SIGMA
JURNAL TEKNOLOGI PELITA BANGSA
Vol. 13 No. 3 September 2022 p-ISSN: 2407-3903 e-ISSN: 2407-3903
Diterima, 03 Oktober 2022 | Direvisi, 10 Oktober 2022 | Dipublikasikan, 14 Oktober 2022

ANALISA DATA MINING UNTUK PREDIKSI PENYAKIT KANKER PARUDENGAN ALGORITMA REGRESI LINEAR

Suherman [1], Faris Muammar[2], Irfan Afriantoro[3]

Program Studi Teknik Informatika Fakultas Teknik Universitas Pelita Bangsa
[1]suherman@pelitabangsa.ac.id, [2]farismuammar20@gmail.com, [3]irfan@pelitabangsa.ac.id

Abstraksi

Kanker paru-paru sering kali tidak menimbulkan gejala pada tahap awal. Gejala baru muncul ketika kanker sudah cukup besar atau telah menyebar ke jaringan dan organ sekitar. Sehingga penderita kanker paru baru akan merasakan sakit setelah kanker menyebar ke lapisan pleura, lapisan tipis yang menutupi paru-paru.. Penelitian ini bertujuan untuk menganalisa penyakit kanker paru- paru dalam pencegahan dini. Penelitian ini menggunakan teknik prediksi dan tahapan-tahapan pada data mining untuk memprediksi data pasienyang menderita penyakit kanker paru-paru dengan metode algoritma regresi linear menggunakan *tools* rapidminer, pengolahan data yang di jadikan dataset dalam penelitian ini, dataset dibagi menjadi dua yaitu90% data training dan 10% data testing. hasil pengujian yang telah dilakukan bahwa variabel atau atribut yang digunakan dalam penelitian ini (usia, merokok, dan hasil_test) berpengaruh signifikan terhadap penelitian ini terbukti dengan menggunakanalgoritma regresi linear mampu memberikan hasil yang baik dengan nilai *Root Mean Squared Error*: 0.379 +/- 0.000 dan *Squared Error*: 0.144 +/- 0.229. Kesimpulan dari penelitian yang dilakukan dengan menerapkan algoritma regresi linear dapat dilakukansuatu prediksi berdasarkan hubungan fungsional padavariable atau atribut didalam data tersebut.

Kata kunci: Kanker paru-paru, Regresi linear,*Rapidminer*

Abstract

Lung cancer often causes no symptoms in its early stages. New symptoms appear when the cancer is large enough or has spread to surrounding tissues and organs. So that patients with lung cancer will only feel pain after the cancer spreads to the pleural layer, the thin layer that covers the lungs. This study aims to analyze lung cancer in early prevention. This study uses prediction techniques and stages in data mining to predict data on patients suffering from lung cancer with a linear regression algorithm method using rapidminer tools. training and 10% data testing. The results of the tests that have been carried out show that the variables or attributes used in this study (age, smoking, and test results) have a significant effect on this study, as evidenced by using a linear regression algorithm to provide good results with a Root Mean Squared Error value: 0.379 +/- 0.000 and Squared Error: 0.144 +/- 0.229. The conclusion of the research conducted by applying the linear regression algorithm can be made a prediction based on the functional relationship onthe variables or attributes in the data.

Keywords: Lung cancer, linear regression algorithm, Rapidminer

1. Pendahuluan

Kanker adalah salah satu dari penyebab kematian terbesar di dunia. Kanker merupakan penyakit kompleks yang melibatkan pertumbuhan sel abnormal atau sel tidak biasa yang dikenal sebagai tumor ganas. Dari semua kanker, kanker paru-paru menjadi penyakit paling umum yang menyebabkan kematian [1].

Kanker paru adalah semua penyakit keganasan di paru, mencakup keganasan yang berasal dari paru sendiri (primer) Dalam pengertian klinik yang dimaksud dengan kanker paru primer adalah tumor ganas yang berasal dari epitel bronkus (karsinoma bronkus = bronchogenic carcinoma). Kanker paru merupakan penyebab utama keganasan di dunia, mencapai hingga 13 persen dari semua diagnosis kanker. Selain itu, kanker paru juga menyebabkan 1/3 dari seluruh kematian akibat kanker pada laki-laki. Di Amerika

Serikat, diperkirakan terdapat sekitar 213.380 kasus baru pada tahun 2007 dan 160.390 kematian akibat kanker paru. Berdasarkan data WHO, kanker paru merupakan jenis kanker terbanyak pada laki-laki di Indonesia, dan terbanyak kelima untuk semua jenis kanker pada perempuan Kanker paru juga merupakan penyebab kematian akibat kanker terbanyak pada laki-laki dan kedua pada perempuan [2].

Data adalah kumpulan fakta yang terekam atau sebuah entitas yang tidak memiliki arti dan selama initerabaikan sedangkan mining adalah proses penambangan. Sehingga data mining dapat diartikan proses penambangan data yang menghasilkan output (keluaran) yang berupa pengetahuan. Klasifikasi proses pengidentifikasian objek ke dalam sebuah kelas atau kelompok berdasarkan atribut data yang akan digunakan yang bertujuan untuk menempatkan objek dan variabel data. Regresi Linear digunakan untuk mengetahui bagaimana variabel dependen/kriteria dapat diprediksikan melalui variabel independen atau variabel prediktor secara individual. Dampak dari penggunaan analisis regresi dapat digunakan untuk memutuskan apakah naik dan menurunnya variabel dependen dapat dilakukan melalui menaikan dan menurunkan keadaan variabel independen, atau meningkatkan keadaan variabel dependen dapat dilakukan dengan meningkatkan variabel independen dan sebaliknya.

Penelitian ini akan dijadikan acuan yang mengenai analisa prediksi kanker paru-paru. Hasil dari pengolahan data kanker paru-paru tersebut menjadi sebuah informasi dan pengetahuan yang diharapkan, sehingga dapat digali suatu potensi atau pengetahuan yang lebih baik atau akurat dalam pembacaan datanya, tepat dan cepat dari data tersebut sehingga dapat menganalisa prediksi kanker paru-paru dan menemukan peluang-peluang yang baru serta menemukan rencana strategis dan untuk analisa dan prediksi kanker paru-paru, selain itu bisa digunakan sebagai sarana untuk mengambil keputusan.

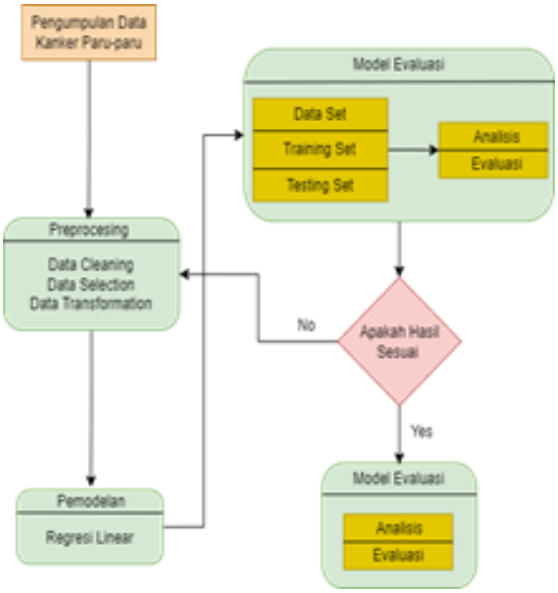
Berdasarkan uraian masalah diatas, maka dalam penelitian ini digunakan algoritma regresi linear sebagai proses identifikasi data dengan judul "ANALISA DATA MINING UNTUK PREDIKSI PENYAKIT KANKER PARU DENGAN ALGORITMA REGRESI LINEAR".

2. Landasan Pemikiran

Pada penelitian ini, tahapan yang akan digunakan dalam melakukan prediksi terhadap data kanker paru-paru dan penentuan atribut pasien atau pemeriksaan untuk mempermudah penelitian sehingga penelitian dapat berjalan dengan baik dan sistematis, serta memenuhi tujuan yang diinginkan. Berikut ini adalah langkah-langkah dalam tahapan yang dilakukandalam penelitian ini.

2.1 Tahapan Penelitian

Dalam melakukan prediksi data kanker paru-paruyang akan di uji sesuai pemodelan data yang akan digunakan agar mempermudahkan penelitian dan berjalan sesuai dinginkan maka dibuat alur atau tahapan dalam penelitian ini sebagai berikut:



Gambar 1. Tahapan Penelitian

2.2 Data Transfomation

Tahap Data Transformation merupakan proses mengubah format data awal menjadi sebuah format data untuk proses dengan algoritma pada program maupun tools yang digunakan. Berikut adalah hasil pengolahan data awal setelah melewati tahapan diatas untuk dijadikan dataset pada tahap selanjutnya, ditunjukkan pada Tabel 1

Tabel 1.Data *Transfomation*

Usia	Merokok
69	1
74	2
59	1
63	2
63	1
75	1
52	2
51	2
68	2
53	2
61	2
72	1
60	2
58	2

Y: peubah tak-bebas
X: peubah bebas
a: konstanta b: kemiringan

Y = a + bX

2.3 Pemodelan

Pemodelan pada penelitian ini dilakukan dengan data mining teknik estimasi menggunakan algoritmaregresi linear. Regresi Linear digunakan untuk mengetahui bagaimana variabel dependen/kriteria dapat diprediksikan melalui variabel independen atauvariabel prediktor, secara individual. Dampak dari penggunaan analisis regresi dapat digunakan untuk memutuskan apakah naik dan menurunnya variabel dependen dapat dilakukan melalui menaikan dan menurunkan keadaan variabel independen, atau meningkatkan keadaan variabel dependen dapat dilakukan dengan meningkatkan variabel *independen*/dan sebaliknya. Metode Kuadrat terkecil(*least square method*) metode yang paling populer untuk menetapkan persamaan regresi linier sederhana. Bentuk Umum Regresi Linear Sederhana:

Penetapan Persamaan Regresi Linier Sederhana nilai

a dan b

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \bar{y} - b\bar{x} \quad \text{sehingga} \quad a = \frac{\sum_{i=1}^{n} y_i}{n} - b\frac{\sum_{i=1}^{n} x_i}{n}$$

Menghitung persamaan regresi liniernya adalah
Y = a + bX
Y: peubah tak-bebas
X: peubah bebas
a: konstanta
b: kemiringan

3.Metode Penelitian

3.1.Pengujian dan Validasi

Pengujian metode dilakukan dengan maksud mengetahui hasil perhitungan yang dianalisa dan mengukur metode serta prosedur pemecahan yang digunakan apakah berfungsi dengan baik atau tidak. Proses pengujian menggunakan tools rapidminer dan melihat data apakah sesuai dengan yang akan terjadi yang diperoleh melalui tool tersebut.

3.2.Evaluasi dan Hasil

Evaluasi dilakukan dengan maksud mengetahui hasil perhitungan yang dianalisa dan mengukur metode serta algoritma regrensi linier yang digunakan apakah berfungsi dengan baik atau tidak. Proses pengujian menggunakan tools repidminer dan melihat data apakah sesuai dengan hasil yang diperoleh melalui tools tersebut, mengetahui bagaimana variable dependen/kriteria dapat diprediksikan melalui variabel independen atau variabel prediktor, secara individual. Dampak dari penggunaan analisis regresi dapat digunakan untuk memutuskan apakah naik dan menurunnya variabel dependen dapat dilakukan melalui menaikan dan menurunkan keadaan variabel independen, atau meningkatkan keadaan variabel dependen dapat dilakukan dengan meningkatkan variabel independen/dan sebaliknya Metode Kuadrat terkecil (least square method).

3.3.Data Uji

Penelitian ini menggunakan algoritma Regresi Linear, untuk mengidentifikasikan penyakit kanker paru-paru dan akan mendapatkan hasil Root Mean Square Error (RMSE) serta prediksi yang dapat digunakan dalam pengambilan keputusan ketika pasien terkena penyakit atau tidak. Sumber data sebagai objek pada penelitian ini adalah data historis yang diambil dari situs Kaggle.com. Data yang digunakan dalam penelitian ini terdiri dari atribut atau variabel seperti usia, merokok, dan hasil test.

3.4.Spit Validation

Teknik validasi yang membagi data menjadi dua bagian secara acak, sebagian sebagai data training dan sebagian lainnya sebagai data testing. Dengan menggunakan Split Validation akan dilakukan percobaan training berdasarkan split ratio yang telah ditentukan sebelumnya, untuk kemudian sisa dari split ratio data training akan dianggap sebagai data testing.

Tabel 2. Dataset Kanker Paru

Usia	Merokok	Hasil	Usia		Merokok
69	1	1	48		1
74	2	1	75		2
59	1	0	57		2
63	2	0	68		2
63	1	0	61		1
75	1	1	44		2
52	2	1	64		1
51	2	1	21		2
68	2	0	60		2
53	2	1	72		2
61	2	1	65		1
72	1	1	61		2
60	2	0	69		1
58	2	1	…..		…….

]=

4. Pembahasaan

Tabel 3 Prediksi Penyakit Kanker Paru

No	Usia (X_1)	Merokok (X_2)	Hasil_Test (Y)
1	58	2	?
2	21	2	?
3	69	1	?

Menghitung Persaman Regresi LinearY = $a + b_1X_1 + b_2X_2$

y = 0,6685241936 + (−0,0001746768 x 58) + (0,109751122 *x* 2)

= **0,8778951832**

y = 0,6685241936 + (−0,0001746768 x 21) + (0,109751122 *x* 2)

= 0,8843582248

y = 0,6685241936 + (−0,0001746768 x 69) + (0,109751122 x 1)

= **0,7662226164**

4.1. Proses Pengujian Rapid Mineer

Melakukan select attributes yaitu untuk mengetahui hasil prediksi dari rapidminer, hasil perhitungan manual dan hasil uji di rapidminer. Dapat dilihat pada gambar 2.

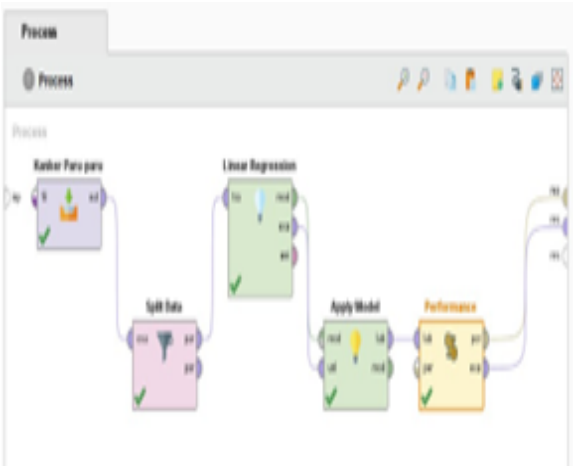
Gambar 2. Proses *Rapidminer*

Pada proses ini untuk memasukan data training dan data testing yang akan diuji untuk menghasilkan prediksi pada atribut class yang ditampilkan pada gambar 3.

Row No.	Hasil	prediction(H...	Usia	Merokok
1	1	0.806	69	1
2	1	0.986	74	2
3	0	0.711	59	1
4	0	0.882	63	2
5	0	0.749	63	1
6	1	0.863	75	1
7	1	0.778	52	2
8	1	0.769	51	2
9	0	0.930	68	2
10	1	0.788	53	2
11	1	0.834	72	1
12	0	0.854	60	2
13	1	0.835	58	2
14	0	0.939	69	2

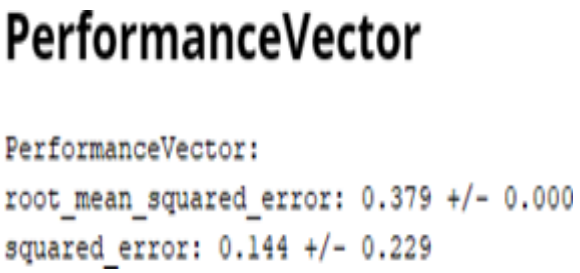
Ganbar 3 . Hasil Prediksi

Ketika prediksi sudah dicari langkah selanjutnya yaitu mengukur seberapa akurat hasil prediksi yang telah kita buat pada



Gambar 4. Proses Pencarian *Root Mean SquaredError* dan *Squared Error*

Untuk mempermudah dalam pembacaan data kanker paru-paru, maka perlu di masukan *tools performance* untuk mencari *Root Mean Squared Error* dan*Squared Error*. Berikut ini hasilnya.



Gambar 5. Hasil Pengujian *Root Mean Square Error* dan *Square Error*

Langkah kedua, dilakukan implementasi algoritma regresi linear dengan menggunakan *tools rapidminer*. Berikut adalah tahapan dalam penerapan algoritma regresi linear:

1. Menentukan prediksi data test yang dilakukan oleh rapidminer dan menghasilkan nilai confidence yang telah diprediksi.
2. Menentukan *performance* dengan *output* untuk mencari *Root Mean Squared Error* dan *Squared Error*.

Pada permodelan *split validation* didalamnyaterdapat dua bagian, yaitu bagian *training* (digunakan untuk algoritma klasifikasi) dan bagian *testing* (menggunakan fitur *Apply Model* untuk mengaplikasikan model pada data *testing* dan fitur *Performance* untuk menampilkan *Root MeanSquared Error* dan *Squared Error*).

4.2. Analisa Hasil

Berdasarkan hasil pengujian yang telah dilakukan bahwa variabel atau atribut yang digunakan dalam

penelitian ini (usia, merokok, dan hasil_test) berpengaruh signifikan terhadap penelitian ini terbukti dengan menggunakan algoritma regresilinear mampu memberikan hasil yang baik dengan nilai *Root Mean Squared Error*: 0.379 +/- 0.000 dan *Squared Error*: 0.144 +/- 0.229. Hal ini dikarenakan adanya korelasi atau hubungan fungsional (sebab – akibat) antara variabel yang satu (dependen atau kriteria) dengan variabel yang lain (independen atau predictor). Proses pengujian ini dilakukan untuk mengidentifikasi penyakit kanker paru-paru dengan algoritma regresi linear.

5. Penutup

Berdasarkan hasil pengujian yang telah dilakukan dalam penelitian ini, maka dapat diambil suatu kesimpulan yaitu:

Pada penelitian ini dengan memanfaatkan beberapa data pasien penderita penyakit kanker paru-paru yang telah tersimpan dalam basis data menggunakan beberapa atribut diantaranya: usia, merokok dan hasil test. Sehingga dengan menerapkan algoritma regresi linear dapat dilakukan suatu prediksi berdasarkan hubungan fungsional pada variable atau atribut didalam data tersebut.

Mengolah data kanker paru-paru menggunakan algoritma regresi linear dimulai dari tahap seleksi data (atribut yang digunakan dan penentuan data training serta data testing), tahap pengujian algoritma (regresi linear), dan tahap uji akurasi (menggunakan split validation).

Proses pengujian data pada penelitian ini menggunakan algoritma regresi linear mampu memberikan hasil yang baik dengan nilai Root Mean Squared Error: 0.379 +/- 0.000 dan Squared Error: 0.144 +/- 0.229.

Daftaar Pustaka

[1] I. W. Septiani, A. C. Fauzan, and M. M. Huda, "Implementasi Algoritma K-Medoids Dengan Evaluasi Davies-Bouldin- Index Untuk Klasterisasi Harapan Hidup Pasca Operasi Pada Pasien Penderita Kanker Paru-Paru," *J. Sist. Komput. dan Inform.*, vol. 3, no. 4, pp. 556–566, 2022, doi: 10.30865/json.v3i4.4055.

[2] A. F. Iedam and D. Riana, "Prediksi Harapan Hidup Pasien Kanker Paru-Paru Pasca Operasi Bedah Thoraks Menggunakan Boosted Neural Network Dan Smote," *J. Infomedia Tek. Inform. Multimed. Jar.*, vol. 6,no. 1, pp. 9–15, 2021.

[3] A. Fitri Boy, "Implementasi Data Mining Dalam Memprediksi Harga Crude Palm Oil (CPO) Pasar Domestik Menggunakan Algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi SumateraUtara)," *J. Sci. Soc. Res.*, vol. 4307, no. 2, pp.78–85, 2020, [Online]. Available: http://jurnal.goretanpena.com/index.php/JSS R

[4] R. T. Prasetio and S. Susanti, "Prediksi Harapan Hidup Pasien Kanker Paru Pasca Operasi Bedah Toraks Menggunakan Boosted k-Nearest Neighbor," *J. Responsif*, vol. 1, no.1, pp. 64–69, 2019, [Online]. Available: http://ejurnal.univbsi.id/index.php/jti

[5] Hafizah, Tugiono, and W. R. Maya,"Penerapan Data Mining Dalam Memprediksi Jumlah Penumpang Pada CV . Surya MandiriSukses Dengan Menggunakan Metode Regresi Linier," *J. Teknol. Inf. dan Sist. Komput. TGD*, vol. 2, no. 1, pp. 54–61, 2019.

[6] M. Yunianto *et al.*, "Klasifikasi Kanker Paru Paru menggunakan Naïve Bayes dengan Variasi Filter dan Ekstraksi Ciri GLCM," *Indones. J. Appl. Phys.*, vol. 11, no. 2, p. 256, 2021, doi: 10.13057/ijap.v11i2.53213.

[7] F. Fajriah, "Adenocarcinoma Dengan Mutasi Egfr," *J. Kedokt. Syiah Kuala*, vol. 18, no. 1, pp. 66–68, 2018, doi:10.24815/jks.v18i1.11216.

[8] H. Kenang, C. Alivian, W. Suharso, and A. Qurrota, "Pengklasifikasian Kanker Payudara Dan Kanker Paru-Paru Dengan Metode Gaussian Naïve Bayes , Multinomial Naïve Bayes , Dan Bernoulli Naïve Bayes Classification Of Breast Cancer And Lung Cancer Using The Gaussian Naïve Bayes Multinomial Nave Bayes And Berno," *J. Smart Teknol.*, vol. 3, no. 4, pp. 350–355, 2022.

[9] A. A. Aljumah, M. G. Ahamad, and M. K. Siddiqui, "Application of data mining: Diabetes health care in young and old patients," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 25, no. 2, pp. 127–136, 2013, doi: 10.1016/j.jksuci.2012.10.003.

[10] F. Aris and Benyamin, "Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi," *Router Res.*, vol. 1, no. 1, pp. 1–6, 2019, [Online]. Available: https://www.ejournal.stipwunaraha.ac.id/index.php/router/article/view/313

[11] Kusrini and E. T. Luthfi, *Algoritma Data Mining*. Yogyakarta: ANDI, 2009.

[12] M. S. Mustafa and I. W. Simpen,"Implementasi Algoritma K-NearestNeighbor (KNN) Untuk Memprediksi

Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba," *Pros.Semin. Ilm. Sist. Inf. Dan Teknol. Inf.*, vol.VIII, no. 1, pp. 1–10, 2019, [Online]. Available: https://ejurnal.dipanegara.ac.id/index.php/sisiti/article/view/1 -10

[13] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika, 2017.