Jurnal Ilmiah
ILMU KOMPUTER
Universitas Udayana

Vol. X, No. 2, September 2017 **ISSN 1979 - 5661**

PREDIKSI KESUBURAN (FERTILITY) DENGAN MENGGUNAKAN PRINCIPAL COMPONENT ANALYSIS DAN KLASIFIKASI NAIVE BAYES

Gede Agus Irawan [1] , Agus Muliantara [2]

Program Studi Teknik Informatika, Jurusan Ilmu Komputer,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
E-mail: gede.irawan@cs.unud.ac.id[1]

ABSTRAK

Fertilitas pada pria menjadi sebuah masalah dalam beberapa dekade ini di bidang kesehatan. Pada penelitian di beberapa jurnal didapatkan bahwa gaya dan pola hidup sangat mempengaruhi fertilitas pada pria. Disamping itu ada juga faktor lain yang mempengaruhi tingkat fertilitas seperti penyakit bawaan, operasi, demam tinggi, kerusakan testis, dsb. Apabila testis mengalami kerusakan yang cukup parah, testis tidak dapat dipulihkan fungsinya seperti semula. Maka dari itu dibutuhkan pencegahan dini agar kerusakan testis tidak terlanjur parah dan masih dapat ditanggulangi. Pencegahan dapat dilakukan dengan melakukan prediksi kesuburan (fertility).

Dalam memprediksi digunakanlah salah metode klasifikasi yaitu *Naive Bayes Classifier*, yang didukung oleh *Principal Component Analysis* untuk melakukan reduksi fitur. Data fertilitas yang digunakan berasal dari UCI (University of California, Irvine) Penelitian ini menggunakan aplikasi untuk melakukan simulasi dengan RapidMiner dan didapatkan hasil akurasi yaitu 80% untuk prediksi fertilitas pada pria.

Kata Kunci: *Fertilitas, Naive Bayes Classifier, Prediksi, Principal Component Analysis, RapidMiner.*

ABSTRACT

Fertility in men has become a problem this decade in health department. Research in several journal found that lifestyle is greatly affect the fertility in men. Besides of that, there are also another factor that affect fertility rates such as child disease, surgery, high fever, testicular damage, etc. If testicles is severely damaged, the testicles can not be restored to its original function. Therefore it is necessary early prevention so that testicular damage is not too severe and can still be overcome. Prevention can be done by predicting fertility.

In predicting it is used one classification method that is Naive Bayes Classifier, which is supported by Principal Component Analysis to perform feature reduction. The fertility data used is from UCI (University of California, Irvine). This research use application to perform simulation with RapidMiner and got result of accuracy that is 80% for fertility prediction in man..

Keywords: *Fertility, Naive Bayes Classifier, Prediction, Principal Component Analysis, RapidMiner.*

1 PENDAHULUAN

Kesuburan (fertility) pada pria merupakan salah satu faktor penting dalam proses melanjutkan keturunan pada pasangan suami istri. Beberapa dekade belakangan dibidang kesehatan di mancanegara terjadi permasalahan yaitu tingkat kesuburan pria. Beberapa penelitian tentang tingkat kesuburan menyatakan bahwa faktor-faktor yang mempengaruhi tingkat kesuburan seperti hormon, penyakit bawaan, pernah tidaknya dioperasi[4], faktor lain seperti konsumsi minuman beralkohol, duduk terlalu lama, dan merokok[6], dsb, mengambil peran penting dalam mengukur tingkat kesuburan. Dengan berubahnya pola dan gaya hidup masyarakat yang tidak sehat, secara tidak langsung, gaya hidup seperti ini mempengaruhi kualitas sperma pada pria yang dapat mengakibatkan sulitnya pasangan suami istri mendapatkan keturunan[1][3]. Karena sekali testis mengalami kerusakan, tidak akan bisa dipulihkan fungsinya seperti semula. Di Bali sendiri pola hidup seperti yang ditunjukkan sangat sering kita temui. Untuk melakukan pencegahan awal dibuatlah suatu simulasi untuk memprediksi kesuburan pria berdasarkan faktor-faktor yang telah disebutkan sebelumnya dengan bantuan *data mining*.

Data mining adalah suatu proses yang memiliki tujuan untuk menemukan suatu pola otomatis atau semi-otomatis dari data yang sudah kita dapat/miliki di dalam basis data yang dimanfaatkan untuk penyelesaian suatu masalah. *Data mining* memiliki beberapa teknik, diantaranya klasifikasi dan *clustering*. Teknik klasifikasi adalah teknik pembelajaran yang digunakan untuk memprediksi nilai dari atribut kategori target. Metode yang paling populer digunakan untuk teknik klasifikasi adalah *Decision Trees*, *Naïve Bayes Classifiers* (NBC), *Statistical analysis*, dan lain lain.

Naïve Bayes Classifier merupakan salah satu dari metode pengklasifikasian. *NBC* dipilih karena merupakan metode klasifikasi yang simpel dan efisien. *Naïve Bayes Classifier* dapat diterapkan pada data yang lumayan besar/banyak jumlahnya[7], dan dapat menangani data yang tidak lengkap (memiliki *missing value*). Namun asumsi independen atributnya membatasi dalam penerapan pada data aktual. Dimana dibutuhkan bantuan metode lain untuk melakukan seleksi atau pengubahan fitur atribut. Maka dari itu dipilihlah Principal Component Analysis untuk membantu meningkatkan keakuratan dan kinerja dari *Naïve Bayes Classifier* itu sendiri.

2 METODOLOGI PENELITIAN

2.1 Data

Dataset yang digunakan diperoleh dari website UCI (University of California, Irvine) tentang fertilitas. Jumlah data sebanyak 100 data dengan pembagian 70 data *training* dan 30 *testing*. Data berisi 9 atribut yaitu :

1. Musim yang mana saat analisa dilakukan yang berisi 1) winter, 2) spring, 3) Summer, 4) fall. Dengan nilai (-1, -0.33, 0.33, 1)
2. Umur saat analisa dilakukan 18-36 dengan rentang nilai (0 – 1)
3. Penyakit bawaan 1)Ya, 2) Tidak (0,1)
4. Kecelakaan atau Trauma 1)Ya, 2) Tidak (0,1)
5. Pernah tidaknya mengalami operasi 1)Ya, 2) Tidak (0,1)
6. Demam tinggi selama tahun saat analisis 1) Kurang dari tiga bulan yang lalu, 2) lebih dari tiga bulan yang lalu, 3) Tidak (-1, 0, 1)
7. Frekuensi minum minuman beralkohol 1) Beberapa kali sehari, 2) Setiap hari, 3) Beberapa kali seminggu, 4) Sekali seminggu, 5) Sangat jarang atau tidak pernah sama sekali (0 – 1)
8. Kebiasaan merokok 1) Tidak pernah, 2) Sesekali, 3) Setiap hari (-1, 0, 1)
9. Lamanya duduk perhari 0 hingga 16 jam (0 – 1)

Dan kelas yang berisi kelas N (normal) atau O (*altered*) yang merupakan penentu apakah pria tersebut normal atau terjadi perubahan (*altered*) pada tingkat kesuburannya.

2.2 *Principal Component Analysis* (PCA)

Principal Component Analysis (PCA) melibatkan prosedur matematis yang mengubah fitur dalam jumlah yang besar menjadi fitur yang jumlahnya lebih sedikit yang disebut *principal component*. *Principal component* dipilih untuk menjelaskan sebanyak mungkin informasi, variansi, di dalam data set. Fitur baru akan mengeliminasi informasi berlebih dan menyaring noise dari data set asli [5]. Keuntungan lainnya dari PCA adalah saat kita

telah menemukan pola datanya dan kita meringkas datanya, misalnya dengan mengurangi jumlah dimensi dari fitur itu sendiri, tanpa kehilangan banyak informasi. *Principal Component* didapat dengan memproyeksikan vektor fitur asli pada ruang yang ditentukan oleh eigenvektor. Cara menghitung algoritma PCA dapat dinyatakan sebagai berikut :

1) Cari *mean* dari tiap fitur :

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (1)$$

Dimana $\overline{X}$ (disebut "X bar") mengindikasikan mean dari fitur set X dan n adalah jumlah data yang ada pada fitur tersebut.

2) Cari standar deviasi dengan rumus :

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})}{(n-1)}} \quad (2)$$

Dimana s merupakan standar deviasi dari sebuah sampel.

3) Hitung kovariansi dari sampel, dimana pada rumus dibawah akan digunakan 2 contoh fitur yang kita sebut fitur X dan fitur Y :

$$cov(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)} \quad (3)$$

4) Selanjutnya analisa eigen pada hasil kovariansi matriks. Dimana ini akan melibatkan penghitungan nilai eigen dan vektor eigen.
5) Nilai eigen diurutkan dan diberi label dari besar ke kecil. Nilai eigen k yang terbesar dipilih.
6) Vektor eigen yang terkait dimasukkan ke dalam matriks transformasi **W**.
7) Transormasi dari dataset dilakukan dengan memproyeksikan data asli, **X**, kedalam sub-ruang **Y** sebesar k-dimensi, dengan persamaan berikut :

$$Y = W'X \quad (4)$$

Jumlah dari *principal component* ditentukan jumlah nilai eigen yang dipilih. Misalnya jika kita hanya menginginkan satu *principal component* yang mengandung variabilitas paling banyak, kita memilih yang memiliki nilai eigen terbesar. Memilih jumlah nilai eigen yang optimal bisa dibilang merupakan suatu seni, daripada sains.

2.3 *Naive Bayes Classifier*

Dalam klasifikasi *Naïve Bayes* sendiri, sebuah keputusan klasifikasi ditentukan oleh sebuah probabilitas. Dimana probabilitas yang paling besar atau yang paling sering muncul itulah yang dipilih. Probabilitas ini, *p(x)*, ditentukan dengan menggunakan Persamaan Diskriminan Umum sebagai berikut [5]:

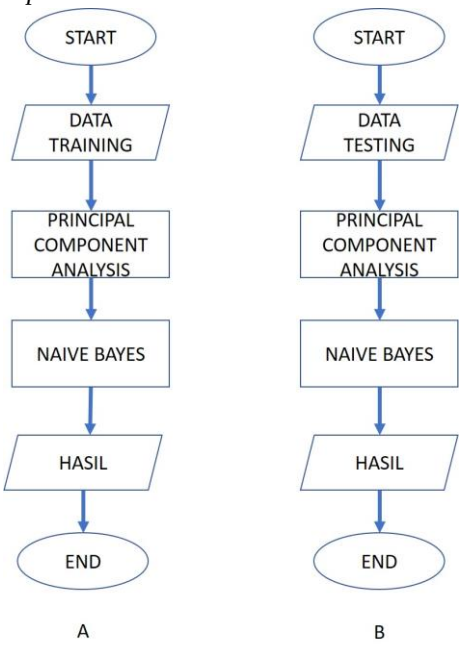
$$p(x) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right] \quad .....(6)$$

Dimana *d* adalah jumlah dari fitur, Σ adalah matriks kovariansi, dan μ adalah *mean* dari vektor.

2.4 Metode Penelitian

Dalam penelitian ini kita menggunakan sebuah aplikasi simulasi menggunakan Rapidminer versi 7.5, dimana data yang kita dapat dari UCI diinputkan terlebih dahulu ke dalam basis data yang ada pada Rapidminer. Dimana seperti yang disebutkan pada bagian A, data di pecah menjadi 2 bagian yaitu data *training* dan data *testing* dengan jumlah masing-masing 70 dan 30 data. Dimasukkan ke dalam basis data yang berbeda.

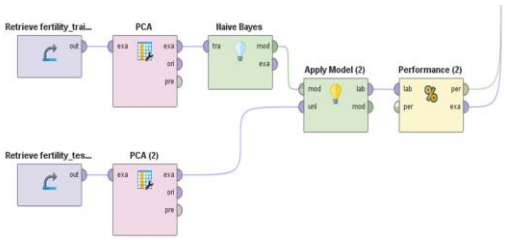
Alur data penelitian dapat dilihat pada gambar 1. Dimana kedua data terlebih dahulu melalui proses reduksi fitur dengan menggunakan *Principal Component Analysis* yang nantinya akan menghasilkan *Principal Component*. Setelah direduksi kemudian data



Gambar. 1. Flowchart penelitian penggunaan Principal Component Analysis untuk reduksi fitur dan Naive Bayes untuk klasifikasi, dimana A merupakan alur penlatihan dan B adalah alur pengujian.

masuk kedalam *Naïve Bayes Classifier* untuk mendapatkan hasil. Dimana pada bagian A, hasil yang didapat adalah probabilitas dari setiap Principal Component yang nantinya akan di bawa ke bagian B untuk dilakukan proses testing. Pada bagian B, hasil yang didapat adalah hasil klasifikasi dari data testing yang diujicobakan kedalam sistem.

Di dalam Rapidminer kita persiapkan fungsi-fungsi dan juga database yang akan digunakan dimana dalam hal ini yang kita gunakan adalah database fertility_training yang berisi data training sejumlah 70 dan fertility_testing yang berisikan data testing sebanyak 30 data. Fungsi yang digunakan adalah *Principal Component Analysis*, *Naive Bayes Classifier*, *Apply Model*



Gambar. 2. Proses penggunaan RapidMiner dengan urutan paling kiri adalah data training (atas), data testing (bawah), *Principal Component Analysis*, *Naive Bayes*, *Apply Model*, dan *Performance*

dan *Performance*. Dimana *Apply Model* bertugas sebagai penghubung antara data *testing* dengan *Naive Bayes* yang sudah melakukan pelatihan dengan data *training* yang tentunya harus melewati *Principal Component Analysis* terlebih dahulu. Fungsi Performance adalah memberikan tampilan berupa akurasi dan presisi hasil dari PCA dan klasifikasi *Naive Bayes*. Proses dapat dilihat pada gambar 2.

3 HASIL DAN PEMBAHASAN

Setelah menggunakan simulasi dengan RapidMiner, kita mendapatkan hasil reduksi seperti yang diperlihatkan oleh Tabel I. Dimana dari 9 fitur atribut yang ada, PCA mereduksi fitur tersebut menjadi 7 *Principal Component*. Dipilihnya 7 *Principal Component* dikarenakan dari perhitungan yang dilakukan menggunakan RapidMiner, setelah melihat ROC (*Reciever Operating Characteristic*) dan menghitung AUC (*Area Under Curve*) dari penggunaan 1 *Principal Component* hingga 9 *Principal Component*, didapatkan bahwa menggunakan 7 *Principal Component* memiliki nilai AUC 0,712, dimana angka ini termasuk *Fair Classification*[2]. Hasil perbandingan dapat

TABEL I
HASIL PRINCIPAL COMPONENT PADA 10 DATA TESTING

PC1	PC2	PC3	PC4	PC5	PC6	PC7
-0,269	0,267	-0,898	-0,142	-0,607	0,269	0,445
-0,299	1,484	-0,309	-0,376	0,438	-0,172	0,015
-0,235	-0,505	0,741	-0,29	-0,178	0,022	0,079
-0,366	-0,666	-0,801	-0,09	-0,561	0,079	-0,114
-0,442	-0,654	0,279	0,264	0,075	0,616	0,058
-0,317	0,49	-0,179	-0,353	0,407	-0,358	0,187
-0,081	-0,69	0,25	-1,361	-0,336	0,469	-0,133
-0,592	-0,611	-0,543	0,246	0,315	0,064	-0,046
0,837	-0,579	0,282	1,229	0,048	0,209	-0,165
1,075	-0,599	-0,538	-0,431	-0,134	-0,461	-0,367

TABEL II
PERBANDINGAN AUC DAN AKURASI SETIAP PRINCIPAL COMPONENT

PC	Nilai AUC	Akurasi
1	0,452	86,67%
2	0.529	86,67%
3	0,500	86,67%
4	0,490	86,67%
5	0,519	86,67%
6	0,558	80%
7	0,712	80%
8	0,712	80%
9	0,712	80%

dilihat pada tabel II. Nantinya *Principal Component* inilah yang akan masuk ke proses berikutnya di klasifikasi *Naive Bayes*.

Setelah diklasifikasi oleh *Naive Bayes*, didapatkan akurasi dari percobaan ini. Akurasi dihitung dengan menggunakan *Confussion matrix*. Dimana hal ini sudah dihitung di performance dan hasilnya dapat dilihat pada

TABEL III
HASIL PERHITUNGAN AKURASI

	Result N	Result O	Class Precision
Pred. N	23	3	88,46%
Pred. O	3	1	25%
Class recall	88,46%	25%	

Tabel III. Untuk menghitung akurasi dari simulasi ini, digunakanlah persamaan seabagai berikut :

TABEL IV
PERBANDINGAN METODE NAIVE BAYES DAN NAIVE BAYES DENGAN PRINCIPAL COMPONENT ANALYSIS

Metode	Akurasi	Nilai AUC dari ROC
Naive Bayes	80%	0,462
Naive Bayes + Principal Component Analysis	80%	0,712

$$Accuracy = \frac{TrueN + TrueO}{total\_data} * 100\%$$

Dimana TrueN adalah jumlah prediksi N (normal) yang benar-benar normal, TrueO adalah jumlah prediksi O (*altered*) yang memang benar *altered*, total_data adalah jumlah total data *testing* yang digunakan. Dari 30 data *testing* yang diuji pada simulasi yang telah dibuat, didapatkan hasil akurasi 80%.

Jika kita lihat pada Tabel IV, *Naive Bayes* saja dengan menggunakan data *training* dan *testing* yang sama, mendapatkan akurasi yang sama dengan *Naive Bayes* yang telah dikombinasikan dengan PCA yaitu 80%. Akan tetapi dari nilai AUC kita bisa melihat perbedaan dimana AUC dari Naive Bayes saja mempunyai nilai 0,462 yang dimana menurut Gorunescu, F, bahwa nilai 0,462 ini termasuk *Failure*[2]. Sedangkan saat Naive Bayes digabungkan dengan PCA nilai AUC dari ROC nya adalah 0,712 dimana ini termasuk *fair classification*.

4 KESIMPULAN

Dari hasil yang dicapai dapat disimpulkan bahwa *Principal Component Analysis* dan *Naïve Bayes* dapat melakukan reduksi fitur dan melakukan klasifikasi dengan cukup baik. Terbukti dengan tercapainya akurasi 80% pada simulasi menggunakan RapidMiner.

Pengembangan kedepannya akan terfokus pada penerapan metode reduksi yang lain untuk mendapatkan akurasi yang lebih baik. Dan dapat juga mulai menggunakan data asli dari masyarakat di Bali untuk pembuatan program yang kedepannya dapat membantu memprediksi fertilitas secara langsung.

5 DAFTAR PUSTAKA

[1] Daftar Carlsen, E., Giwercman, A., Keiding, N., & Skakkebaek, N. E. (1992). Evidence for decreasing quality of semen during past 50 years. BMJ, 305(6854), 609–613.

[2] Gorunescu, F. (2011). Data Mining Concepts, Model and Technique. Berlin : Springer

[3] Irvine DS.(1998). Epidemiology and aetiology of male infertility. Hum. Reprod.;Vol 13(1):33-44.

[4] Irvine, D. S. (2000). Male reproductive health: Cause for concern? Andrologia, 32(4–5), 195–208.

[5] Gupta, Gopal Khrisna. (2004). Principal Component Analysis and Bayesian Classifier Based Character Recognition

[6] Martini, A. C., Molina, R. I., Estofan, D., Senestrari, D., Fiol de Cuneo, M., & Ruiz, R. D. (2004). Effects of alcohol and cigarette consumption on human seminal quality. Fertility and Sterility, 82(2), 374–377.

[7] Wu, Xindong and Kumar, Vipin. (2009). The Top Ten Algorithms in Data Mining. Boca Raton: CRC Press.