Jurnal Dinamika Informatika
Volume 15, No 1, Maret 2026
ISSN 1978-1660 : 1-15
ISSN *online* 2549-8517

Analisis Prediksi Gender pada Data Member Gym Menggunakan Algoritma Logistic Regression

Riki Hasan

Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Pancasakti Tegal, Indonesia

Email: **rikihasan626@gmail.com**

Intisari

Data kesehatan dan kebugaran fisik memiliki pola yang berbeda dan unik di setiap jenis kelamin. Namun, pemanfaatan data tersebut untuk identifikasi seringkali belum mencapai hasil yang optimal. Penelitian ini bertujuan untuk menerapkan teknik data mining untuk mengklasifikasikan jenis kelamin berdasarkan atribut berat badan, jenis olahraga dan *Body Mass Index* (BMI). Algoritma klasifikasi yang digunakan untuk membangun model yaitu algoritma *Logistic Regression*. Evaluasi pada model dilakukan dengan menggunakan metode *Cross Validation* untuk memastikan bahwa model valid saat pengujiannya. Berdasarkan hasil penelitian, model yang dibuat mampu menghasilkan kinerja yang baik dengan akurasi mencapai 96,61%. Hasil evaluasi *confusion matrix* juga menunjukkan hasil nilai *recall* dan presisi memiliki nilai yang seimbang di atas 95% untuk kedua kelas prediksi. Hal ini membuktikan bahwa atribut fisik dan kebiasaan olahraga yang diimplementasikan merupakan variabel yang signifikan dalam penentuan klasifikasi jenis kelamin.
Kata kunci: Data Mining, Klasifikasi, Logistic Regression, Indeks Massa Tubuh (BMI), cross validation.

Abstract

Health and physical fitness data exhibit distinct and unique patterns across genders. However, utilizing this data for identification often fails to achieve optimal results. This study aims to apply data mining techniques to classify gender based on weight, exercise type, and body mass index (BMI). The classification algorithm used to build the model is the Logistic Regression algorithm. The model was evaluated using the Cross Validation method to ensure its validity during testing. The research results show that the model produced by Mamou performed well with an accuracy of 96.61%. The confusion matrix evaluation also showed balanced recall and precision values above 95% for both prediction classes. This demonstrates that physical attributes and exercise habits are significant variables in determining gender classification.

Keywords: Data Mining, Gym, Classification, Logistic Regression, Body Mass Index (BMI), cross validation.

PENDAHULUAN

Saat ini sudah tidak asing lagi dengan pusat kebugaran atau yang sering disebut dengan gym. Banyak orang yang berlomba-lomba ingin mendapatkan badan yang proporsional dan kekar berotot. Tapi tidak jarang juga orang-orang pergi ke pusat kebugaran hanya untuk dijadikan hobi dan tempat untuk melepas penat yang tidak melulu untuk membentuk otot. Dari banyaknya orang yang datang berolahraga, tentunya memiliki karakteristik tubuh yang berbeda dan jenis latihan yang bermacam-macam pula.

Penelitian ini akan mengklasifikasikan dataset anggota pusat kebugaran sebagai data acuan untuk melihat bahwa setiap individu memiliki karakteristik yang berbeda. Metode *logistic regression* berguna untuk memodelkan variabel dependen dengan variabel independen[4]. Metode ini akan mengklasifikasikan karakteristik profil anggota gym (laki-laki dan perempuan) untuk personalisasi program diet ataupun program latihan. Data mining berguna untuk mengamati apakah terdapat pola fisik yang jelas yang membedakan keduanya.

Agar memudahkan saat mengolah data, pada penelitian ini menggunakan bantuan dari *software* RapidMiner yang sudah banyak digunakan dan menjadi rekomendasi saat akan mengolah data mining menggunakan metode apapun. Dataset dapat berupa file berformat csv, xls, dbf, dml dan lain-lain. File tersebut dapat diambil dari *website* penyedia dataset.

METODE PENELITIAN

1. Pengumpulan Data

   Kumpulan data atau yang biasa disebut dengan dataset dapat dicari melalui *website* penyedia dataset yang banyak beredar di internet. Ada beberapa *website* penyedia dataset secara gratis di internet, misalnya Kagle, UCI Machine Learning Repository dan lain-lain. Pada penelitian ini, *website* yang dipakai untuk mencari dataset yaitu Kagle. Penyeleksian data dilakukan untuk mengambil data dari banyaknya dataset yang akan diolah[1]. Dataset yang digunakan dibuat oleh Sayid Vala Khorasani dan dataset tersebut dapat diunduh dari tautan berikut ini https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset.

   Untuk mempermudah dalam penelitian, dataset perlu diterjemahkan ke dalam Bahasa Indonesia. Data yang diubah dari dataset tersebut mencakup nama atributnya. Hal ini dilakukan untuk meminimalisir bahasa asing dan memudahkan pemahaman.

Jurnal Dinamika Informatika
Volume 13, No 2, Oktober 2024
ISSN 1978-1660 : 32-42
ISSN *online* 2549-8517
DOI: https://doi.org/10.31316/jdi.v14i1.386

Tabel 1. Terjemahan Dataset Asli ke Dalam Bahasa Indonesia

Atribut asli (dataset)	Atribut yang diterjemahkan	Satuan / keterangan
Age	Usia	Tahun
Gender	Jenis kelamin	Laki-lai atau perempuan
Weight	Berat badan	Kg
Height	Tinggi badan	Meter
Max bpm	Denyut jantung maksimal	Detak per menit
Avg bpm	Rata-rata detak jantung	Detak per menit
Resting bpm	Detak jantung istirahat	Detak per menit
Session duration	Durasi latihan	Jam
Caloried burned	Kalori terbakar	Kkal
Workout type	Jenis olahraga	*Cardio, strength, yoga, hiit*
Fat persentage	Persentasi lemak	Persentase
Water intake	Asupan air	Liter
Workout frequency	Frekuensi latihan	Hari / minggu
Experience level	Tingkat pengalaman	Numerik
Body mass index	Indeks massa tubuh	Numerik

Sumber: Dokumentasi Pribadi

Setiap atribut memiliki jenis data yang berbeda, terdapat dua jenis data yaitu data numerik dan data non-numerik. Data numerik yaitu data yang berisi bilangan bulat ataupun bilangan pecahan, sedangkan data non-numerik berisi sebuah teks, simbol, deskripsi dan lain-lain yang tidak dapat dimanipulasi secara matematis. Berikut merupakan jenis data yang dihimpun dari setiap atribut.

Tabel 2. Kumpulan Data Latihan Anggota Pusat Kebugaran

Kolom	Jenis	Keterangan
Usia	Numerik	Usia anggota pusat kebugaran
Jenis kelamin	Non-numerik	Jenis kelamin anggota pusat kebugaran
Berat badan	Numerik	Berat anggota dalam satuan kg
Tinggi badan	Numerik	Tinggi anggota dalam satuan meter
Denyut jantung maksimal	Numerik	Denyut jantung maksimum selama sesi latihan
Rata-rata detak jantung	Numerik	Denyut jantung rata-rata selama sesi latihan
Detak jantung istirahat	Numerik	Denyut jantung saat istirahat sebelum latihan
Durasi latihan	Numerik	Durasi setiap sesi latihan dalam jam

Kolom	Jenis	Keterangan
Kalori terbakar	Numerik	Total kalori yang terbakar setiap sesi
Jenis olahraga	Non-numerik	Jenis latihan yang dilakukan
Persentasi lemak	Numerik	Persentase lemak tubuh anggota tubuh
Asupan air	Numerik	Asupan air harian selama latihan
Frekuensi latihan	Numerik	Jumlah sesi latihan per minggu
Tingkat pengalaman	Numerik	Tingkat pengalaman anggota pusat kebugaran
Indeks massa tubuh	Numerik	Kalkulasi dari tinggi dan berat badan

Sumber: Dokumentasi Pribadi

2. *Logistic Regression*

Regresi Logistik adalah suatu metode analisis statistika untuk mendeskripsikan hubungan antara respon *dependent* variabel yang memiliki dua kategori atau lebih dengan satu atau lebih peubah penjelas *independent* variabel berskala kategori atau interval[8]. Secara teori, *logistic regression* merupakan regresi *non-linear* yang berguna sebagai penghitung hubungan antara X dan Y yang bersifat *non-linear*.

Karena metode ini bersifat *biner*, maka akan memunculkan hasil antara *false* dan *true*. Pada implementasinya secara manual akan menggunakan rumus seperti berikut:

$$y = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

keterangan:

a. y merupakan peluang atau kemungkinan dari kejadian sukses atau *true*
b. e merupakan bilangan *euler* yang bernilai sekitar 2.71828
c. β_0 merupakan intersep
d. β_1,β_2…β_n merupakan koefisien regresi sebagai penunjuk variabel independen

3. Aplikasi Rapidminer

RapidMiner yaitu sebuah aplikasi atau perangkat lunak untuk menganalisis data dan pengembangan dari kumpulan data yang dirangkum menjadi data[6]. Aplikasi ini bersifat *open source* dan bisa digunakan secara gratis di sistem operasi manapun seperti Windows, Linux, Mac OS dan lain-lain. Penelitian ini menggunakan bantuan RapidMiner untuk mempermudah penghitungan dan juga mempermudah analisis dataset anggota pusat kebugaran. Dengan menggunakan teknologi Java maka dataset akan dianalisis menggunakan operator-operator yang telah disediakan oleh RapidMiner,

Jurnal Dinamika Informatika
Volume 13, No 2, Oktober 2024
ISSN 1978-1660 : 32-42
ISSN *online* 2549-8517
DOI: https://doi.org/10.31316/jdi.v14i1.386

operator-operator tersebut berguna sebagai blok untuk menyusun struktur-struktur data mining yang akan digunakan[2].

RapidMiner memberikan tampilan *User Interface* (UI) yang *user friendly* yang membuatnya unggul karena mudah melacak *error* saat ada kesalahan dan mudah diimplementasikan. Tata letaknya sederhana, tidak berbelit-belit dan semua aksesnya terdapat pada layar utama. Banyak terdapat operator pada RapidMiner yang mendukung proses *end-to-end: import, evaluasi, modelling, preprocessing*[5]. Terdapat empat menu utama yang memudahkan para pengguna, yaitu *repository, operators, process, dan parameters.*

4. Operator Dalam Rapidminer

Operator dalam RapidMiner yaitu sebuah fungsi rumus atau *block kode* yang telah dirangkai menjadi sebuah *shortcut* yang siap digunakan untuk keperluan proses data mini. Setiap operator memiliki *port input* dan *output* serta memiliki operator untuk memberi nilai serta acuan sebagai pengendali nilai *input* atau *output*. yang bisa terhubung satu sama lain menjadi rangkaian sebuah blok analisis. Pada *software* RapidMiner ini terdapat lebih dari 1000 operator yang bisa digunakan.

Saat akan melakukan validasi menggunakan salah satu metode data mining, terkadang akan dihadapkan dengan tiga operator. Operator tersebut seringkali dipakai dan memang sudah menjadi *template* di *software* tersebut. Misalnya ingin menggunakan metode *logistic regression*, maka operator yang digunakan yaitu *logistic regression*, *apply models* dan *performance*. Untuk melakukan hal tersebut RapidMiner telah menyiapkan sebuah pengelompokan operator-operator metode data mining yang disebut dengan *a cross-validation* yang berguna untuk mempermudah dalam mengelompokkan operator secara ringkas, pengelolaan operator yang kompleks, serta memudahkan dalam pemantauan data.

HASIL DAN PEMBAHASAN

1. Pembagian Dataset dan Skema Validasi

Sebelum melakukan pengujian dataset, data harus diolah terlebih dahulu agar mudah saat diprediksi. Pengolahan data menghasilkan *label* dan pembagian data. *Label* berfungsi sebagai target yang ingin diprediksi dan akan diuji nantinya. Data harus dibagi terlebih dahulu menjadi dua bagian, yaitu *data training* dan *data testing* [7]. Data *training* akan digunakan untuk acuan yang akan di proses dan dicari polanya untuk diprediksi, sedangkan data *testing* digunakan untuk evaluasi kinerja model yang sudah di cari polanya. Tujuan dari data testing yaitu untuk mengukur seberapa baik prediksi yang diambil dari data training setelah diproses.

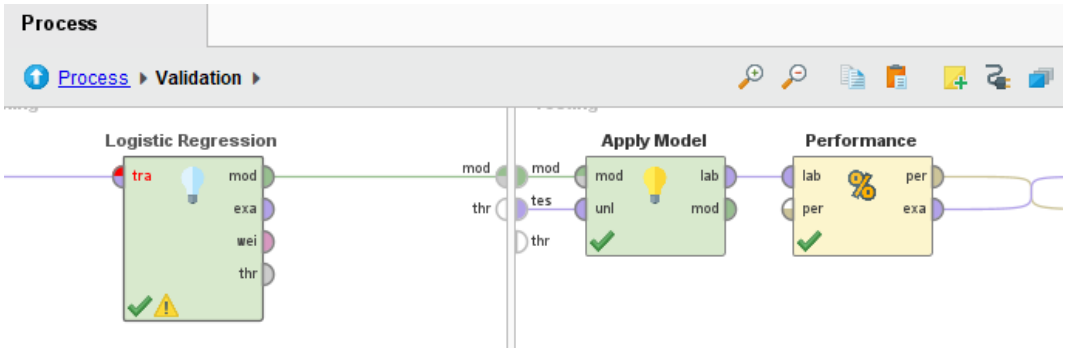
Pada penelitian ini mengimplementasikan metode *10-Fold Cross Validation,* metode tersebut akan secara otomatis membagi dataset menjadi

90:10. Perbandingan itu mencakup data training sebesar 90% dan data testing 10%. Penggunaan 10-Fold Cross Validation dipilih karena setiap data akan mendapat kesempatan untuk menjadi data uji, sehingga hasil akurasinya merupakan hasil rata-rata yang stabil dan tidak bias.

2. Pengujian Dataset

Pada bagian ini dataset sudah siap untuk dilakukan pengujian menggunakan metode *logistic regression*. pengujian ini bisa berjalan dengan menggunakan tiga operator utama yaitu operator *logistic regression*, *apply models* dan *performance*. Pengujian dataset dilakukan untuk mengukur akurasi dan kehandalan model yang dibangun. Pada penelitian ini, metode validasi yang digunakan yaitu *cross validation*. Metode tersebut dipilih untuk memastikan pengujian dilakukan secara objektif dengan membagi seluruh data pada dataset menjadi data testing dan training.

Proses pengujian dilakukan dengan membagi dataset ke dalam lipatan (*fold*) yang banyak datanya sama besar. Proses evaluasi dilakukan sebanyak jumlah *fold* yang telah ditentukan. Pada setiap *fold*, satu blok data akan digunakan menjadi data uji untuk validasi model, sedangkan blok data sisanya digunakan sebagai data *training*. Proses ini akan berulang hingga setiap blok pernah menjadi data uji satu kali.



Sumber : Dokumentasi Pribadi
Gambar 1. Proses cross validation

3. *Performance Vector*

Evaluasi kinerja model dilakukan dengan menganalisis *performace vector* yang menghasilkan sebuah *confusion matrix*. Matrix ini memberikan gambaran detail mengenai perbandingan prediksi sistem dengan label kelas yaitu atribut jenis kelamin. Matrix yang digunakan sebagai tolak ukur keberhasilan model meliputi *Accuracy*, *Precision*, dan *Recall*.

Jurnal Dinamika Informatika
Volume 13, No 2, Oktober 2024
ISSN 1978-1660 : 32-42
ISSN *online* 2549-8517
DOI: https://doi.org/10.31316/jdi.v14i1.386

Hasil dari pengujian dapat dilihat pada gambar berikut:

accuracy: 96.61% +/- 2.38% (micro average: 96.61%)

	true Male	true Female	class precision
pred. Male	490	12	97.61%
pred. Female	21	450	95.54%
class recall	95.89%	97.40%	

Sumber: Dokumentasi Pribadi
Gambar 2. Akurasi dan Confusion Matrix

Berdasarkan gambar *performance vector* diatas, menunjukkan bahwa pengujian mendapatkan tingkat akurasi sebesar 96,61% dengan *deviasi* standar +/- 2.38%. Nilai akurasi ini menunjukan bahwa model sangat andal dalam mengklasifikasikan ke dalam kelas yang valid.
Rincian distribusi prediksi pada confusion matriks sebagai berikut:

a. Prediksi kelas dengan jenis kelamin Laki-Laki
   Model telah berhasil mengklasifikasikan sebanyak 490 data sebagai laki-laki (*true male*). Akan tetapi, terdapat kesalahan prediksi yang dimana sebanyak 21 data yang seharusnya laki-laki terprediksi sebagai perempuan. Nilai *class precision* untuk kelas ini mencapat 97,61% sedangkan *class recall* sebesar 95,89%.
b. Prediksi kelas dengan jenis kelamin perempuan
   Model berhasil mengklasifikasikan sebanyak 450 data sebagai perempuan (*true female*). Akan tetapi, terdapat kesalahan prediksi dimana 12 data yang seharusnya berjenis kelamin perempuan terprediksi sebagai laki-laki. Nilai *class precision* untuk kelas ini yaitu 95,54% dengan *class recall* yang lebih tinggi yaitu sebesar 97,40%.

KESIMPULAN

Berdasarkan hasil penelitian, penerapan data mining menggunakan algoritma *logistic regression* berhasil dalam mengenali pola hubungan antara data fisik seperti data berat badan dan BMI serta aktivitas olahraga terhadap label jenis kelamin. Pengujian performa model menggunakan metode *cross validation* menghasilkan akurasi sebesar 96,61% dengan *deviasi* standar +/- 2,38%. Hasil ini menunjukkan bahwa model memiliki tingkat kehandalan yang tinggi dan stabil saat melakukan prediksi.

Berdasarkan hasil analisis *confusion matrix*, sistem mampu mengenali kelas laki-laki dan perempuan dengan sangat baik yang dibuktikan dengan persentase *class precision* dan *class recall* yang stabil di atas 95%, sehingga minim kesalahan dalam prediksi.

SARAN

Guna pengembangan lebih lanjut agar mendapatkan hasil yang lebih komprehensif. Penelitian selanjutnya disarankan untuk menambah variabel atau atribut lain yang relevan seperti tinggi badan, usia atau durasi latihan untuk melihat pengaruhnya terhadap akurasi prediksi. Diharapkan penelitian selanjutnya dapat menambahkan dataset latih yang lebih besar dan bervariasi untuk meningkatkan kemampuan model terhadap data baru yang lebih kompleks dan dapat juga dilakukan perbandingan kinerja dengan algoritma klasifikasi yang lain untuk mengetahui metode mana yang paling optimal untuk karakteristik data kebugaran ini.

DAFTAR PUSTAKA

[1] Alghifari, F., & Juardi, D. (2021). Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes. Jurnal Ilmiah Informatika, 9.

[2] Apriyadi, Lubis, M. R., & Damanik, B. E. (2022). PENERAPAN ALGORITMA C5.0 DALAM MENENTUKAN TINGKAT PEMAHAMAN MAHASISWA TERHADAP PEMBELAJARAN DARING. Jurnal Ilmiah Komputer Dan Informatika, 11(1)..

[3] Darmawan, R., & Amini, S. (2022). Perbandingan Hasil Sentimen Analysis Menggunakan Algoritma Naïve Bayes dan K-Nearest Neighbor pada Twitter. Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) Jakarta-Indonesia, 495–501. https://senafti.budiluhur.ac.id/index.php.

[4] Noviandri, Adhyanti, Ansar Mursaha, & Nur M Ali. (2024). Analisis Regresi Logistik Biner Pada Karakteristik Demografi Sosial Ekonomi Dan Intervensi Gizi Spesifik Balita Stunting Di Wilayah Kerjapuskesmas Jambula Kota Ternate. Jurnal Ilmiah Kesehatan Sekolah Tinggi Ilmu Kesehatan Majapahit, 16, 127–278.

[5] Pangestu, B., Nugroho, B. A., Ramadhani, D., Juliana, M. F., Hidayat, S., & Fansyuri, M. (2025). Penerapan Algoritma K-Nearest Neigh bor Menggunakan RapidMiner Pada Kepuasan Hidup Pekerja Commuter di Indonesia. PT Jurnal Cendekia Indonesia, 1(1), 1–6.

[6] Prasetyo, V. R., Lazuardi, H., Mulyono, A. A., & Lauw, C. (2021). Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Linear Regression. Jurnal Nasional Teknologi Dan Sistem Informasi, 7(1), 8–17. https://doi.org/10.25077/teknosi.v7i1.2021.8-17.

[7] Pratama, Y., & Hidayat, N. (2024). Komparasi Algoritma Support Vector Machine dan Logistic Regression dalam Klasifikasi Data Kesehatan Masyarakat. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK), 8(2), 245-254. Universitas Brawijaya.

[8] Suprayogi, M. A. (2022). Analisis Regresi Logistik Biner Pada Faktor-Faktor Yang Memengaruhi Evaluasi Kinerja Barang Milik Negara Di Provinsi Dki

Jurnal Dinamika Informatika
Volume 13, No 2, Oktober 2024
ISSN 1978-1660 : 32-42
ISSN *online* 2549-8517
DOI: https://doi.org/10.31316/jdi.v14i1.386

Jakarta. Journal of Statictic and Its Applications, 4(1), 35–45. https://doi.org/10.30598/variancevol4iss1.