ANALISIS PENGARUH SELEKSI FITUR PADA KLASIFIKASI KONSENTRASI SPERMA BERDASARKAN FAKTOR FAKTOR LINGKUNGAN, KESEHATAN, DAN GAYA HIDUP

Nasrokhah Noviati[*], Silmi Fauziati[**], Indriana Hidayah[***]
Jurusan Teknik Elektro dan Teknologi Informasi FT UGM, Jalan Grafika No 2,
Kampus UGM Yogyakarta 55281
Email: nasrokhah_s2te12@mail.ugm.ac.id[*], silmi.@ugm.ac.id[**], indriana.h@ugm.ac.id[***]

Abstrak

Menurunnya fertilitas terjadi di banyak negara. Berbagai penyebab yang melatarbelakangi hal tersebut. Beberapa di antaranya adalah disebabkan gaya hidup yang buruk, latar belakang kesehatan yang tidak baik, dan juga lingkungan yang tidak sehat. Menggunakan metode data mining, dapat mengklasifikasikan konsentrasi sperma apakah normal atau tidak. Fitur yang banyak dalam dataset akan menimbulkan banyak permasalahan, sehingga perlu melakukan seleksi fitur. Keuntungan menggunakan seleksi fitur antara lain dapat mengingkatkan akurasi suatu klasifikasi, dan membantu mengurangi fitur-fitur yang tidak relevan. Algoritme seleksi fitur yang digunakan dalam penelitian ini Principal Component Alanalisys (PCA) yang diterapkan pada metode klasifikasi Multilayer Perceptron (MLP), Decision Tree, dan Support Vector Machines (SVM). Dataset yang digunakan diambil dari dataset fertility pada UCI Maching Learning Repository untuk mengklasifikasikan konsentrasi sperma. Kesimpulan pada penelitian ini adalah menggunakan seleksi fitur PCA mampu mengurangi fitur yang kurang relevan dari 9 fitur menjadi 5 fitur terbaik yaitu musim, penyakit, kecelakaan, demam, dan rokok. Serta 6 fitur terbaik yaitu musim, penyakit, kecelakaan, demam, rokok, dan operasi. Penggunaan 5 atau 6 fitur terpilih terbukti mampu meningkatkan akurasi dari hasil klasifikasi tanpa seleksi fitur.

***Kata kunci**: data mining, fertilitas, seleksi fitur, PCA.*

1. PENDAHULUAN

Tingkat kesuburan (fertilitas) di berbagai negara mengalami penurunan. Seperti pada negara-negara Eropa fertilitas telah menurun secara dramatis selama 50 tahun terakhir (Bloom & Sousa-poza 2010). Berbagai penelitian dilakukan untuk mengetahui apa penyebab penurunan tersebut, salah satunya adalah faktor lingkungan. Jurewichz, dkk. meneliti tentang studi epidemiologis menunjukkan kesadaran factor lingkungan yang dapat mempengaruhi kualitas sperma. Penelitian tersebut menyebutkan bahwa sejumlah pestisida mampu mempengaruhi jumlah sperma dan motilitasnya (pergerakan), polusi udara mempengaruhi karakteristik sperma, serta ponsel mungkin mempengaruhi kualitas semen dengan mengurangi sebagian besar motilitas tetapi juga jumlah sperma, viabilitas (vitalitas), dan morfologi (bentuk) sperma (Jurewicz et al. 2009).

Masalah infertilitas (ketidak suburan) seringnya ditujukan pada pihak wanita, padahal pria juga memiliki peluang infertilitas sebesar 30-40%. Berbagai macam hal yang dapat mempengaruhi fertilitas pada pria di antaranya adalah gaya hidup yang tidak sehat seperti merokok, menkonsumsi alkohol, berat badan dan olahraga yang berlebihan, suplementasi vitamin serta obat-obatan yang belebihan, serta stres (HIFERI et al. 2013).

Beberapa penelitian yang memprediksi fertilitas pria berdasarkan gaya hidup, status kesehatan, dan faktor lingkungan telah dilakukan dengan metode kecerdasan buatan. Gill, dkk. melakukan penelitian pada pria usia 18-37 tahun yang berjumlah 100 orang dan 34 fitur yang menjadi penilaian dengan standar penilaian WHO 2010. Hasil dari penelitian menghasilkan 88 '*normal*', dan 12 '*altered*' untuk memprediksi konsentrasi sperma, dan motilitas sperma. Gill, dkk. melakukan perbandingan tiga metode yaitu *Decision Tree*, *Multilayer Perceptron* (MLP) dan *Support Vector Machines* (SVM). Penilaian akurasi tertinggi dihasilkan oleh sebesar 86% untuk konsentrasi sperma dan 73%-76% untuk motilitasnya. Sedangkan *Decision Tree* menghasilkan akurasi sebesar 84% untuk konsentrasi sperma dan 70% untuk motilitas (Gil et al. 2012).

Girella kemudian bersama Gil, dkk. melanjutkan penelitian tersebut dengan algoritme MLP dengan 123 sukarelawan yang mendonorkan sampel spermanya yang dianalisis menggunakan standar WHO 2010 dengan usia 18-36 tahun. Namun hanya untuk konsentrasi sperma dan motilitas

ISBN 978-602-99334-4-4

sperma. Hasil akurasi pada konsentrasi sperma sebesar 90% dan motilitas sebesar 82% (Girela et al. 2013).

Kemudian Wang, dkk. dengan pendekatan *Ensemble Learning*, yaitu *Clustering-Based Decision Forest* (CBDF) untuk mengatasi masalah kelas yang tidak seimbang dalam prediksi kualitas sperma dari data *fertility* pada UCI *Machine Learning Repository*. Metode tersebut memiliki hasil yang lebih baik pada data set kesuburan pada sperma dibanding *Decision Tree*, *Support Vector Machines* (SVM), *Random Forest*, *Multilayer Perceptron* (MLP) *Neural Network* dan *Logistic Regression*. CBDF juga dapat digunakan untuk mengevaluasi variabel yang penting dan lima faktor penting yang dapat mempengaruhi konsentrasi sperma yang diperoleh dalam penelitian ini, yaitu umur, trauma serius, waktu duduk, musim ketika sampel sperma diproduksi, dan demam tinggi pada tahun sebelumnya (Wang et al. 2014).

2. METODOLOGI

2.1 DATA DAN ALAT

Data set yang digunakan dalam penelitian ini diambil dari UCI *Machine Learning Repository* yang terdiri dari 100 *record* data dan 9 fitur. Data diperoleh dengan pembagian kuesioner pada relawan muda yang sehat, dan juga menggunakan hasil analisis sperma untuk menilai akurasinya. *Software* yang digunakan untuk mengklasifikasikan dan melakukan seleksi fitur adalah *Rapidminer*. Fitur dalam penelitian ini mencakup faktor lingkungan, kesehatan, dan gaya hidup yang digunakan untuk mengklasifikasikan konsentrasi sperma. Fitur-fitur tersebut yaitu musim dilakukannya analisis, usia, penyakit saat kecil, kecelakaan atau trauma yang serius, operasi, demam tinggi pada tahun sebelumnya, frekuensi konsumsi alkohol, kebiasaan merokok, jumlah jam duduk per hari, serta kelas yang dibagi 2 yaitu N: normal, dan O: altered (Gil 2013).

2.2 DATA MINING

Data mining adalah metode pengambilan pola dari data-data yang tidak bermakna menjadi informasi yang bernilai di dalam *database*. Data mining mencakup penggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstrak dan mengidentifikasi informasi yang bermanfaat serta pengetahuan yang terjalin dari berbagai *database* besar (Kusrini & Luthfi 2009; E., Turban 2005)(E., Turban 2005). Pengelompokan data mining berdasarkan tugasnya dibagi menjadi enam yaitu deskrispi, estimasi, prediksi, klasifikasi, pengklusteran, dan asosiasi.

2.2.1 PCA

Penelitian ini akan melakukan klasifikasi dengan algoritme SVM, MLP, dan *Decision Tree* serta melakukan seleksi fitur dengan menggunakan *Principal Component Analisys* (PCA). Penggunaan seleksi fitur sangat berguna untuk mencegah terjadinya efek dimensionalitas, mengurangi jumlah waktu dan memori yang dibutuhkan oleh algoritme yang digunakan, memudahkan dalam memvisualisasikan data, dan membantu mengurangi fitur-fitur yang tidak relevan (derau) tanpa kehilangan informasi dari data aslinya (Hermawati 2013). PCA mentransformasikan sejumlah besar variabel yang berkorelasi, menjadi beberapa variabel yang tidak berkorelasi tanpa menghilangkan informasi penting di dalamnya (Jatra et al. 2011). Cara untuk mendapatkan *principal component* adalah dengan mencari nilai *eigenvalue* dan *eigenvector* dari matriks *covariance*-nya (Santosa 2007).

Gambar 1. menjelaskan alur dari proses PCA untuk mengurangi dimensi data dengan dataset hipotetis variabel m (Deshpande 2011). Kovarian digunakan untuk mengukur bagaimana sebaran data dengan dimensi yang bervariasi terhadap nilai tengahnya yang berkaitan dengan hubungan antar data (Hermawati 2013). Nilai covarian didapatkan menggunakan persamaan (1).

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

Berlaku persamaan cov(X,Y) = cov(Y,X).

$\bar{X}$: nilai rata-rata dari data X; $\bar{Y}$: nilai rata-rata dari data Y; n : jumlah data dalam data set.

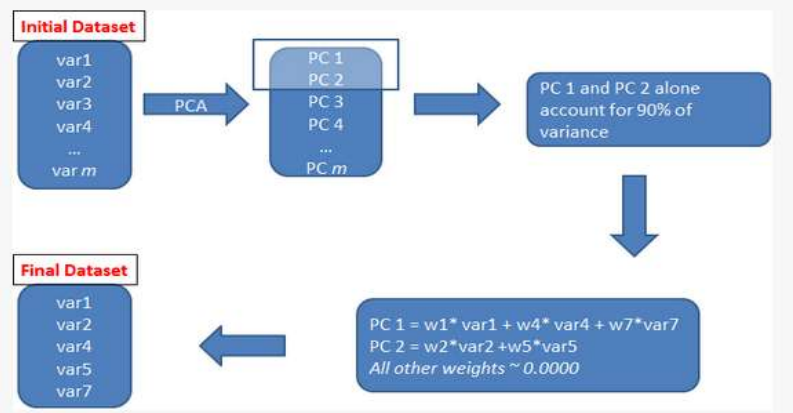
Nilai *eigenvalue* dan *eigenvector* dapat dicari menggunakan persamaan (2) dan (3) (Prasetyo 2012; Santosa 2007; Astuti 2014).

$$(A-\lambda I)= 0 \quad (2)$$

$$(A- \lambda I)x= 0 \quad (3)$$

λ: bilangan riil; *x*: matriks yang nilainya tidak nol; A:matriks utama.

P*rinciple component* ditentukan dengan mengambil *eigenvalue* dan *eigenvector* yang terbesar. Kemudian kalikan antara variabel asli dengan matriks *eigenvector*-nya (Prasetyo 2012; Jatra et al. 2011).



Gambar 1. Alur proses PCA.

2.2.2 DECISION TREE

Dalam data mining, *decision tree* termasuk algoritme yang kuat dan banyak digunakan untuk melakukan klasifikasi dan regresi. *Decision tree* mengubah fakta yang besar menjadi pohon keputusan yang mempresentasikan aturan sehingga dapat mudah dipahami. Selain itu *Decision tree* dapat digunakan untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel masukan dengan sebuah variabel target. Metode ini sangat bagus digunakan untuk langkah awal proses pemodelan,karena mampu memadukan antara eksplorasi data dan pemodelan (Kusrini & Luthfi 2009).

Nilai gain dapat dilakukan dengan menggunakan persamaan (4), sedangkan untuk mencari nilai *entropy* menggunakan persamaan (5).

$$Gain\ (S,A) = Entropy\ (S) \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (Si) \quad (4)$$

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (5)$$

S: himpunan kasus, sedangkan A adalah atribut; n : jumlah kasus pada partisi ke-i;
|Si| : jumlah kasus pada partisi ke-i; |S| : jumlah kasus dalam S;
pi : proporsi dari S_i terhadap S.

2.2.3 MULTILAYER PERCEPTRON (MLP) BACKPROPAGATION

Multilayer Perceptron (MLP) adalah salah satu algoritma jaringan saraf tiruan yang mengadopsi cara kerja jaringan saraf pada mahluk hidup (Mentari et al. 2014). *Backpropagation* adalah bagian dari MLP yang paling umum digunakan. Metode ini bekerja melalui proses secara iteratif dengan menggunakan sekumpulan *data training*, membandingkan nilai prediksi dari jaringan dengan setiap *data training*. Di dalam setiap prosesnya, bobot relasi dalam jaringan dimodifikasi untuk meminimalkan nilai Mean Squared Error (MSE) antara nilai prediksi dari

ISBN 978-602-99334-4-4

jaringan dengan nilai sesungguhnya. Modifikasi relasi jaringan saraf tersebut dilakukan dengan cara mundur dari *output layer* hingga layer pertama dari *hidden layer*, oleh karena itu metode ini disebut sebagai *backpropagation* (Kusrini & Luthfi 2009).

Menghitung *error* dalam *output layer* dapat dilihat pada persamaan (6), dan *error* pada *hidden layer* dapat dilihat pada persamaan (7). Semua nilai *error* pada setiap *node* sudah dihitung. Selanjutnya adalah melakukan modifikasi terhadap bobot jaringan sebagaimana terdapat pada persamaan (8) (Kusrini & Luthfi 2009),

$$Err_i = O_i(1 - O_i)(T_i - O_i) \quad (6)$$

$$Err_i = O_i(1 - O_i)\sum_j Err_j w_{ij} \quad (7)$$

$$w_{ij} = w_{ij} + l.Err_j.O_i \quad (8)$$

Err_i : nilai error dalam *node* (neuron) *i;*

O_i : keluaran dari *output node* unit *i;*

T_i : nilai sebenarnya dari *output node* dalam contoh kasus data (*training data);*

w_{ij} : bobot antara kedua *node;*

l: *learning rate* dengan nilai antara 0 hingga 1. Jika nilai *l* kecil, maka perubahan bobot akan menjadi sedikit dalam setiap iterasi, begitu juga jika terjadi sebaliknya. Nilai *learning rate* biasanya akan berkurang selama proses pembelajaran.

2.2.4 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) adalah metode yang dapat digunakan untuk klasifikasi maupun regresi. SVM sangat baik bekerja pada data dengan dimensi yang tinggi (Prasetyo 2012). Namun waktu pelatihan SVM cenderung lambat, meskipun begitu SVM sangat akurat untuk menangani model-model nonlinear yang kompleks. Kelemahan SVM adalah rentan terhadap *overfitting* jika diandingkan dengan metode lain (Widodo et al. 2013).

SVM untuk kasus klasifikasi linear terdapat pada persamaan (9) dan (10) (Santosa 2007).

$$min\ \frac{1}{2}\ ||w||^2 \quad (9)$$

Subject to $y_i(wx_i + b) \geq 1, i - 1, \ldots, \lambda$ (10)

x_i: data *input* ;

y_i: *output* data x_i, *w,* dan *b*

x_i, *w,* dan *b*: parameter-parameter yang nilainya dicari.

2.2.5 AKURASI

Akurasi adalah salah satu parameter untuk mengetahui mengetahui performa dari proses data mining. Alat evaluasi tersebut digunakan untuk memilih metode mana yang mempunyai tingkat akurasi yang paling baik, dengan melakukan estimasi berdasarkan hasil rata-rata akurasi dari cross validation. Untuk menentukan akurasi dapat dilihat pada persamaan (11) (Firqiani et al. 2003).

$$Akurasi = \frac{\sum data\ uji\ benar\ diklasifikasi}{\sum data\ uji} \quad (11)$$

3. HASIL DAN PEMBAHASAN

Hal yang pertama dilakukan adalah mengidentifikasi data. Karena d*ataset fertility* yang berasal dari UCI *machine learning repository* sudah dinormalisasi dan bersih dari *missing value,* maka dilanjutkan pada proses klasifikasi data dengan semua fitur (9 fitur) menggunakan Decision Tree, SVM, dan MLP. Hasil klasifikasi dengan semua fitur ini menghasilkan akurasi Decision Tree 86.67%, SVM 96.67%, dan MLP 93.33%.

Tabel 1 menunjukkan nilai *eigenvalue* dari Rapidminer. Langkah selanjutnya adalah

menentukan *Principal Component* (PC) yang akan digunakan sesuai kebutuhan. Apabila pemilihan *cummulative variance* 50% masih belum memberikan akumulasi yang signifikan, maka dapat memilih *cummulative variance* yang lebih dari 50% hingga hasilnya signifikan (F.H. 2010; Long 2010). PC2 mampu menampilkan 57% dari variasi yang ada, sedangkan PC3 73%, PC4 83%, dan PC5 92%.

Tabel 2 menunjukkan nilai dari *eigenvector*. Kemudian mencari nilai terbesar dapat urutan yang memiliki nilai *eigenvector* terbesar hingga yang terkecil dari *principal component* yasng sudah dipilih sesuai dengan nilai *cummulative variance* yang ada.

Tabel 1. Nilai eigenvalue

Component	Srandart Deviation	Propontion of Variance	Cumulative Variance
PC 1	0.831	0.298	0.298
PC 2	0.803	0.278	0.576
PC 3	0.604	0.157	0.734
PC 4	0.484	0.101	0.835
PC 5	0.463	0.093	0.927
PC 6	0.321	0.044	0.972
PC 7	0.181	0.014	0.986
PC 8	0.154	0.010	0.996
PC 9	0.096	0.004	1.000

Tabel 2.Eigenvector.

No.	Attribute	PC1	PC2	PC3	PC4	PC5
1.	Musim	0.749	0.580	-0.234	0.184	-0.091
2.	Usia	-0.004	0.013	0.015	0.072	-0.088
3.	Penyakit	-0.101	-0.018	-0.018	0.160	0.177
4.	Kecelakaan	-0.117	0.045	0.336	0.896	0.159
5.	Operasi	0.057	-0.006	0.536	-0.023	-0.814
6.	Demam	-0.219	-0.204	-0.736	0.338	-0.493
7.	Alkohol	0.019	-0.039	-0.024	-0.081	0.029
8.	Rokok	-0.603	0.786	-0.035	-0.102	-0.066
9.	Duduk	0.017	-0.018	0.012	-0.027	0.131

Tabel 3 menunjukkan 5 fitur terpilih dihasilkan dari 53% *cummulative variance* dan 6 fitur dari 73% *cummulative variance.* Proses terakhir adalah membandingkan hasil akurasi dari klasifikasi non-seleksi fitur dan dibandingkan dengan klasifikasi yang menggunakan seleksi fitur PCA. Hasil perbandingan ditampilkan pada Tabel 4. Semua akurasi meningkat, kecuali SVM yang hasil akurasinya tetap tinggi dan tidak terpengaruh seleksi fitur.

Tabel 3. Hasil seleksi fitu dengan PCA.

Cummulative Variance	Principal Component	Fitur Terpilih PC>=0,100 (+/-)	Jumlah Fitur
53%	PC1, PC2	Musim, penyakit, kecelakaan, demam, rokok.	5
73%	PC1, PC2, PC3.	Musim, penyakit, kecelakaan, operasi, demam, rokok.	6

Tabel 4. Hasil perbandingan klasifikasi.

Klasifikasi	Non-Seleksi Fitur	PCA 5 Fitur	PCA 6 Fitur
Decision Tree	86,67%	96,67%	96,67%
MLP	93,33%	96,67%	96,67%
SVM	96,67%	96,67%	96,67%

4. KESIMPULAN

Kesimpulan pada penelitian ini adalah menggunakan seleksi fitur PCA mampu mengurangi fitur yang kurang relevan dari 9 fitur menjadi 5 fitur terbaik yaitu musim, penyakit, kecelakaan, demam, dan rokok. Serta 6 fitur terbaik yaitu musim, penyakit, kecelakaan, demam, rokok, dan operasi. Penggunaan 5 atau 6 fitur terpilih terbukti mampu meningkatkan akurasi dari hasil klasifikasi tanpa seleksi fitur. Decision tree dari 86,67% meningkat menjadi 96,67% baik pada 5 maupun 6 fitur. MLP dari 93,33% meningkat juga menjadi 96,67% pada 5 maupun 6 fitur. Sedangkan SVM tetap mendapatkan akurasi tertinggi dan tidak terpengaruh oleh seleksi fitur.

Untuk penelitian selanjutnya dapat dikembangkan dengan menambah fitur lain selain untuk memprediksi konsentrasi sperma. Penggunaan metode seleksi fitur yang lain juga perlu dilakukan untuk mengetahui apakah sesuai untuk *dataset fertility* sehingga mampu meningkatkan akurasinya.

ISBN 978-602-99334-4-4

DAFTAR PUSTAKA

Astuti, T., 2014. *Estimasi Missing Value Dataset Hepatitis Berdasarkan Kombinasi Seleksi Fitur dan Algoritme Machine Learning*. Gadjah Mada.

Bloom, D.E. & Sousa-poza, A., 2010. Economic Consequences od Low Fertility in Europe. *FZID Discussion Papers CC, Universität Hohenheim*.

Deshpande, B., 2011. How to run Principal Component Analysis with RapidMiner - Part 1. Available at: http://www.simafore.com/blog/bid/62910/How-to-run-Principal-Component-Analysis-with-RapidMiner-Part-1 [Accessed April 27, 2015].

E., Turban, D., 2005. *Decision Support Systems and Intelligent Systems*, Yogyakarta: Andi Offset.

F.H., I.N., 2010. Analisis Component Utama (Principal Component Analysis). Available at: http://www.slideshare.net/Nurinhapsari/penjelasan-lengkap-analisis-komponen-utama-pca.

Firqiani, H.N., Kustiyo, A. & Giri, E.P., 2003. Seleksi Fitur Menggunakan Fast Correlation Based Filter pada Algoritma Voting Feature Intervals 5. *Portal Garuda*, pp.1–12. Available at: http://download.portalgaruda.org/article.php?article=85697&val=235.

Gil, D. et al., 2012. Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, 39(16), pp.12564–12573. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0957417412007269 [Accessed September 10, 2014].

Gil, D., 2013. UCI Machine Learning Repository Fertility Data Set. Available at: Https://archive.ics.uci.edu/ml/datasets/Fertility [Accessed October 11, 2014].

Girela, J.L. et al., 2013. Semen parameters can be predicted from environmental factors and lifestyle using artificial intelligence methods. *Biology of reproduction*, 88(4), p.99. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23446456 [Accessed August 27, 2014].

Hermawati, F.A., 2013. *Data Mining*, Yogyakarta: Penerbit Andi.

HIFERI et al., 2013. *Konsensus Penanganan Infertilitas*, HIFERI, Himpunan Endokrinologi Reproduksi dan Fertilitas Indonesia PERFITRI, Perhimpunan Fertilisasi In Vitro Indonesia. IAUI, Ikatan Ahli Urologi Indonesia POGI, Perkumpulan Obstetri dan Ginekologi Indonesia.

Jatra, M., Isnanto, R. & Santoso, I., 2011. Identifikasi iris mata menggunakan metode analisis komponen utama dan perhitungan jarak euclidean. , pp.1–9. Available at: http://www.crc.uri.edu/download/JournalPLVol4No3-O.pdf#page=4.

Jurewicz, J. et al., 2009. Environmental factors and semen quality. *International journal of occupational medicine and environmental health*, 22(4), pp.305–29. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20053623 [Accessed May 29, 2015].

Kusrini & Luthfi, E.T., 2009. *Algoritma Data Mining*, Yogyakarta: Penerbit Andi.

Long, J., 2010. Principal Component Analysis (PCA) vs Ordinary Least Squares (OLS): A Visual Explanation. Available at: http://www.cerebralmastication.com/2010/09/principal-component-analysis-pca-vs-ordinary-least-squares-ols-a-visual-explination/.

Mentari, M., Sari, E.K.R. & Mutrofin, S., 2014. Klasifikasi Menggunakan Kombinasi Multilayer Perceptron dan Aligment Particle Swarm Optimization. *SENASTIK*, 2014(September), pp.10–11.

Prasetyo, E., 2012. *Data Mining Konsep dan Aplikasi Menggunakan Matlab*, Penerbit Andi.

Santosa, B., 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Yogyakarta: Graha Ilmu.

Wang, H., Xu, Q. & Zhou, L., 2014. Seminal Quality Prediction Using Clustering-Based Decision Forests. *Algorithms*, 7(3), pp.405–417. Available at: http://www.mdpi.com/1999-4893/7/3/405/ [Accessed September 24, 2014].

Widodo, P.P., Handayanto, R.T. & Herlawati, 2013. *Penerapan Data Mining dengan Matlab*, Bandung: Rekayasa Sains.