Versi Online tersedia di : https://jurnal.buddhidharma.ac.id/index.php/algor/index

JURNAL ALGOR

|2715-0577 (Online)| 2715-0569 (Print)

Artikel

Penerapan Data Mining Untuk Prediksi Kualitas Sperma (*Fertility*) Berbasis *Android*

Steven Hosana[1], *Dicky Surya Dwi Putra*[2]

[1, 2] *Universitas Buddhi Dharma, Teknik Informatika, Banten, Indonesia*

SUBMISSION TRACK

Recieved: 28 Agustus 2019
Final Revision: 25 September 2019
Available Online: 30 September 2019

KATA KUNCI

Fertilitas, Sperma, Data *Mining*, Algoritma *Naïve Bayes*

KORESPONDENSI

[1]E-mail: stevenhosana18@gmail.com

[2]E-mail: dsurya_eboh@yahoo.co.id

A B S T R A K

Kesuburan (Fertilitas) adalah dapat bekerjanya secara optimal organ-organ reproduksi baik, pada pria maupun wanita, sehingga dapat melakukan fungsi fertilisasi dengan baik. Sperma dibuat di testis dalam proses yang dinamakan spermatogenesis, kemudian disimpan jauh sampai sperma matang dan siap untuk meninggalkan tubuh. Ketika seorang laki-laki berejakulasi, sperma matang (dikombinasikan dengan sisa cairan yang membentuk air mani) dikeluarkan dari uretra, terletak di ujung penis.Beberapa studi klinis mengamati sekitar 10% pasangan mengalami infertilitas. Faktor penyebab infertilitas dapat berasal dari suami, istri atau keduanya. Menurut penelitian yang dilakukan oleh Lim dan Ratnam, faktor-faktor penyebab yang berasal dari suami mencapai 33%, sedangkan dari penelitian *WHO* pada tahun 1989 menunjukkan bahwa infertilitas pada pria menyumbang 40% dari kasus infertilitas dan dibandingkan dengan studi Arsyad terhadap 246 pasangan infertil, kasus infertilitas pada pria berjumlah 48,4%. Atas dasar ini infertilitas pada pria merupakan faktor signifikan sebagai penyebab infertilitas. Teknik data *mining* yang dapat digunakan untuk mengklasifikasi ataupun prediksi yang dapat menggunakan bantuan dari aplikasi *RapidMiner*. Penulis tertarik untuk melakukan penelitian terhadap prediksi kualitas sperma menggunakan algoritma *Naive Bayes* dan menerapkannya ke dalam *smartphone Android*.

I. PENDAHULUAN

Indonesia sebagai negara berkembang yang memiliki populasi sebanyak 237.641.326 jiwa pada tahun 2010 yang mencakup mereka yang bertempat tinggal di daerah perkotaan sebanyak 118.320.256 jiwa (49,79 persen) dan di daerah pedesaan sebanyak 119.321.070 jiwa (50,21 persen). 119.630.913 jiwa (50,34 persen) dari penduduk Indonesia adalah laki-laki sedangkan perempuan berjumlah 118.010.413 jiwa (49,66 persen)[1].

Kesuburan (*Fertilitas*) adalah dapat bekerjanya secara optimal organ-organ reproduksi baik, pada pria maupun wanita, sehingga dapat melakukan fungsi fertilisasi dengan baik. Banyak faktor yang mempengaruhi kesuburan dan keberhasilan pembuahan sel telur oleh sperma, serta tumbuh kembang janin agar lahir sebagai bayi yang normal dan sehat[2].

Pada saat ini perkembangan teknologi informasi sangatlah cepat, sehinga dengan teknologi-teknologi yang ada saat ini dapat mempermudah kegiatan yang dilakukan oleh manusia di berbagai macam bidang. Salah satunya adalah teknik data *mining* yang dapat digunakan untuk mengklasifikasi ataupun prediksi. Data *mining* adalah proses ekstraksi kumpulan data-data untuk mendapatkan informasi penting yang sebelumnya tidak diketahui. Data *mining* dilakukan untuk mencari sebuah pola di dalam data yang nantinya apat menghasilkan prediksi yang akurat pada data di masa yang akan datang. Menentukan metde data mining yang tepat tidaklah mudah karena setiap metode memiliki kekurangan dan kelebihan masing-masing berdasarkan penggunaannya apda data yang ingin diuji[3].

Algoritma *Naive Bayes* adalah classifier yang merupakan sebuah metode *machine learning* yang dapat dimanfaatkan sebuah perhitungan probabilitas dan statistik yang ditemukan oleh *Thomas Bayes*, yaitu memprediksi probabilitas di masa depan berdasarkan pengamatan pada masa sebelumnya[7].

II. METODE

Data Mining

Data *mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar[6]. Berikut adalah langkah-langkah dari data *mining*[6]:

a. Data *Cleaning*
   Sebelum proses *data mining* dapat dilaksanakan perlu dilakukan proses *cleaning* (pembersihan) pada data (untuk menghilangkan *noise* data yang tidak konsisten).
b. Data *Selection*
   Pemilihan (seleksi) data dari sekumpulan data operasional, dimana data yang relevan dengan tugas analisis dikembangkan ke dalam *database*.
c. Data *Transformation*
   Data berubah atau bersatu menjadi bentuk yang tepat untuk menambang dengan ringkasan performa atau operasi agresi.
d. *Knowledge Discovery*
   Proses essensial dimana metode yang intelejen digunakan untuk mengekstrak pola data.
e. *Pattern Discovery*
   Mengidentifikasi pola yang benar-benar baik untuk mewakili pengetahuan berdasarkan data yang ada.
f. *Knowledge Preseentation*
   Merupakan gambaran teknik visualisasi dan pengetahuan digunakan untuk memberikan pengetahuan yang telah diterima oleh *user*.

Algoritma *Naïve Bayes*

Algoritma *Naive Bayes* adalah klasifikasi probabilitas sederhana yang berdasarkan pada penerapan teorama *Bayes* dengan asumsi independensi (ketidaktergantungan) yang kuat. Dengan kata lain, dalam *Naive Bayes* yang digunakan adalah model independen yang berarti bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidak adanya fitur lain dalam data yang sama berkaitan dengan ada atau tidak adanya fitur lain dalam data yang sama[4].

Persamaan dari teorama *Bayes* adalah sebagai berikut [4]:

$$P(D|E) = \frac{P(E|D)\,P(D)}{P(E)}$$

Keterangan:
E = Data dengan kelas yang belum diketahui.
D = Hipotesis data E yang merupakan suatu kelas spesifik.
P(D|E) = Probabilitas E berdasarkan hipotesis D.
P(E|D) = Probabilitas D berdasarkan kondisini E.
P(D) = Probabilitas hipotesis D (Prior probability)
P(E) = Probabilitas dari E.

Proses kerja *Naïve Bayes* adalah sebagai berikut:

a. Sebagai contoh D merupakan *data training*. Tiap data memiliki dimensi-n vektor atribut $X = (x_1, x_2, ..., x_n)$ menggambarkan ukuran pada data dari atribut n.
b. Terdapat *m* kelas, $C_1, C_2, ..., C_m$. Terdapat *X* yang akan diprediksi di mana *X* adalah milik kelas yang memiliki probailitas *pasterior* tertinggi, yang dikondisikan dengan *X*. *Naive Bayes* memprediksi *X* merupakan milik kelas C_i jika dan hanya jika

$P(C_i|X) > P(C_i|X)$ di mana $1 \leq j \leq m, j \neq i$.

Setelah itu, memaksimalkan $P(C_i|X)$. Kelas $C_i$untuk setiap $P(C_i|X)$ yang maksimal disebut hipotesa *posteriori* maksimum. Persamaan untuk mencari $P(C_i|X)$ adalah:

$$P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)}$$

c. Karena $P(X)$ bersifat konsonan untuk semua kelas, hanya $P(X|C_i)\,P(C_i)$ yang perlu dimaksimalkan. Jika kelas probabilitas *prior* tidak diketahui, secara umum dapat diasumsikan bahwa kelas-kelas tersebut mirip, seperti contoh $P(C_1) = P(C_2) = ... = P(C_m)$, dan hanya $P(X|C_i)$ yang perlu dimaksimalkan. Jika tidak, maksimalkan $P(X|C_i)\,P(C_i)$. Perlu diperhatikan bahwa kelas probabilitas *prior* mungkin diestimasi dengan $P(C_1) = |C_{1,D}|/|D|$ di mana $|C_{1,D}|$ merupakan jumlah data *training* kelas C_i dalam D.
d. Jika terdapat kumpulan data dengan banyak atribut, akan menjadi sangat berat dalam komputasi. Untuk mengurangi beban komputasi dalam mengevaluasi $P(X|C_i)$. Asumsi *class-conditional independence* dibuat di mana nilai atribut secara kondisional bebas dari yang lain. Maka

$$P(X|C_1) = \prod_{k=1}^{n} P(X_n|C_k)$$

$$= P(X_1|Ci)\ x P(X_2|Ci)\ x\ ....\ x\ P(Xn|Ci).$$

X_k adalah nilai atribut $A_k$untuk data X. Untuk tiap atribut dilihat apakah atribut tersebut bernilai kategorikal atau kontinyu. Jadi untuk menghitung $P(X|Ci)$ terdapat dua kondisi:

a. Jika $A_k$merupakan kategorikal, maka $P(X_k|Ci)$ adalah jumlah data dari C_i dalam *D* yang memiliki nilai X_k atau A_k, dibagi dengan $|C_{i,D}|$, jumlah data dari kelas $C_i$dalam *D*.
b. Jika $A_k$merupakan nilai kontinyu, maka digunakan distribusi *Gaussian* dengan μ yang merupakan rata-rata dan σ yang merupakan standar deviasi.
Persamaannya adalah:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\Pi\sigma}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Maka

$$P(X_1|C_i) = g(x_i, \mu_{Ci}, \sigma_{Ci})$$

μ_{Ci} *dan* σ_{Ci} merupakan rata-rata dan standar deviasi dari nilai atribut A_k untuk data kelas C_i.

e. Untuk memprediksi label *X* , maka $P(X|C_i)P(C_i)$ dievaluasi untuk setiap kelas C_i,*Naive Bayes* memprediksi label kelas data *X* adalah kelas $C_i$jika dan hanya jika $P(X|C_i)P(C_i) > P(X|C_j)\,P(C_j)$ di mana $1 \leq j \leq m,\ j \neq i$. Dengan kata lain, label kelas yang diprediksi adalah $C_i$untuk $P(X|C_i)P(C_i)$ adalah maksimum.

SPERMA

Sperma adalah sel reproduksi laki-laki. Pada umumnya banyak orang salah mengangap sperma sama dengan air mani tetapi secara teknis mereka berbeda satu sama lain. Air mani adalah nama seluruh cairan yang dikeluarkan dari penis saat ejakulasi, sementara sperma adalah salah satu dari banyak komponen air mani. Meskipun sperma dapat dianggap sebagai bahan paling penting dari air mani karena diperlukan untuk kehamilan, tetapi jumlahnya hanya satu persen dari hasil ejakulasi pria. Sperma dibuat di testis dalam proses yang dinamakan

spermatogenesis, kemudian disimpan jauh sampai sperma matang dan siap untuk meninggalkan tubuh. Ketika seorang laki-laki berejakulasi, sperma matang (dikombinasikan dengan sisa cairan yang membentuk air mani) dikeluarkan dari uretra, terletak di ujung penis. Cairan dalam air mani mengandung gula untuk mengisi sperma melalui saluran vagina dan bahan kimia untuk mengaktifkan sperma untuk memfasilitasi kehamilan[5].
Salah satu faktor penting dari fungsi testis dan kesuburan pria adalah untuk menentukan jumlah sperma di testis, epididimis, dan semen ejakulasi. Hitungan sperma dalam saluran reproduksi pria mencerminkan normalitas fungsi testis, karena produksi sperma adalah produk akhir dari spermatogenesis. Produksi sperma dipengaruhi oleh satus nutrisi dan hormonal, usia, obat-obatan, dan bahan kimia lingkungan. Motilitas sperma dan morfologi juga mencerminkan normalitas metabolisme sperma dan spermatogenesis. Oleh karena itu, mengurangi jumlah sperma yang motil dan meningkatnya jumlah porphology sperma yang cacat menunjukkan baik sitotoksisitas dan genotoksisitas[8].

III. HASIL

Pada tahap ini penulis telah melakukan data *selection* dan juga data *transformation* dari *dataset* yang penulis dapatkan di situs *web UCI Machine Learning Repository*, pada penelitian ini penulis menggunakan delapan atribut dan satu atribut hasil dari 100 *record* data. Berikut adalah contoh table *dataset* yang telah penulis olah :

Tabel 1. Contoh *Fertility Dataset*

Umur	Childish	Trauma	Bedah	…	Hasil
1	0	1	1	…	N
…	…	…	…	…	…
2	0	1	1	…	N
1	1	1	0	…	O
1	1	0	1	…	N
…	…	…	…	…	…
1	0	1	1	…	N

Berikut ini adalah atribut yang digunakan dalam penelitian ini:
a. Umur
b. Penyakit ketika kecil
c. Trauma
d. Bedah
e. Panas
f. Alkohol
g. Rokok
h. Duduk

Langkah berikutnya adalah menerapkan rumus algoritma *Naïve Bayes* berdasarkan pada jawaban pengguna dalam aplikasi ini yang kemudian masing-masing dari jawaban pengguna akan dihitung dengan rumus *Teorema Bayes*.

IV. Pembahasan

Padah tahap ini adalah langkah hitung manual algoritma *Naïve Bayes* di mana diambil sampel sebagai berikut :
"umur=1, penyakit ketika kecil =1, trauma=0, bedah =1, panas=3, alkohol =3, rokok =1, duduk =1"

Tabel 2. Jumlah Kejadian Atribut

Attribut	Kategori	
	Normal	Altered
Umur-1	61	8
Umur-2	27	4
Childish-0	11	2
Childish-1	77	10
Trauma-0	47	9
Trauma-1	41	3
Bedah-0	44	5
Bedah-1	44	7
Panas-1	7	2
Panas-2	55	8
Panas-3	26	2
Alkohol-1	17	4
Alkohol-2	33	6
Alkohol-3	38	2
Rokok-1	50	6
Rokok-2	20	3

Rokok-3	18	3
Duduk-0	70	11
Duduk-1	18	1

$$P(H|E) = \frac{P(E|H)\ x\ P(H)}{P(E)}$$

Dimana:

H = Kelas

E = Atribut yang digunakan

Jumlah data = 100

Jumlah kemunculan normal = 88

⇨ P(normal) = 88/100 = 0,88

Jumlah kemunculan altered = 12

⇨ P(altered) =12/100 = 0,12

Jika P(H | normal) :

P(umur = 1 | normal) 61/88 = 0.693

P(childish = 1| normal) 77/88 = 0.875

P(trauma = 0 | normal) 47/88 = 0.534

P(bedah = 1 | normal) 44/88 = 0.5

P(panas = 3 | normal) 26/88 = 0.295

P(alkohol = 3 | normal) 38/88 = 0.431

P(rokok = 1 | normal) 50/88 = 0.568

P(duduk = 1 | normal) 18/88 = 0.204

Maka

P(H | normal) = 0.693 x 0.875 x 0.534 x 0.5 x 0.295 x 0.431 x 0.568 x 0.204

= 0.002385

Jika P(H | Altered) :

P(umur = 1 | altered) 8/12 = 0.667

P(childish = 1 | altered) 10/12 = 0.833

P(trauma = 0 | altered) 9/12 = 0.75

P(bedah = 1 | normal) 7/12 = 0.583

P(panas = 3 | altered) 2/12 = 0.167

P(alkohol = 3 | altered) 2/12 = 0.167

P(rokok = 1 | altered) 6/12 = 0.5

P(duduk = 1 | altered) 1/12 = 0.083

Maka

P(H | Altered) = 0.667 x 0.833 x 0.75 x 0.583 x 0.167 x 0.167 x 0.5 x 0.083

= 0.0002826

Karena nilai dari P(H | normal) lebih tinggi dibandingkan dengan P(H | altered) maka dapat disimpulan hasil prediksi kualitas sperma untuk sampel yang sebelum telah disebutkan adalah “Normal”.

Berikut ini adalah tabel untuk menguji valid atau tidaknya prediksi dari aplikasi.

Tabel 3. Hasil Pengujian *Data Testing*

Masukan pengguna (umur,childish,trauma,bedah,panas,alkohol,rokok,duduk)	Prediksi yang diharapkan	Hasil Pengujian
2,1,0,1,2,2,3,0	Altered	Valid
1,1,0,1,2,2,1,0	Altered	Valid
1,0,0,1,2,1,2,0	Altered	Valid
2,1,0,0,3,3,3,0	Altered	Tidak Valid
1,1,0,1,2,2,2,0	Normal	Tidak Valid
2,1,1,1,2,1,2,0	Normal	Tidak Valid
2,1,0,0,2,1,2,0	Normal	Tidak Valid
1,1,0,0,2,3,1,0	Normal	Valid
1,1,1,0,2,2,3,0	Normal	Valid
1,1,1,0,3,3,1,1	Normal	Valid
1,1,0,0,3,3,3,0	Normal	Valid
1,1,0,0,3,3,1,0	Normal	Valid
2,1,1,1,3,2,2,0	Normal	Valid
2,1,0,0,2,3,3,0	Normal	Valid
1,0,0,0,3,2,3,0	Normal	Tidak Valid

Berikut adalah tabel *confusion matrix* yang dihasilkan dari tabel pengujian.

Tabel 4. *Confusion Matrix*

	True Normal	**True Altered**	**Class Precision**
Pred. Normal	7	1	87.50%
Pred. Altered	4	3	42.85%
Class recall	63.63%	75.00%	
Accuracy	66.67%		

Setelah melakukan perhitungan secara manual, penulis melakukan tahap desain aplikasi. Berikut adalah tampilan aplikasi yang telah dibuat:

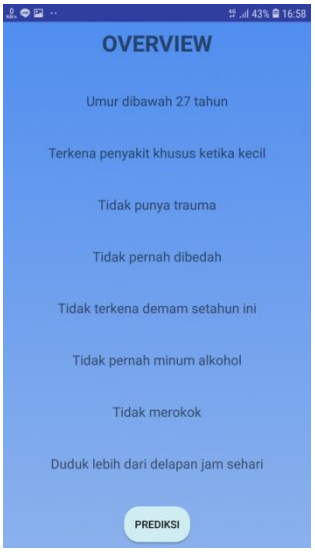
1. Menu utama adalah halaman awal dimana pengguna dapat melanjutkan ke menu prediksi atau melihat halaman tentang aplikasi. Berikut adalah tampilannya:

Gambar 1. Tampilan Menu Utama

2. Menu pertanyaan adalah menu di mana pengguna memilih jawaban dan nantinya aplikasi akan melakukan perhitungan sesuai dengan jawaban yang pilih oleh pengguna. Tampilan untuk menu pertanyaan dapat di lihat pada gambar dibawah ini :

Gambar 2. Tampilan Pertanyaan

3. Menu ringkasan jawaban
Halaman ini menampilkan jawaban yang telah pengguna pilih sebelumnya. Berikut adalah tampilan dari menu ringkasan jawaban :



Gambar 3. Tapilan Ringkasan Jawaban

4. Halaman hasil prediksi
Halaman ini menampilkan hasil prediksi yang telah aplikasi jalankan. Hasil yang muncul berupa teks dan persentase dari hasil prediksi beserta keterangan prediksi.

Gambar 4. Tampilan Hasil Prediksi

5. Halaman tentang aplikasi
Halaman ini menampilkan kapan, dimana, dan siapa yang membuat aplikasi.

Gambar 5. Tampilan Tentang

V. Simpulan

Berdasarkan data yang dijadikan data training, algoritma *Naïve Bayes* dapat memprediksi 10 dari 15 data pengujian untuk prediksi kualitas sperma sehingga menghasilkan akurasi sebesar 66.67%, dalam proses klasifikasi akan semakin akurat dengan bertambahnya jumlah data latih, serta algoritma *Naïve Bayes* dapat diterapkan pada aplikasi berbasis *Android*.

REFERENSI

[1] Badan Pusat Statistik, "Sensus Penduduk 2010," 2010. [Online]. Available: http://sp2010.bps.go.id. [Accessed 18 10 2018].

[2] N. M. Dewantari, "Peranan Gizi Dalam Kesehatan Reproduksi," *SKALA HUSADA*, pp. 219-224, 2013.

[3] I. H. Witten, E. Frank and M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques (3rd Edition), USA: Elsevier, 2011.

[4] H. Markus and K. Raff, RapidMiner : Data Mining User Cases and Business Analythics Application, London: CRP PRESS, 2013.

[5] University of California Santa Barbara, "What is sperm?," [Online]. Available: https://sexinfo.soc.ucsb.edu/what-sperm. [Accessed 10 Oktober 2018].

[6] E. Lutfi, Algoritma Data Mining, Yogyakarta: C.V Andi Offset, 2009.

[7] E. Saputra, "Keberhasilan Telemarketing Bank Untuk Mencari Algoritma Dengan Performa Terbaik," *Jurnal Ilmu Pengetahuan dan Teknologi Komputer*, 2017.

[8] T. Lobl and E. Hafez, Male Fertility and Its Regulation, Lancaster: MTP PRESS LIMITED, 2012.

BIOGRAFI

Steven Hosana lahir di Tangerang pada 22 Agustus 1997. Menyelesaikan Pendidikan di SMK Negeri Kabupaten Tangerang 1 pada tahun 2015, dan mendaftar di Universitas Buddhi Dharma pada tahun 2015 jurusan Teknik Informatika.

Dicky Surya Dwi Putra lahir di Tangerang pada 27 Juni 1987. Menyelesaikan pendidikan S1 (S.Kom.) di STMIK Buddhi dan pendidikan S2 (M.Kom.) di Eresha School of IT pada tahun 2012. Sekarang aktif sebagai dosen tetap di Universitas Buddhi Dharma sejak tahun 2009.