunakan metode *Naïve Bayes* dilakukan dengan memerhatikan peluang yang didapat dari perhitungan menggunakan variabel independen. Hasil klasifikasi stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya pada data *testing* ditampilkan sebagai berikut.

Tabel 10. *Confusion Matrix* Data *Testing* Metode *Naïve Bayes*

Aktual	Prediksi				Jumlah
	Stadium I	Stadium II	Stadium III	Stadium IV	
Stadium I	5	0	0	1	6
Stadium II	2	0	0	1	3
Stadium III	0	0	0	3	3
Stadium IV	0	1	0	3	4
Jumlah	7	1	0	8	16

Berdasarkan Tabel 10 dapat diamati bahwa klasifikasi data *testing* menggunakan metode *Naïve Bayes* telah mampu tepat memprediksi 5 dari 6 pasien stadium I dan 3 dari 4 pasien stadium IV. Kriteria ketepatan klasifikasi yang digunakan dalam penelitian ini adalah *AUC.* Nilai *AUC* yang dihasilkan untuk klasifikasi data *testing* sebesar 0,7326. Nilai ini menunjukkan bahwa klasifikasi data *testing* dengan metode *Naïve Bayes* telah termasuk dalam klasifikasi cukup atau *acceptable.*

Tahap selanjutnya adalah melakukan klasifikasi untuk data *training* dengan langkah-langkah yang sama namun menggunakan keseluruhan variabel independen pada data *training.* Hasil klasifikasi data *training* ditampilkan sebagai berikut.

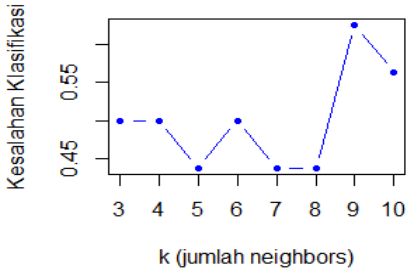
Tabel 11. *Confusion Matrix* Data *Training* Metode *Naïve Bayes*

Aktual	Prediksi				Jumlah
	Stadium I	Stadium II	Stadium III	Stadium IV	
Stadium I	43	1	1	3	48
Stadium II	8	4	0	6	18
Stadium III	5	0	0	18	23
Stadium IV	8	0	1	27	36
Jumlah	64	5	2	54	125

Tabel 11 menunjukkan klasifikasi data *training* menggunakan metode *Naïve Bayes* telah mampu tepat memprediksi 43 dari 48 pasien stadium I, 4 dari 18 pasien stadium II, dan 27 dari 36 pasien stadium IV. Nilai *AUC* yang dihasilkan untuk klasifikasi data *training* sebesar 0,7185. Nilai ini juga termasuk dalam klasifikasi cukup atau *acceptable.*

D. Analisis Klasifikasi Stadium Kanker Tiroid dengan Metode k-Nearest Neighbor

Analisis klasifikasi menggunakan metode *k-Nearest Neighbor* (*k-NN*) dilakukan dengan memerhatikan jarak terdekat antara data *testing* dengan *k neighbor* terdekatnya dalam data *training.* Tidak ada penentuan khusus untuk nilai *k* yang digunakan dalam proses klasifikasi menggunakan metode *k-NN.* Oleh karena itu, dalam analisis ini digunakan nilai *k* mulai dari 3 hingga 10 untuk melakukan prediksi data *testing* kemudian dihitung kesalahan klasifikasinya untuk masing-masing nilai *k.* Nilai *k* yang memiliki kesalahan klasifikasi terkecil yang akan digunakan.



Gambar 4. Perbandingan Nilai *k* dan Kesalahan Klasifikasinya

Gambar 4 menunjukkan bahwa nilai *k* yang memiliki kesalahan klasifikasi terkecil adalah *k* = 5, 7, dan 8. Selanjutnya digunakan nilai *AUC* untuk memilih *k* terbaik dari ketiga nilai *k* tersebut.

Tabel 12. Perbandingan *AUC* Masing-Masing Nilai *k*

k	Data *Training*	Data *Testing*
	AUC	***AUC***
5	0,7620	0,7037
7	0,7405	0,8102
8	0,7634	0,7130

Berdasarkan Tabel 12, nilai *AUC* yang dihasilkan telah termasuk dalam kategori cukup atau *acceptable* dan pada data *testing* dengan *k* = 7 nilai *AUC* yang dihasilkan termasuk dalam kategori klasifikasi baik. Nilai *k* = 7 menghasilkan nilai *AUC* tertinggi baik pada data *testing* jika dibandingkan dengan nilai *k* = 5 dan 8 maka digunakan *k* = 7 pada klasifikasi stadium kanker tiroid menggunakan metode *k-NN*. Hasil klasifikasi pada data *testing* dengan nilai *k* = 7 ditampilkan sebagai berikut.

Tabel 13. *Confusion Matrix* Data *Testing* Metode *k-NN*, *k* = 7

Aktual	Prediksi				Jumlah
	Stadium I	Stadium II	Stadium III	Stadium IV	
Stadium I	4	0	1	1	6
Stadium II	2	0	1	0	3
Stadium III	0	0	1	2	3
Stadium IV	0	0	0	4	4
Jumlah	6	0	3	7	16

Berdasarkan Tabel 13, dapat diamati bahwa klasifikasi data *testing* menggunakan metode *k-NN* dengan *k* = 7 telah mampu tepat memprediksi 4 dari 6 pasien stadium I, 1 dari 3 pasien stadium III, dan seluruh pasien stadium IV. Hasil klasifikasi pada data *training* dengan nilai *k* = 7 ditampilkan pada Tabel 14. Dapat diamati pada Tabel 14 bahwa klasifikasi data *training* menggunakan metode *k-NN* dengan *k* = 7 telah mampu tepat memprediksi 40 dari 48 pasien stadium I, 4 dari 18 pasien stadium II, 2 dari 23 pasien stadium III, dan 28 dari 36 pasien stadium IV.

Tabel 14. *Confusion Matrix* Data *Training* Metode *k-NN*, *k* = 7

Aktual	Prediksi				Jumlah
	Stadium I	Stadium II	Stadium III	Stadium IV	
Stadium I	40	2	1	5	48
Stadium II	7	4	1	6	18
Stadium III	2	3	2	16	23
Stadium IV	5	1	2	28	36
Jumlah	54	10	6	55	125

E. Perbandingan Hasil Ketepatan Klasifikasi Metode Naïve Bayes dan k-Nearest Neighbor

Metode yang lebih baik digunakan adalah metode yang memiliki nilai *AUC* tertinggi, baik pada data *training* maupun data *testing*. Berikut perbandingan nilai *AUC* antar metode *Naïve Bayes* dan *k-NN*.

Tabel 15. Perbandingan *AUC* Antara Metode *Naïve Bayes* dan *k-NN*

Metode	Data *Training*	Data *Testing*
	AUC	***AUC***
Naïve Bayes	0,7185	0,7326
k-NN (*k* = 7)	0,7405	0,8102

Tabel 15 menunjukkan bahwa klasifikasi menggunakan metode *k-NN* dengan *k* = 7 memiliki nilai *AUC* yang lebih tinggi dari pada metode *Naïve Bayes*, sehingga dapat dikatakan metode yang lebih baik digunakan untuk mengklasifikasikan stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya adalah metode *k-NN* dengan *k* = 7.

F. Perbandingan Hasil Ketepatan Metode Terbaik dengan Penelitian Sebelumnya

Klasifikasi stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya pada penelitian sebelumnya menggunakan metode regresi logistik ordinal pada keseluruhan data menghasilkan nilai akurasi sebesar 57,45%. Pada penelitian ini, metode terbaik adalah metode *k-NN* dengan *k* = 7 maka akan dilakukan klasifikasi stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya menggunakan metode *k-NN* dengan *k* = 7 pada keseluruhan data. Hasil klasifikasi ini ditampilkan pada *confusion matrix* berikut.

Tabel 16. *Confusion Matrix* Data Keseluruhan Metode *k-NN*, *k* = 7

Aktual	Prediksi				Jumlah
	Stadium I	Stadium II	Stadium III	Stadium IV	
Stadium I	44	2	2	6	54
Stadium II	9	4	2	6	21
Stadium III	2	1	7	16	26
Stadium IV	4	1	3	32	40
Jumlah	59	8	14	60	141

Berdasarkan Tabel 16, nilai akurasi dapat dihitung menggunakan persamaan (9) sebagai berikut.

$$Akurasi = \frac{n_{11} + n_{22} + n_{33} + n_{44}}{n_{..}} = \frac{44 + 4 + 7 + 32}{141} = 0{,}6170$$

Klasifikasi stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya menggunakan metode *k-NN* dengan *k* = 7 pada keseluruhan data menghasilkan nilai akurasi sebesar 61,70%, nilai akurasi ini lebih tinggi jika dibandingkan dengan hasil penelitian sebelumnya.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan analisis yang telah dilakukan, diperoleh kesimpulan sebagai berikut.

1. Pasien kanker tiroid yang melakukan pengobatan di Rumah Sakit Onkologi Surabaya paling banyak didiagnosis pada stadium I dan stadium IV. Rata-rata pasien berusia 48 tahun, dengan usia paling muda 16 tahun dan paling tua 86 tahun. Pasien berjenis kelamin perempuan lebih banyak dijumpai daripada laki-laki dengan perbandingan 7:3. Mayoritas pasien memiliki status menikah, dengan persentase sebesar 81,6%. Keluhan yang paling banyak dialami oleh pasien adalah adanya benjolah pada leher/tiroid, yaitu sebesar 74,5%. Pasien yang memiliki riwayat kanker pada keluarganya hanya sebesar 10,6% dari keseluruhan pasien.

2. Hasil klasifikasi stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya dengan metode *Naïve Bayes* menghasilkan nilai *AUC* sebesar 0,7185 untuk data *training* dan 0,7326 untuk data *testing.* Sedangkan metode *k-NN* dengan $k = 7$ menghasilkan nilai *AUC* sebesar 0,7405 untuk data *training* dan 0,8102 untuk data *testing.* Metode terbaik dipilih berdasarkan nilai *AUC* terbesar maka metode *k-NN* dengan $k = 7$ dipilih sebagai metode terbaik untuk mengklasifikasikan stadium kanker tiroid pasien di Rumah Sakit Onkologi Surabaya. Klasifikasi menggunakan metode *k-NN* dengan $k = 7$ menghasilkan nilai akurasi yang lebih tinggi jika dibandingkan dengan penelitian sebelumnya.

B. Saran

Berdasarkan hasil analisis yang telah dilakukan, berikut saran yang dapat diberikan kepada peneliti selanjutnya dan masyarakat.

1. Bagi Peneliti Selanjutnya

   Menambahkan metode klasifikasi lain untuk mendapatkan nilai *AUC* yang lebih baik serta mempertimbangkan adanya pengaruh dari *imbalance data* terhadap hasil klasifikasi.

2. Bagi Masyarakat

   Meningkatkan kesadaran tentang bahaya kanker tiroid dan melakukan tindakan preventif jika mengalami gejala-gejala yang mengarah pada kanker tiroid, seperti benjolan pada leher agar lebih cepat ditangani dengan perawatan yang tepat.

DAFTAR PUSTAKA

[1] S. S. Sunaryati, 14 Penyakit Paling Sering Menyerang dan Mematikan, Yogyakarta: Flash Books, 2011.

[2] UICC, "GLOBOCAN 2020: New Global Cancer Data," 2020. [Online]. Available: https://www.uicc.org/news/globocan-2020-new-global-cancer-data. [Accessed 5 Februari 2021].

[3] Kemenkes, "Kementrian Kesehatan Republik Indonesia - Hari Kanker Sedunia 2019," 31 Januari 2019. [Online]. Available: https://www.kemkes.go.id/article/view/19020100003/hari-kanker-sedunia-2019.html. [Accessed 5 Februari 2021].

[4] GCO, "Estimated Number of New Cases in 2020, Worldwide, Females, All Ages," 2020. [Online]. Available: https://gco.iarc.fr/today/online-analysis-pie. [Accessed 5 Februari 2021].

[5] Oktahermoniza, W. A. Harahap, Tofrizal and R. Rasyid, "Analisis Ketahanan Hidup Lima Tahun Kanker Tiroid yang Dikelola di RSUP Dr. M. Djamil Padang," *Jurnal Kesehatan Andalas,* vol. 2, no. 3, pp. 151-157, 2013.

[6] NCCN, "NCCN Guidelines for Patients Thyroid Cancer," 2020. [Online]. Available: https://www.nccn.org/patients/guidelines/content/PDF/thyroid-patient.pdf. [Accessed 5 Februari 2021].

[7] ASCO, "Thyroid Cancer : Risk Factor," 2019. [Online]. Available: https://www.cancer.net/cancer-types/thyroid-cancer/risk-factors. [Accessed 5 Februari 2021].

[8] R.-l. Shi, N. Qu, Z.-w. Lu, T. Liao, Y. Gao and Q.-h. Ji, "The Impact of Marital Status at Diagnosis on Cancer Survival in Patients with Differentiated Thyroid Cancer," *National Center of Biotechnology Information,* vol. 5, no. 8, pp. 2145-2154, 2016.

[9] L. Hegedus, "Clinical Practice. The Thyroid Nodule," *The New England Journal of Medicine,* vol. 351, no. 17, pp. 1764-1771, 2004.

[10] Y. I. Mir and S. Mittal, "Thyroid Disease Prediction Using Hybrid Machine Learning Techniques: An Effective Framework," *International Journal of Scientific and Technology,* vol. 9, no. 2, pp. 2868-2874, 2020.

[11] W. A. Somali and R. A. Shammari, "Comparison of Algorithms in Data Mining of Thyroid Disease Datasets," in *HEALTHINF 2018 - 11th International Conference on Health Informatics*, Riyadh, 2018.

[12] L. Shalini and M. R. Ghalib, "A Hypothyroidism Prediction Using Supervised Algorithm," *International Journal of Engineering and Advanced Technology (IJEAT),* vol. 9, no. 1, pp. 7285-7288, 2019.

[13] A. A. S. Husna, "Analisis Regresi Logistik Ordinal pada Faktor-Faktor yang Mempengaruhi Stadium Kanker Tiroid," ITS, Surabaya, 2020.

[14] R. A. Johnson and G. K. Bhattacharyya, Statistics Principles & Methods 6th Edition, New York: John Wiley & Sons Inc, 2006.

[15] Walpole, Pengantar Statistika Edisi ke 3, Jakarta: Gramedia Pustaka Utama, 1995.

[16] A. Agresti, Categorical Data Analysis, New York: Inc. John Wiley and Sons, 2002.

[17] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, 3rd ed., USA: Elsevier Inc., 2012.

[18] F. Gorunescu, Data Mining: Concepts, Models and Techniques, Jerman: Springer, 2011.

[19] J. Sreemathy and P. S. Balamurugan, "An Efficient Text Classification using KNN and Naive Bayesian," *International Journal on Computer Science and Engineering (IJCSE),* vol. 4, no. 3, pp. 392-396, 2012.

[20] A. G. Karegowda, M. A. Jayaram and A. S. Manjunath, "Cascading K-Means Clustering and k-Nearest Neighbor Classifier for Categorization of Diabetic Patients," *International Journal of Engineering and Advanced Technology (IJEAT),* vol. 1, no. 3, pp. 147-151, 2012.

[21] R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis 6th Edition, New Jersey: Prentice Hall Inc, 2007.

[22] M. Bekkar, D. K. Djemaa and D. A. Alitouche, "Evaluation Measures for Model Assesment over Imbalanced Data Sets," *Journal of Information Engineering and Application,* vol. 3, no. 10, pp. 27-38, 2013.

[23] D. J. Hand and R. J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," *Machine Learning,* vol. 45, no. 2, pp. 171-186, 2001.

[24] S. H. S. Handayani and S. W. Purnami, "Pendekatan Metode Classification and Regression Tree untuk Diagnosis Tingkat Keganasan Kanker pada Pasien Kanker Tiroid," *Jurnal Sains dan Seni Pomits,* vol. 3, no. 1, pp. 24-29, 2014.

[25] N. Toulasik, "Analisis Faktor yang Berhubungan dengan Kualitas Hidup Wanita Penderita Kanker di RSUD Prof. Dr. W. Z. Johannes Kupang," Universitas Airlangga, Surabaya, 2019.

[26] FKUI, "12 Penyebab Benjolan di Leher yang Sering Terjadi, Begini Cara Mengatasinya," 2021. [Online]. Available: https://fk.ui.ac.id/infosehat/12-penyebab-benjolan-di-leher-yang-sering-terjadi-begini-caramengatasinya/. [Accessed 13 Agustus 2021].

[27] ATA, "Thyroid Cancer: Does papillary thyroid cancer run in families?," *Clinical Thyroidology for the Public,* vol. 7, no. 3, p. 10, 2014.