



Prediksi Penyakit Kanker Paru-Paru Dengan Algoritma Regresi Linier

M. Abdul Rahman Wahid, Agung Nugroho, Abdul Halim Anshor*

Fakultas Teknik, Teknik Informatika, Pelita Bangsa, Bekasi, Indonesia

Email : 1abdulcasilas1@gmail.com, 2agung@pelitabangsa.ac.id, 3*abdulhalimanshor@pelitabangsa.ac.id

Email Penulis Korespondensi: abdulcasilas1@gmail.com

Abstrak– Kanker paru-paru adalah salah satu jenis kanker yang paling mematikan di seluruh dunia. Oleh karena itu, upaya untuk memprediksi kemungkinan terkena kanker paru-paru sangat penting dalam pencegahan dan penanganannya. Salah satu cara untuk memprediksi kemungkinan terkena kanker paru-paru adalah dengan menggunakan algoritma regresi linier. Tudi ini bertujuan untuk mengembangkan model prediksi yang dapat mengidentifikasi kemungkinan seseorang terkena kanker paru-paru berdasarkan faktor-faktor tertentu, seperti usia, perokok pasif dan level atau tingkat parah. Data yang digunakan dalam penelitian ini dikumpulkan dari 100 pasien yang didiagnosis dengan kanker paru-paru beserta tingkat ke-parahannya. Hasil analisis menunjukkan bahwa algoritma regresi linier dapat digunakan untuk memprediksi kemungkinan terkena kanker paru-paru dengan akurasi sekitar 90% dan mampu memberikan hasil yang baik dengan nilai Root Mean Squared Error: 0.686 +/- 0.000 dan Squared Error: 0.471 +/- 0.546.

Kata Kunci: Data Mining, Regresi Linier, Penyakit Kanker Paru-Paru

Abstract– Lung cancer is one of the deadliest types of cancer worldwide. Therefore, efforts to predict the likelihood of developing lung cancer are very important in its prevention and treatment. One way to predict the likelihood of getting lung cancer is to use a linear regression algorithm. This study aims to develop a predictive model that can identify a person's likelihood of developing lung cancer based on certain factors, such as age, passive smoker and level or severity. The data used in this study were collected from 100 patients diagnosed with lung cancer and their severity. The results of the analysis show that the linear regression algorithm can be used to predict the probability of getting lung cancer with an accuracy of about 90% and is able to give good results with a Root Mean Squared Error: 0.686 +/- 0.000 and Squared Error: 0.471 +/- 0.546.

Keywords: Data Mining, Linear Regression, Lung Cancer

1. PENDAHULUAN

Kanker paru-paru adalah salah satu jenis kanker yang paling mematikan di seluruh dunia. Setiap tahun, jutaan orang didiagnosis dengan kanker paru-paru dan ribuan orang meninggal akibat penyakit ini. Oleh karena itu, deteksi dini dan prediksi penyakit ini sangat penting untuk meningkatkan peluang kesembuhan dan menyelamatkan nyawa. Dalam era digital saat ini, teknologi komputer dan analisis data memberikan kemampuan untuk menganalisis data kesehatan dengan cara yang lebih efektif dan efisien. Salah satu teknik yang digunakan untuk prediksi penyakit adalah algoritma regresi linier. Algoritma ini memungkinkan untuk memodelkan hubungan antara variabel independen dan variabel dependen untuk memprediksi nilai yang diinginkan. Dalam konteks prediksi kanker paru-paru, algoritma regresi linier dapat digunakan untuk memprediksi risiko seseorang terkena kanker paru-paru berdasarkan faktor-faktor risiko tertentu seperti usia, jenis kelamin, riwayat merokok, paparan asap rokok, dan faktor lingkungan lainnya. [1]

Data mining (penggalian data) adalah proses menemukan pola atau informasi yang berguna dari sejumlah besar data yang tersimpan dalam database, data warehouse, atau sumber data lainnya. Tujuan dari data mining adalah untuk mengekstraksi informasi yang berharga dari data, termasuk pola, keteraturan, tren, dan anomali yang dapat membantu dalam pengambilan keputusan bisnis. Data mining melibatkan penggunaan teknik-teknik analisis statistik, matematika, kecerdasan buatan, dan mesin learning untuk mengidentifikasi pola dan tren dalam data. Proses ini dapat melibatkan beberapa tahapan, termasuk preprocessing data, eksplorasi data, pemodelan, dan evaluasi. [2]

Prediksi adalah proses membuat perkiraan atau ramalan tentang suatu kejadian atau hasil di masa depan berdasarkan informasi atau data yang tersedia saat ini. Tujuan dari prediksi adalah untuk memperkirakan kemungkinan hasil atau kejadian di masa depan, sehingga dapat membantu pengambilan keputusan yang lebih baik dan lebih tepat. Prediksi dapat dilakukan dengan berbagai cara, termasuk penggunaan statistik, analisis data, atau teknik kecerdasan buatan seperti machine learning. Dalam banyak kasus, prediksi melibatkan penggunaan model matematika atau statistik yang dibuat berdasarkan data historis atau informasi lain yang relevan. [3]

Regresi linier adalah teknik analisis statistik yang digunakan untuk menemukan hubungan fungsional antara dua variabel, di mana satu variabel (variabel independen) mempengaruhi atau memprediksi nilai variabel lainnya (variabel dependen). Tujuan regresi linier adalah untuk menemukan garis terbaik (best fit line) yang dapat digunakan untuk memprediksi nilai variabel dependen berdasarkan nilai variabel independen yang diberikan. Regresi linier dapat digunakan dalam berbagai bidang, termasuk ilmu sosial, ekonomi, ilmu lingkungan, dan ilmu alam. Misalnya, regresi linier dapat digunakan untuk memprediksi penjualan produk berdasarkan biaya iklan, atau untuk menentukan hubungan antara suhu, penyakit dan kepadatan air di lingkungan alam. [4]

Algoritma regresi linier adalah serangkaian langkah atau prosedur yang digunakan untuk menemukan garis terbaik (best fit line) yang dapat digunakan untuk memprediksi nilai variabel dependen berdasarkan nilai variabel independen dalam analisis regresi linier. Algoritma ini melibatkan beberapa tahapan, termasuk: 1. Preprocessing data: Langkah awal adalah mengumpulkan data dan mempersiapkannya untuk analisis. Preprocessing data meliputi pembersihan data, penghapusan nilai kosong, dan normalisasi data jika diperlukan. 2. Pemilihan model: Tahap berikutnya adalah

memilih model yang akan digunakan untuk analisis regresi linier. Model yang paling umum digunakan adalah regresi linier sederhana atau regresi linier berganda, tergantung pada jumlah variabel independen yang digunakan. 3. Menentukan garis terbaik: Setelah model dipilih, garis terbaik harus ditentukan. Garis terbaik adalah garis yang paling dekat dengan titik data dalam model. Garis terbaik biasanya ditentukan dengan menggunakan metode minimum kuadrat terkecil. 4. Evaluasi model: Tahap terakhir adalah evaluasi model untuk memastikan bahwa garis terbaik yang ditemukan adalah model yang paling sesuai dengan data. Evaluasi model meliputi uji hipotesis, uji signifikansi, dan pengukuran kesalahan seperti koefisien determinasi. [5]

Data mining sering digunakan dalam berbagai bidang, seperti perbankan, e-commerce, kesehatan, dan ilmu pengetahuan. Contohnya, data mining dapat digunakan untuk menganalisis perilaku pelanggan, menemukan obat baru, mengoptimalkan rantai pasokan, dan memprediksi penyakit. Penting untuk dicatat bahwa data mining melibatkan pengolahan data pribadi dan perlu dilakukan dengan etika dan privasi yang memadai. [6]

Tujuan utama dari data mining adalah untuk menemukan informasi yang berharga dan tersembunyi dari data besar yang tersedia. Data mining memungkinkan pengguna untuk: 1. Menemukan pola dan hubungan dalam data: Data mining dapat membantu pengguna untuk menemukan pola dan hubungan yang tersembunyi dalam data besar yang sulit dipahami dengan cara konvensional. Contohnya, data mining dapat membantu dalam menemukan pola pembelian produk oleh pelanggan atau menemukan korelasi antara variabel dalam data kesehatan. 2. Memprediksi hasil: Data mining dapat membantu dalam memprediksi hasil berdasarkan pola dan hubungan yang ditemukan dalam data. Contohnya, data mining dapat membantu dalam memprediksi harga saham di masa depan berdasarkan pola harga saham masa lalu. 3. Meningkatkan efisiensi dan produktivitas: Data mining dapat membantu dalam meningkatkan efisiensi dan produktivitas dengan menemukan cara-cara baru untuk mengatur data dan proses bisnis. Contohnya, data mining dapat membantu dalam mengoptimalkan proses produksi untuk mengurangi biaya produksi. 4. Mendukung pengambilan keputusan: Data mining dapat memberikan wawasan yang berharga dan terperinci untuk membantu dalam pengambilan keputusan yang lebih baik dan berbasis data. Contohnya, data mining dapat membantu dalam menentukan strategi pemasaran berdasarkan data pelanggan dan pola pembelian produk. Dalam keseluruhan, tujuan dari data mining adalah untuk membantu pengguna memahami data yang tersedia dengan lebih baik dan mendapatkan wawasan yang berguna untuk mengambil keputusan bisnis yang lebih baik dan lebih efektif. [7]

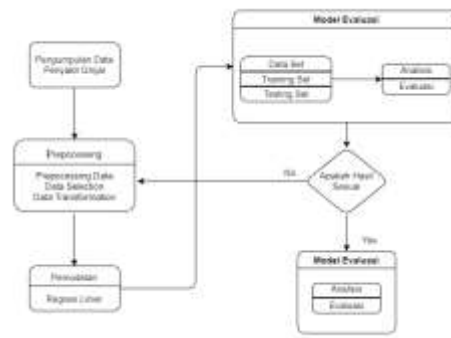
Proses tahapan data mining melibatkan enam tahapan, yaitu: 1. Pemahaman Bisnis: Tahap ini melibatkan pemahaman tentang masalah bisnis yang ingin dipecahkan atau tujuan yang ingin dicapai melalui analisis data. Pada tahap ini, kita perlu memahami tujuan dari proyek data mining, data yang tersedia, dan kebutuhan pengguna untuk mendefinisikan masalah bisnis yang ingin dipecahkan. 2. Pemahaman Data: Tahap ini melibatkan pengumpulan dan pemahaman tentang data yang tersedia. Pada tahap ini, kita perlu memahami struktur dan karakteristik data, melihat apakah ada data yang hilang atau tidak lengkap, dan mempersiapkan data untuk analisis. 3. Pre-processing Data: Tahap ini melibatkan pembersihan dan persiapan data yang akan digunakan dalam proses analisis data mining. Pada tahap ini, kita melakukan pemrosesan data seperti menghapus data yang tidak relevan, menangani data yang hilang atau tidak lengkap, dan transformasi data seperti normalisasi dan pengkodean data. 4. Analisis Data: Tahap ini melibatkan penerapan algoritma data mining pada data yang telah diproses. Pada tahap ini, kita melakukan eksplorasi data dan analisis statistik untuk mengidentifikasi pola dan hubungan yang berguna dalam mencapai tujuan bisnis. 5. Evaluasi Model: Tahap ini melibatkan evaluasi dan validasi model yang dihasilkan pada tahap analisis data. Pada tahap ini, kita menguji model dengan data yang belum pernah digunakan sebelumnya dan mengevaluasi performanya dalam memprediksi hasil. 6. Penerapan Model: Tahap ini melibatkan penggunaan model yang telah diuji pada tahap evaluasi untuk mengambil keputusan atau menghasilkan informasi berguna dalam masalah bisnis yang ingin dipecahkan. Tahap-tahap ini umumnya dilakukan secara berulang-ulang hingga mencapai hasil yang memuaskan dan sesuai dengan tujuan bisnis yang ingin dicapai. [8]

Tujuan utama prediksi adalah untuk memprediksi nilai atau kejadian masa depan berdasarkan informasi dan data yang tersedia. Dalam konteks bisnis, prediksi dapat membantu mengidentifikasi tren pasar, perilaku pelanggan, atau hasil bisnis yang mungkin terjadi di masa depan. Dengan memprediksi hasil masa depan, perusahaan dapat membuat keputusan yang lebih baik dalam hal pengelolaan stok, pemasaran, pengembangan produk, dan alokasi sumber daya. [9]

2. METODOLOGI PENELITIAN

2.1 Proses Tahapan Data Mining

Untuk memprediksi pada penderita penyakit kanker paru-paru yang akan di uji dengan memanfaatkan model data yang akan digunakan supaya memudahkan dalam proses penelitian, maka dibuat sebuah alur tahapan data mining sebagai berikut :


Gambar 1. Tahapan Penelitian

2.2 Data Selection

Data Selection (pemilihan data) adalah proses pemilihan data yang relevan dan bermakna dari suatu sumber data besar untuk digunakan dalam analisis lebih lanjut.

Tabel 1. Data Selection

Age	Passive Smoker	Smoking	Alcohol use	Obesity	Level
33	3	2	4	4	1
17	2	4	1	2	2
35	2	3	5	7	3
37	7	7	7	7	3
46	8	7	8	7	3
35	2	3	5	7	3
52	3	2	4	4	1
28	1	4	1	3	1
35	6	6	5	5	2
46	2	3	3	3	2
44	7	8	7	7	3
64	7	8	8	7	3
39	6	6	5	6	2
34	7	7	7	7	3

2.3 Data Transformation

Data Transformation (Transformasi Data) adalah proses mengubah format atau struktur data menjadi format yang berbeda untuk meningkatkan kualitas data dan memudahkan analisis lebih lanjut. Transformasi data melibatkan perubahan nilai data yang ada dalam satu atau beberapa variabel, serta mengubah struktur data, menggabungkan data dari berbagai sumber, dan menghapus data yang tidak relevan. Tujuan utama transformasi data adalah untuk meningkatkan kualitas data dan mempersiapkan data untuk analisis lebih lanjut.

Tabel 2. Data Transformation

Age	Passive Smoker	Level
33	3	1
17	2	2

35	2	3
37	7	3
46	8	3
35	2	3
52	3	1
28	1	1
35	6	2
46	2	2
44	7	3
64	7	3
39	6	2
34	7	3

2.4 Pemodelan

Pemodelan dalam Regresi Linier melibatkan proses menentukan hubungan antara variabel independen (x) dan variabel dependen (y) dalam bentuk persamaan linear. Tujuan dari pemodelan ini adalah untuk menggambarkan hubungan antara variabel independen dan dependen dan menggunakan model tersebut untuk memprediksi nilai dependen berdasarkan nilai independen. Terdapat beberapa tahap dalam proses pemodelan regresi linier, yaitu:

1. Penentuan Model: Tahap pertama dalam pemodelan regresi linier adalah menentukan model regresi yang paling sesuai. Ada beberapa jenis model regresi, seperti regresi linier sederhana, regresi linier berganda, regresi non-linier, dan regresi logistik.
2. Pengumpulan Data: Setelah model regresi dipilih, selanjutnya adalah mengumpulkan data yang diperlukan. Data ini harus terdiri dari setidaknya satu variabel independen dan satu variabel dependen, serta memiliki jumlah sampel yang cukup untuk melakukan analisis.
3. Persiapan Data: Data perlu dipersiapkan sebelum digunakan untuk membuat model. Tahap ini melibatkan pembersihan data, penghapusan data yang hilang, dan transformasi data.
4. Pengujian Model: Setelah persiapan data selesai, selanjutnya adalah menguji model untuk menentukan apakah model tersebut sesuai dengan data. Ada beberapa metode untuk menguji model, seperti uji signifikansi koefisien regresi, uji goodness-of-fit, dan uji heteroskedastisitas.
5. Evaluasi Model: Setelah model diuji, tahap selanjutnya adalah mengevaluasi model dan mengidentifikasi apakah ada masalah dalam model. Evaluasi model melibatkan analisis residual, analisis multikolinearitas, dan analisis outlier.
6. Pemilihan Model: Setelah model dievaluasi, langkah selanjutnya adalah memilih model terbaik berdasarkan kriteria yang telah ditentukan, seperti koefisien determinasi, signifikansi statistik, dan kesesuaian model dengan data.
7. Prediksi: Setelah model dipilih, model dapat digunakan untuk memprediksi nilai dependen berdasarkan nilai independen. Prediksi dilakukan dengan menghitung nilai y dari persamaan regresi menggunakan nilai x yang telah diberikan.

Bentuk Rumus Regresi Linear Sederhana:

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

$$a = \bar{y} - b\bar{x} \quad (2), \text{ sehingga}$$

$$a = \frac{\sum_{i=1}^n y_i}{n} - b \frac{\sum_{i=1}^n x_i}{n} \quad (3)$$



Keterangan :

n: banyaknya jumlah data

yi: variabel dependen (nilai yang diprediksikan)

xi: variabel independen

b : Koefisiensi Regresi (hasil perhitungan menggunakan rumus)

a : Konstanta (untuk menghitung hasil dari koefisien regresi)

Menghitung persamaan regresi linearnya

$$Y = a + bX$$

2.5 Evaluasi dan Hasil

Pada Evaluasi dan Hasil ini ditujukan untuk menganalisa sebuah dataset yang ada untuk diprediksi menggunakan metode regresi linier, cara yang pertama menggunakan cara hitung manual yaitu dengan rumus sederhana regresi linier, dan untuk memastikan tingkat akurasi menggunakan *tools* RapidMiner.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

3.1.1 Data Uji

Penelitian ini di uji dan diteliti menggunakan metode algoritma regresi linier. Hasil dari klasifikasi data penyakit kanker paru-paru menggunakan *tools* Rapid Miner mendapatkan hasil RMSE (*Root Mean Square Error*) yang baik dan akurat, maka dapat digunakan pada pasien penderita penyakit kanker paru-paru dan tingkat keparahan penderitanya. Sumber data berasal dari situs Kaggle.com, dan data yang digunakan dalam penelitian ini terdiri dari atribut atau variabel seperti usia, perokok pasif dan level.

3.1.2 Split Validation

Split validation adalah salah satu teknik evaluasi model yang digunakan dalam pembelajaran mesin (*machine learning*) dan data mining. Teknik ini digunakan untuk memeriksa seberapa baik model yang dibuat dapat memprediksi data baru yang belum pernah dilihat sebelumnya.

Tabel 3. Dataset

Age	Passive Smoker	Level
33	3	1
17	2	2
35	2	3
37	7	3
46	8	3
35	2	3
52	3	1
28	1	1
35	6	2
46	2	2
44	7	3
64	7	3
39	6	2
34	7	3
27	2	1



73	8	2
17	2	2
34	7	3
36	7	3
14	5	2
24	8	3
53	2	3
62	8	3
29	7	3
36	7	3
65	1	2

3.1.3 Perhitungan Regresi Linier

Didalam perhitungan Regresi Linier, untuk mencari nilai :

X_1Y : Usia X Level

X_2Y : Perokok Pasif X Level

X_1X_2 : Usia X Perokok Pasif

X_1^2 : Usia²

X_2^2 : Perokok Pasif²

Tabel 4. Perhitungan Regresi Linier

Usia (X_1)	Perokok Pasif (X_2)	Level (Y)	X_1Y	X_2Y	X_1X_2	(X_1) ²	(X_2) ²
33	3	1	33	3	99	1089	9
17	2	2	34	4	34	289	4
35	2	3	105	6	70	1225	4
37	7	3	111	21	259	1369	49
46	8	3	138	24	368	2116	64
35	2	3	105	6	70	1225	4
52	3	1	52	3	156	2704	9
28	1	1	28	1	28	784	1
35	6	2	70	12	210	1225	36
46	2	2	92	4	92	2116	4
44	7	3	132	21	308	1936	49
64	7	3	192	21	448	4096	49
39	6	2	78	12	234	1521	36
34	7	3	102	21	238	1156	49
27	2	1	27	2	54	729	4



BULLETIN OF INFORMATION TECHNOLOGY (BIT)

Vol 4, No 1, Maret 2023, Hal 63 - 74

ISSN 2722-0524 (media online)

DOI [10.47065/bit.v3i1.501](https://doi.org/10.47065/bit.v3i1.501)

<https://journal.fkpt.org/index.php/BIT>

73	8	2	146	16	584	5329	64
17	2	2	34	4	34	289	4
34	7	3	102	21	238	1156	49
36	7	3	108	21	252	1296	49
14	5	2	28	10	70	196	25
24	8	3	72	24	192	576	64
53	2	3	159	6	106	2809	4
62	8	3	186	24	496	3844	64
29	7	3	87	21	203	841	49
36	7	3	108	21	252	1296	49
65	1	2	130	2	65	4225	1
38	1	2	76	2	38	1444	1
19	2	2	38	4	38	361	4
33	4	3	99	12	132	1089	16
28	3	2	56	6	84	784	9
35	6	3	105	18	210	1225	36
42	7	3	126	21	294	1764	49
32	3	2	64	6	96	1024	9
33	3	1	33	3	99	1089	9
25	1	1	25	1	25	625	1
14	5	2	28	10	70	196	25
27	2	1	27	2	54	729	4
28	4	1	28	4	112	784	16
32	4	1	32	4	128	1024	16
45	6	1	45	6	270	2025	36
27	1	1	27	1	27	729	1
26	3	1	26	3	78	676	9
48	1	1	48	1	48	2304	1
17	3	1	17	3	51	289	9
22	6	1	22	6	132	484	36
42	1	1	42	1	42	1764	1
24	8	3	72	24	192	576	64
35	2	1	35	2	70	1225	4
24	2	1	24	2	48	576	4



BULLETIN OF INFORMATION TECHNOLOGY (BIT)

Vol 4, No 1, Maret 2023, Hal 63 - 74

ISSN 2722-0524 (media online)

DOI [10.47065/bit.v3i1.501](https://doi.org/10.47065/bit.v3i1.501)

<https://journal.fkpt.org/index.php/BIT>

38	3	1	38	3	114	1444	9
18	1	1	18	1	18	324	1
23	2	1	23	2	46	529	4
24	4	1	24	4	96	576	16
35	2	1	35	2	70	1225	4
38	2	1	38	2	76	1444	4
47	2	1	47	2	94	2209	4
52	7	1	52	7	364	2704	49
53	2	3	159	6	106	2809	4
44	7	1	44	7	308	1936	49
38	4	1	38	4	152	1444	16
62	6	1	62	6	372	3844	36
61	3	1	61	3	183	3721	9
55	3	1	55	3	165	3025	9
45	3	2	90	6	135	2025	9
38	1	2	76	2	38	1444	1
44	1	2	88	2	44	1936	1
45	2	2	90	4	90	2025	4
33	3	2	66	6	99	1089	9
62	8	3	186	24	496	3844	64
32	3	2	64	6	96	1024	9
44	3	2	88	6	132	1936	9
62	1	2	124	2	62	3844	1
38	1	2	76	2	38	1444	1
33	2	2	66	4	66	1089	4
22	1	2	44	2	22	484	1
35	1	2	70	2	35	1225	1
23	2	2	46	4	46	529	4
48	3	2	96	6	144	2304	9
46	3	2	92	6	138	2116	9
29	7	3	87	21	203	841	49
52	3	2	104	6	156	2704	9
52	1	2	104	2	52	2704	1
48	1	2	96	2	48	2304	1

36	1	2	72	2	36	1296	1
31	2	2	62	4	62	961	4
38	3	2	76	6	114	1444	9
35	1	2	70	2	35	1225	1
44	7	3	132	21	308	1936	49
33	7	3	99	21	231	1089	49
45	4	3	135	12	180	2025	16
65	1	2	130	2	65	4225	1
53	8	3	159	24	424	2809	64
35	2	3	105	6	70	1225	4
46	8	3	138	24	368	2116	64
27	7	3	81	21	189	729	49
26	7	3	78	21	182	676	49
37	7	3	111	21	259	1369	49
28	7	3	84	21	196	784	49
19	7	3	57	21	133	361	49
29	2	3	87	6	58	841	4
3757	384	202	7677	875	14712	157479	2074

Telah dilakukan sebuah pencarian dan olahan data untuk mencari nilai regresi linier sederhana dengan memanfaatkan atribut yang ada. Setelah di dapatkan semua angka dijumlahkan lah di hasil akhir dengan rumus (=SUM) untuk menemukan sebuah nilai untuk menghitung prediksi menggunakan cara manual atau hitung sesuai menggunakan rumus regresi linier sederhana.

$$b_1 = \frac{n \sum(x_1y) - (\sum x_1)(\sum y)}{n(\sum x_1)^2 - (\sum x_1)^2}$$

$$b_1 = \frac{100(7677) - (3757)(202)}{100(157479) - (3757)^2}$$

$$b_1 = \frac{767700 - 758914}{15747900 - 14115049}$$

$$b_1 = \frac{8786}{1632851}$$

$$b_1 = 0,00538077264857$$

Setelah mendapatkan nilai b1, selanjutnya mencari nilai b2

$$b_2 = \frac{n \sum(x_2y) - (\sum x_2)(\sum y)}{n(\sum x_2)^2 - (\sum x_2)^2}$$

$$b_2 = \frac{100(875) - (384)(202)}{100(2074) - (384)^2}$$

$$b_2 = \frac{87500 - 77568}{207400 - 147456}$$

$$b_2 = \frac{9932}{59944}$$

$$b_2 = 0,165687975443747$$

Setelah mendapatkan nilai b1 dan b2, selanjutnya mencari nilai a menggunakan rumus yang ada dan hasil sebagai berikut :

$$a = \frac{\sum y - b_1 \sum x_1 + b_2 \sum x_2}{n}$$

$$a = \frac{202 - ((0,00538077264857) * (3757)) + ((0,165687975443747) * (384))}{100}$$

$$a = \frac{202 - (20,21556284067749) + 63,62418257039885}{100}$$

$$a = \frac{202 - 83,83974541107634}{100}$$

$$a = \frac{118,1602545889237}{100}$$

$$a = 1,181602545889237$$

Diambil dari beberapa contoh data random untuk memprediksi data penyakit kanker paru-paru menggunakan cara hitung manual untuk mengetahui akurat atau tidaknya sebagai berikut :

Tabel 5. Kolom data random prediksi penyakit ginjal

No	Usia (X ₁)	Perokok Pasif (X ₂)	Level (Y)
1	33	3	?
2	29	2	?
3	35	2	?

Menghitung Persaman Regresi Linear

$$Y = a + b_1X_1 + b_2X_2$$

$$y = 1,181602545889237 + (0,00538077264857x \ 33) + (0,165687975443747 \ x \ 3)$$

$$= 1,856231969623288$$

$$y = 1,181602545889237 + (0,00538077264857 \ x \ 29) + (0,165687975443747 \ x \ 2)$$

$$= 1,669020903585261$$

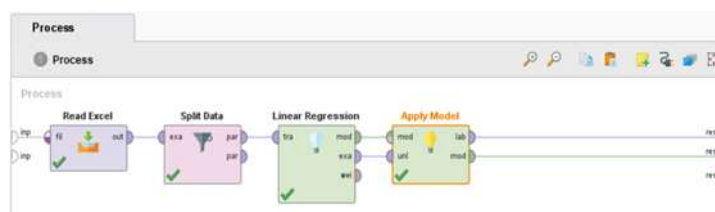
$$y = 1,181602545889237 + (0,00538077264857 \ x \ 35) + (0,165687975443747 \ x \ 2)$$

$$= 1,701305539476681$$

3.2 Pembahasan

3.2.1 Proses Pengujian Data (RapidMiner)

Melakukan select *attributes* yaitu untuk mengetahui hasil prediksi dari rapidminer, hasil perhitungan manual dan hasil uji di rapidminer.



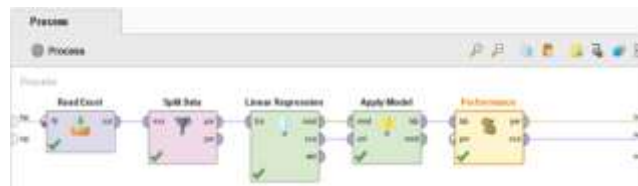
Gambar 2. Proses RapidMiner

Pada proses ini untuk memasukan data training dan data testing yang akan diuji untuk menghasilkan prediksi pada *atribute class* yang ditampilkan pada tabel 6.

Tabel 6. Hasil Prediksi

Row No.	Level	prediction(L	Age	Smoking
1	1	1.861	33	3
2	2	1.695	17	2
3	3	1.695	35	2
4	3	2.526	37	7
5	3	2.692	46	8
6	3	1.695	35	2
7	1	1.861	52	3
8	1	1.529	28	1
9	2	2.360	35	6
10	2	1.695	46	2
11	3	2.526	64	7
12	2	2.360	39	6
13	3	2.526	34	7
14	1	1.695	27	2
15	2	2.692	73	8
16	2	1.695	17	2

Ketika prediksi sudah ketemu, langkah selanjutnya mencari tingkat akurasi yang ada menggunakan Performance pada gambar 3.



Gambar 3. Proses Pencarian *Root Mean Squared Error* dan *Squared Error*

Untuk mempermudah dalam pembacaan akurasi data penyakit ginjal, maka perlu di masukan *tools performance* untuk mencari RMSE (*Root Mean Squared Error*) dan *Squared Error*.

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.686 +/- 0.000
squared_error: 0.471 +/- 0.546
```

Gambar 4. Hasil Pengujian *Root Mean Square Error* dan *Square Error*

3.2.2 Analisa Hasil

Berdasarkan hasil pengujian yang telah dilakukan menggunakan *tools* RapidMiner bahwa atribut (usia, perokok pasif dan level) berpengaruh signifikan terhadap penelitian ini terbukti dengan menggunakan algoritma regresi linier mampu memberikan hasil yang sangat baik dan sangat akurat dengan nilai *Root Mean Squared Error*: 0.686 +/- 0.000 dan *Squared Error*: 0.471 +/- 0.546.

4. KESIMPULAN

Algoritma regresi linier ini sangat efektif untuk digunakan dalam memprediksi penderita penyakit kanker paru-paru berdasarkan analisis dari dataset yang ada. Kemampuan meng-klasifikasi algoritma regresi linier sangat baik dalam memprediksi penderita penyakit kanker paru-paru dengan memanfaatkan data atribut (usia, perokok pasif dan level). Hasil prediksi dari algoritma regresi linier sangat baik dan akurat dibandingkan dengan metode lain dalam memprediksi risiko penyakit kanker paru-paru. Proses pengujian data pada penelitian ini menggunakan algoritma regresi linear mampu memberikan hasil yang baik dengan nilai *Root Mean Squared Error*: 0.686 +/- 0.000 dan *Squared Error*: 0.471 +/- 0.546.



UCAPAN TERIMA KASIH

Tidak luput saya mengucapkan terima kasih kepada segenap jajaran pihak dosen yang telah membantu untuk menyelesaikan penelitian ini, orang tua dan teman-teman yang selalu men-*support* saya dalam keadaan suka maupun duka.

REFERENCES

- [1] N. Aifa Zahara and M. Fakultas Tarbiyah dan Keguruan Program Studi Pendidikan Biologi FAKULTAS TARBIYAH DAN KEGURUAN, “PENGEMBANGAN MEDIA PEMBELAJARAN AUDIOVISUAL BERBASIS NILAI ALQURAN PADA MATERI SISTEM PERNAPASAN DI SMP NEGERI 2 KABUPATEN ACEH BESAR SKRIPSI”, 2022
- [2] Y. Mahena, M. Rusli and E. Winarso, “Prediksi Harga Emas Dunia Sebagai Pendukung Keputusan Investasi Saham Emas Menggunakan Teknik Data Mining” JURNAL Sains dan Teknologi, vol. 2, no. 1. 2019
- [3] M. Kafil, “PENERAPAN METODE K-NEAREST NEIGHBORS UNTUK PREDIKSI PENJUALAN BERBASIS WEB PADA BOUTIQ DEALOVE BONDOWOSO,” 2019.
- [4] S. Arafah and Y. Tanjung, “ANALISIS FAKTOR DETERMINAN YANG MEMPENGARUHI PEMAKAIAN METODE JIT (Studi Kasus UD. Pusaka Bakti).” Jurnal Ekonomi Syariah, 2020 [Online]. Available: <http://ejournal.unhasy.ac.id/index.php/bisei/article/view/440>
- [5] S. Kasus, L. Pendidikan Darul Ulum Bantaran Probolinggo, M. Iqbal, and J. Matematika Fakultas Sains Dan, “ANALISIS TREND LINIER DENGAN METODE KUADRAT TERKECIL UNTUK MERAMALKAN PERKEMBANGAN BANYAKNYA SISWA,” 2000.
- [6] S. Anastassia Amellia Kharis and A. Haqqi Anna Zili, “Learning Analytics dan Educational Data Mining pada Data Pendidikan,” *Jurnal Riset Pembelajaran Matematika Sekolah*, vol. 6, 2022.
- [7] D. A. Manalu and G. Gunadi, “IMPLEMENTASI METODE DATA MINING K-MEANS CLUSTERING TERHADAP DATA PEMBAYARAN TRANSAKSI MENGGUNAKAN BAHASA PEMROGRAMAN PYTHON PADA CV DIGITAL DIMENSI,” *Infotech: Journal of Technology Information*, vol. 8, no. 1, pp. 43–54, Jun. 2022, doi: 10.37365/jti.v8i1.131.
- [8] E. A. Firdaus, S. Maulani, A. B. Dharmawan, A. Keperwatan, and R. S. Dustira Cimahi, “PENGUKURAN MINAT BACA MAHASISWA DENGAN METODE CLUSTERING DI PERPUSTAKAAN AKADEMI KEPERAWATAN RS.DUSTIRA CIMAHI MENGGUNAKAN DATA MINING,” vol. 15, no. 1, 2021, [Online]. Available: <https://journal.uniku.ac.id/index.php/ilkom>
- [9] A. R. Wijaya, “MODEL PREDIKSI DATA HARGA MINYAK MENTAH DUNIA DENGAN METODE EXPONENTIAL SMOOTHING,” 2023.