ANALISA PREDIKSI KEKAMBUHAN KANKER PAYUDARA DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR

Noki Ardian Madyaningrum [1], Sulastri [2]
[1,2]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Stikubank
e-mail: [1] Nociardian@gmail.com, [2]Sulastri@edu.unisbank.ac.id

ABSTRAK

Kanker payudara merupakan kanker yang menempati urutan kedua dan paling umum ditemui setelah kanker rahim. Kanker payudara dapat disembuhkan dengan berbagai pengobatan untuk menghambat pertumbuhan sel kanker yaitu dengan operasi, kemoterapi, radioterapi, dan terapi hormonal. Kekambuhan pada kanker payudara adalah penyebab utama kematian pada kanker payudara. Sampai saat ini belum diketahui hal-hal apa saja yang menyebabkan kekambuhan pada kanker payudara. Oleh sebab itu diperlukan prediksi yang tepat dengan menggunakan data mining.

Klasifikasi merupakan salah satu teknik di dalam data mining. Dalam penelitian ini akan dilakukan bagaimana menerapkan algoritma K-Nearest Neighbor. Algoritma ini melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) antara data testing dengan data training. Tujuan dari penelitian ini yaitu untuk menganalisa kekambuhan pada data kanker payudara dan mengetahui tingkat akurasi data dengan menggunakan tools RStudio.

Berdasarkan hasil percobaan sebanyak 3 kali dengan komposisi data training dan data testing yang berbeda, didapatkan hasil terbaik pada komposisi 70% data training dan 30% data testing menghasilkan nilai k terbaik pada k=5 dengan nilai akurasi 80%. Pada penitian berikutnya dapat menggunakan algoritma yang lain agar dapat diperoleh hasil yang lebih akurat.

***Kata Kunci**: Data Mining, Kanker Payudara, K-Nearest Neighbor*

1. PENDAHULUAN

Kanker payudara (KPD) merupakan keganasan pada jaringan payudara yang dapat berasal dari epitel duktus maupun lobulusnya. Kanker payudara merupakan salah satu jenis kanker terbanyak di Indonesia. Berdasarkan Pathological Based Registration di Indonesia, KPD menempati urutan pertama dengan frekuensi relatif sebesar 18,6%. Berdasarkan Data Histopatologik Badan Registrasi Kanker Perhimpunan Dokter Spesialis Patologi Indonesia (IAPI) dan Yayasan Kanker Indonesia (YKI) tahun 2010, diperkirakan angka kejadiannya di Indonesia adalah 12/100.000 wanita, sedangkan di Amerika adalah sekitar 92/100.000 wanita dengan mortalitas yang cukup tinggi yaitu 27/100.000 atau 18 % dari kematian yang dijumpai pada wanita. Penyakit ini juga dapat diderita pada laki - laki dengan frekuensi sekitar 1% [1].

Kemajuan teknologi sistem informasi telah mambantu menyelesaikan permasalahan di berbagai bidang terutama bidang kesehatan, salah satunya adalah Data Mining. Data Mining merupakan analisis dari peninjauan kumpulan data untuk menentukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data[2]. Pada data mining terdapat beberapa teknik, salah satunya adalah teknik klasifikasi. Teknik klasifikasi ini digunakan untuk membentuk model yang mendeskripsikan kelas data yang penting, atau model yang memprediksikan trend data[3]. Dalam penelitian ini akan dilakukan bagaimana menerapkan salah satu algoritma dalam data mining, yaitu *K-Nearest Neighbor* guna memprediksi kekambuhan pada data kanker payudara. Algoritma *K-Nearest Neighbor (K-NN)* adalah algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain.

Zemy S. Badu telah melakukan eksperimen menggunakan metode klasifikasi data mining *K-Nearest Neighbor* terhadap data desa yang terkait dengan klasifikasi dana desa. Tujuan dari penelitian tersebut adalah untuk memprediksi seberapa besar tingkat akurasi data dengan komposisi data training 80% dan 20% sebagai data testing. Pada penelitian tersebut diperoleh akurasi tertinggi yaitu pada K=2 sebesar 78,95%[4].

Diina Itsna Annisa telah melakukan penelitian mengenai klasifikasi kehamilan beresiko dengan menggunakan metode K-Nearest Neighbor, dengan adanya klasifikasi ini diharapkan mampu mendeteksi sejak dini dan mengurangi angka kematian ibu, janin, dan bayi akibat kehamilan beresiko. Hasil uji menghasilkan tingkat akurasi sebesar 93% dengan menggunakan nilai K=5[5].

Fitri Yunita telah melakukan kajian algoritma K-NN dan kemudian mengimplementasikannya dalam klasifikasi penderita penyakit diabetes mellitus pasien RSUD Puri Husada Tembilahan. Dengan menggunakan metode *K-Nearest Neighbor*, dalam penelitian ini menggunakan tools data mining rapid miner[6].

2. METODE PENELITIAN

Obyek penelitian yang digunakan adalah data Kanker Payudara yang diperoleh dari Universitas Medical Center, Institut Onkologi, Ljubljana, Yugoslavia (1998) sebanyak 286 record dan terdiri dari 10 atribut yang meliputi Class, Age, Menoupause, Tumor-Size, Inv-Nodes, Node Caps, Deg- Malig, Breast, Breast-Quad dan Irradiant. Data tersebut diperoleh dari UCI Machine Learning dengan link https://archive.ics.uci.edu/ml/datasets/Breast+Cancer.

Algoritma *K-Nearest Neighbor (KNN)* adalah algoritma yang digunakan untuk melakukan klasifikasi terhadap suatu objek, berdasarkan k buah data latih yang jaraknya paling dekat dengan objek tersebut. Syarat nilai k adalah tidak boleh lebih besar dari jumlah data latih, dan nilai k harus ganjil. Secara umum nilai k optimal yang sering digunakan berkisar diantara 3-10 atau $\sqrt{n}$ dimana n merupakan jumlah data latih. Itu akan menghasilkan hasil yang lebih baik dibandingkan dengan 1NN . Euclidian Distance sering digunakan untuk menghitung jarak antara dua titik data.

Berikut urutan proses kerja *K-Nearest Neighbor* :

1. Menentukan parameter *k* (jumlah tetangga paling dekat).
2. Menghitung kuadrat jarak *Euclidean (Euclidean Distance)* masing-masing obyek terhadap data testing yang diberikan.
   Rumus Euclidean Distance dinotasikan pada rumus 1 sebagai berikut :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \qquad (1)$$

Dengan keterangan :

D = Jarak

i = Variabel Data

n = Banyak jumlah data

xi = Sampel data

yi = Data Uji/Testing

3. Mengurutkan obyek-obyek tersebut ke dalam kelompok yang mempunyai jarak *Euclidean* terkecil.
4. Mengumpulkan kategori Y (klasifikasi *nearest neighbor*)

Dengan menggunakan kategori *Nearest Neighbor* yang paling mayoritas maka dapat diprediksi nilai queri instance yang telah dihitung.

Proses data mining ini melalui tahap analisanya berdasarkan metode KDD *(Knowledge Discovery in Databases*) yang dijelaskan sebagai berikut :

2.1 Persiapan Data

Gambar 1 menunjukkan persiapan data yang merupakan tahapan awal sbelum digunakan ke dalam proses data mining :

Class	Age	Menopause	Tumor-Size	Inv-Nodes	Node-Caps	Deg-Malig	Breast	Breast-Quad	Irradiat
no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no
no-recurrence-events	60-69	ge40	15-19	0-2	no	2	left	left_low	no
no-recurrence-events	50-59	premeno	25-29	0-2	no	2	left	left_low	no
no-recurrence-events	60-69	ge40	20-24	0-2	no	1	left	left_low	no
no-recurrence-events	40-49	premeno	50-54	0-2	no	2	left	left_low	no
no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	left_up	no
no-recurrence-events	40-49	premeno	0-4	0-2	no	3	left	central	no
no-recurrence-events	50-59	ge40	25-29	0-2	no	2	left	left_low	no
no-recurrence-events	60-69	lt40	10-14	0-2	no	1	left	right_up	no
no-recurrence-events	50-59	ge40	25-29	0-2	no	3	left	right_up	no
no-recurrence-events	40-49	premeno	30-34	0-2	no	3	left	left_up	no
no-recurrence-events	60-69	lt40	30-34	0-2	no	1	left	left_low	no
no-recurrence-events	40-49	premeno	15-19	0-2	no	2	left	left_low	no
no-recurrence-events	50-59	premeno	30-34	0-2	no	3	left	left_low	no
no-recurrence-events	60-69	ge40	30-34	0-2	no	3	left	left_low	no

Gambar 1 Persiapan Data

2.2. **Pre-Processing/Cleaning**

Pada proses pre-processing/cleaning dilakukan pemeriksaan data yang inkonsisten, dan memperbaiki kesalahan pada data agar mudah dalam proses data mining. Gambar 2 menunjukkan data yang sudah di proses *Cleaning* :

Class	Age	Menopau	TumorSiz	InvNodes	NodeCaps	DegMalig	Breast	BreastQua	Irradiat
0	5	2	2	1	0	1	2	3	0
0	4	1	2	1	0	1	1	1	0
0	4	1	2	1	0	3	2	2	1
0	4	1	2	2	1	2	2	1	0
0	6	2	1	1	0	2	2	2	1
0	6	2	2	2	0	1	2	2	1
0	6	2	2	2	0	1	2	1	1
0	4	1	2	2	0	2	2	2	0
0	4	1	2	2	0	2	2	1	0

Gambar 2. Data setelah dicleaning

2.3 Transformasi Data

Transformasi digunakan untuk mengubah angka menjadi huruf. Data yang diubah adalah :

1. Class yaitu variable yang berisi mengenai status kekambuhan, nilainya adalah no-recurrence-events = 0, recurrence-events = 1
2. Age yaitu Jika pasien berumur 20-29 nilai 2, jika pasien berumur 30-39 nilai 3, jika pasien berumur 40-49 nilai 4, jika pasien berumur 50-59 nilai 5, jika pasien berumur 60-69 nilai 6 dan jika pasien berumur 70-79 nilai 7.
3. Menopause dengan nilai premeno = 1, Ge40 = 2, Lt40 = 3.
4. Tumor-Size, jika tidak ditemukan = 0, ukuran tumor <=20 mm = 1, ukuran tumor<=50 mm = 2, ukuran tumor >50 mm = 3.
5. Inv-Nodes , 0 kgb = 0, 1-3kgb = 1, 4-9kgb = 2, >10kgb =3
6. Node Caps, Penyebaran ke Kelenjar Getah Bening, No = 0, Yes = 1
7. Deg-Malig, tingkat keganasa : 1,2 dan 3
8. Breast, Letak Payudara, Left = 1, Right = 2
9. Breast-Quad, Letak Posisi Tumor Payudara, Low_left = 1, Left_up = 2, Central = 3,Right_low=4,Right_up = 5
10. Irradiant, Status Radiasi, 0 = No, 1 = Yes

Class	Age	Menopause	TumorSize	InvNodes	NodeCaps	DegMalig	Breast	BreastQuad	Irradiat
0	3	1	2	1	0	3	1	1	0
0	4	1	2	1	0	2	2	5	0
0	4	1	2	1	0	2	1	1	0
0	6	2	1	1	0	2	2	2	0
0	4	1	1	1	0	2	2	4	0
0	6	2	1	1	0	2	1	1	0
0	5	1	2	1	0	2	1	1	0
0	6	2	2	1	0	1	1	1	0
0	4	1	3	1	0	2	1	1	0
0	4	1	2	1	0	2	2	2	0
0	4	1	1	1	0	3	1	3	0
0	5	2	2	1	0	2	1	1	0
0	6	3	1	1	0	1	1	5	0
0	5	2	2	1	0	3	1	5	0
0	4	1	2	1	0	3	1	2	0
0	6	3	2	1	0	1	1	1	0

Gambar 3. Data yang sudah ditransformasi

2.2 Pembagian Data

Proses ini digunakan untuk membagi antara *data training* dengan *data testing*. Pembagian data dalam penelitian ini dilakukan sebanyak tiga kali. Tabel 2 menunjukkan pembagian antara *data training* dan *data testing*.

Tabel 2. Pembagian Data Set

Percobaan	*Data Training*	*Data Testing*
Percobaan 1	70% = 201 data	30% = 85 data
Percobaan 2	75% = 215 data	25% = 71 data
Percobaan 3	80% = 229 data	20% = 57 data

2. HASIL DAN PEMBAHASAN

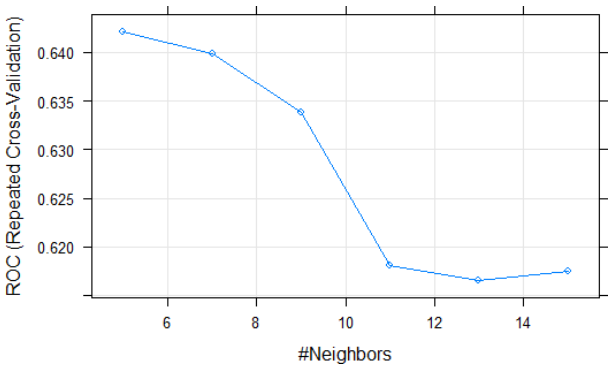
Proses data mining disini diawali membuat model dengan *data training* kemudian mengimplementasikannya ke *data testing* dengan menggunakan algoritma *K-Nearest Neighbor*, sehingga didapatkan hasil akhir berupa nilai akurasi dan hasil klasifikasi. Variabel yang digunakan sebagai varibel penentu klasifikasi yaitu variabel "Class" dimana mempunyai 2 nilai **"Kambuh"** dan **"Tidak"**. Berikut merupakan hasil terbaik pada percobaan 1 dengan komposisi *data training* 70% dan *data testing* 30%.

```
k-Nearest Neighbors

201 samples
  9 predictor
  2 classes: 'No', 'Yes'

Pre-processing: centered (9), scaled (9)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 181, 181, 181, 181, 181, 181, ...
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
   5  0.6420899  0.8728571  0.2222222
   7  0.6398810  0.8942857  0.1944444
   9  0.6337963  0.9107937  0.2111111
  11  0.6180291  0.9106349  0.1722222
  13  0.6165608  0.9390476  0.1722222
  15  0.6174339  0.9436508  0.1833333

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

Gambar 4. Hasil pemilihan nilai K terbaik

Gambar 4 menunjukkan hasil berbagai nilai K ganjil, dimulai dari K= 5 sampai dengan 15.

Adapun visualisasi dengan menggunakan matrix "pROC" ditunjukkan Gambar 5 sebagai berikut :



Gambar 5. Visualisasi dengan *pROC*

Dari model diatas dapat dievaluasi dengan menggunakan *Confusion Matrix* seperti gambar dibawah ini:

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  57  14
       Yes  3  11

               Accuracy : 0.8
                 95% CI : (0.6992, 0.879)
    No Information Rate : 0.7059
    P-Value [Acc > NIR] : 0.03366

                  Kappa : 0.4474
 Mcnemar's Test P-Value : 0.01529

            Sensitivity : 0.9500
            Specificity : 0.4400
         Pos Pred Value : 0.8028
         Neg Pred Value : 0.7857
             Prevalence : 0.7059
         Detection Rate : 0.6706
   Detection Prevalence : 0.8353
      Balanced Accuracy : 0.6950

       'Positive' Class : No
```

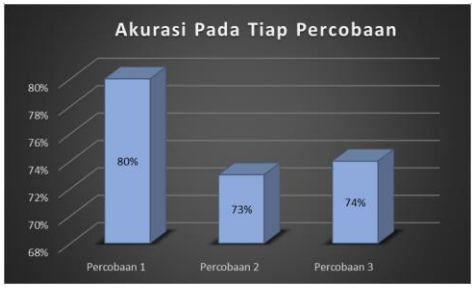
Gambar 6. Hasil Evaluasi

Gambar 6 merupakan *Confusion Matrix* yang menunjukkan hasil dari akurasi penghitungan. Adapun variabel yang paling berpengaruh dapat dilihat pada gambar 7 sebagai berikut :

```
ROC curve variable importance

           Importance
DegMalig      100.000
InvNodes       68.667
NodeCaps       54.793
Irradiat       45.557
TumorSize      36.789
Menopause      11.925
Breast          6.976
Age             2.572
BreastQuad      0.000
```

Gambar 7. Variabel yang paling berpengaruh

Dari gambar 7 diatas dapat diketahui bahwa variabel "DegMalig" merupakan variabel yang sangat berpengaruh dalam memprediksi kambuhnya kanker payudara. Setelah melakukan ketiga percobaan, didapatkan hasil akhir berupa akurasi yang divisualisasikan pada grafik 3D berikut ini :



Gambar 8. Hasil akurasi pada tiap percobaan

Dari gambar 8 dapat diketahui bahwa hasil terbaik diperoleh pada percobaan pertama dengan komposisi *data training* 70% dan *data training* 30% dengan akurasi 80%.

3. KESIMPULAN & SARAN

Berdasarkan hasil penelitian dan pembahasan , maka didapatkan kesimpulan :

1. Berdasarkan percobaan pertama dengan komposisi *data training* 70% dan *data testing* 30% didapatkan hasil nilai K terbaik pada K= 5 dan nilai akurasi 80%. Sedangkan pada percobaan kedua dengan komposisi *data training* 75% dan *data testing* 25% didapatkan hasil nilai K terbaik pada K= 13 dan nilai akurasi 73%. Adapun pada percobaan ketiga dengan komposisi *data training* 80% dan *data testing* 20% didapatkan hasil nilai K terbaik pada K= 15 dan nilai akurasi 74%.
2. Dari ketiga percobaan yang telah dilakukan didapatkan variabel paling berpengaruh yaitu "Deg-Malig". Sedangkan atribut yang dianggap tidak berpengaruh adalah "BreastQuad".

Penelitian yang dilakukan masih terdapat kekurangan. Oleh sebab itu, terdapat beberapa saran yang dapat diaplikasikan pada penelitian dimasa yang akan datang :

1. Melakukan komparasi terhadap algoritma atau metode data mining lainnya dalam menganalisa data kanker payudara, untuk mengetahui algoritma mana yang lebih kuat, efisien dan akurat. Sehingga dapat ditentukan algoritma yang tepat yang dapat digunakan dalam menganalisa data kanker payudara.
2. Dapat menggunakan tools lain yakni Rapid Miner, Matlab, dan lain sebagainya untuk mendapatkan analisis yang lebih akurat lagi.
3. Perlu adanya penelitian untuk mengetahui tingkat *error* dengan metode *K-Nearest Neighbor* untuk klasifikasi data kanker payudara.

DAFTAR PUSTAKA

[1] Komite Penanggulangan Kanker Nasional Kemenkes RI.,2015, Panduan Penatalaksanaan Kanker Payudara. *Kementerian Kesehatan Republik Indonesia. Komite Penanggulangan Kanker Nasional.*, 1–56.
[2] Larose, 2005, *Algoritma Data Mining*, edisi 1, Andi Ofset, Yogyakarta.
[3] Kusrini, 2009, *Algoritma Data Mining*, edisi 1, Andi Ofset, Yogyakarta.
[4] Badu, Z. S. (2016). Penerapan Algoritma K-Nearest Neighbor. *Informatika*, (November).
[5] Kasus, S., Kesehatan, D., Malang, K., Annisa, D. I., Ariyanto, R., Tri, A., & Hayati, R,. 2016, Klasifikasi Kehamilan Beresiko Dengan Menggunakan Metode K-Nearest Neighbor. *Jurnal Informatika Polinema*, v *3*(November), 34–39.
[6] Indrayanti, Sugianti, D., & Al Karomi, M. A., 2017, Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus. *Prosiding SNATIF Ke-4 2017*, 823–829. https://doi.org/10.1007/s10115-007-0114-2
[7] Prasetyo, Eko., 2012, Data Mining – Konsep dan Aplikasi Menggunakan MATLAB.