p-ISSN 2087-1716
e-ISSN 2548-7779

ILKOM Jurnal Ilmiah Volume 11 Nomor 1 April 2019
Terakreditasi peringkat 3 SK. No. 28/E/KPT/2019

KOMPARASI *NAÏVE BAYES*, *SUPPORT VECTOR MACHINE DAN K-NEAREST NEIGHBOR* UNTUK MENGETAHUI AKURASI TERTINGGI PADA PREDIKSI KELANCARAN PEMBAYARAN TV KABEL

Mohamad Efendi Lasulika
fendidsn.ui@gmail.com
Universitas Ichsan Gorontalo

Abstrak

Salah satu kendala dari macetnya pembayaran adalah kurangnya analisis dalam proses penerimaan pelanggan baru yang hanya ditinjau dari formulir yang diberikan pada saat pendaftaran, adapun tujuan penelitian ini untuk mengetahui hasil akurasi tertinggi dari komparasi Algoritma Naïve Bayes, SVM dan K-NN. Dapat diketahui bahwa algoritma *Naïve Bayes* yang memiliki nilai *accuracy* yang paling tinggi yaitu 96 %, sementara algoritma K-*Neural Network* didapatkan *accuracy* tertinggi yaitu pada Nilai K = 3 yaitu 92% , sementara pada *Support Vector Machin*e hanya mendapatkan *accuracy* sebesar 66 %. Hasil ROC *Curve* menunjukkan bahwa *Naïve Bayes* mencapai nilai AUC yang terbaik yaitu 0.99. Komparasi antara algoritma klasfikasi data mining yaitu *Naïve Bayes*, K-*Neural Network* dan *Support Vector Machine* untuk prediksi kelancaran pembayaran dengan menggunakan tipe data multivariat, metode *Naïve Bayes* merupakan algoritma yang akurat dan metode ini juga sangat dominan terhadap metode lain. Berdasarkan *Accuracy*, AUC dan T-tes metode ini masuk dalam kategori klasifikasi terbaik.
Kata Kunci: Komparasi Data Mining, Naïve Bayes, K-Nueral Network dan Support Vector Machine.

Abstract

One obstacle of the default payment is the lack of analysis in the new customer acceptance process which is only reviewed from the form provided at registration, as for the purpose of this study to find out the highest accuracy results from the comparison of Naïve Bayes, SVM and K-NN Algorithms. It can be seen that the Naïve Bayes algorithm which has the highest accuracy value is 96%, while the K-Neural Network algorithm has the highest accuracy at K = 3 which is 92%, while Support Vector Machine only gets accuracy of 66%. The ROC Curve results show that Naïve Bayes achieved the best AUC value of 0.99. Comparison between data mining classification algorithms namely Naïve Bayes, K-Neural Network and Support Vector Machine for predicting smooth payment using multivariate data types, Naïve Bayes method is an accurate algorithm and this method is also very dominant towards other methods. Based on Accuracy, AUC and T-tests this method falls into the best classification category.
Keywords: Comparison of Data Mining, Naïve Bayes, K-Neural Network and Support Vector Machine.

1. Pendahuluan

Pelanggan TV kabel pada PT.Mimoza setiap tahunnya mengalami kenaikan yang signifikan, berdasarkan data jumlah pelangan yang hingga saat ini berkisar 6000 pelanggan dari jumlah pelanggan yang terdaftar terdapat 4000 pelanggan yang aktif dan sisanya pelanggan yang pembayaranya tidak lancar. Salah satu kendala atau penyebabnya adalah kurangnya analisis dalam proses penerimaan pelanggan baru yang hanya ditinjau dari formulir yang diberikan pada saat pendaftaran. Oleh karena itu ,prediksi kelancaran pembayaran TV. Kabel perlu dilakukan dengan akurat. Dataset pelanggan TV kabel yang digunakan dalam penelitian ini merupakan data pelanggan yang di dapatkan langsung dari PT. Mimoza multimedia. Attribut pada dataset terdiri dari : Jenis Pemasangan, Status Rumah, Pekerjaan, Pembayaran, Jenis Pelanggan, Tanggungan, Status Perkawinan, Jenis Penghasilan, dan Penghasilan.

Fokus pada penelitian ini yaitu menguji beberapa algoritma apakah tingkat akurasinya lebih baik atau tidak dalam memprediksi kelancaran pembayaran pada pelanggan TV. Kabel. Ada beberapa metode komputasi yang dapat digunakan untuk menyelesaikan masalah prediksi ataupun klasifikasi namun pada penelitian ini peneliti hanya fokus pada beberapa Algoritma yang akan dibandingkan yaitu *Naïve Bayes*, *Support Vector Machine* dan K-NN untuk mengetahui mana yang lebih akurat dari ketiga metode tersebut apabila tipe datanya multivariat.

DOI: http://dx.doi.org/10.33096/ilkom.v11i1.408.11-16

p-ISSN 2087-1716
e-ISSN 2548-7779

ILKOM Jurnal Ilmiah Volume 11 Nomor 1 April 2019
Terakreditasi peringkat 3 SK. No. 28/E/KPT/2019

Naïve Bayes merupakan metode yang mudah diimplementasikan, proses klasifikasi pada Naïve Bayes didasarkan pada probabilitas bersyarat dari salah satu fitur ke fitur seleksi menggunakan algoritma yang ada [1]. Salah satu metode yang dapat melakukan prediksi yaitu SVM, metode ini merupakan salah satu metode yang memiliki performa lebih baik di berbagai aplikasi. Selain itu SVM juga bisa diterapkan untuk data yang berdimensi tinggi akan tetapi sulit digunakan untuk data dengan jumlah yang besar [2]. Sementara itu K-NN merupakan metode yang digunakan untuk analisis klasifikasi, akan tetapi belakangan ini metode ini juga digunakan untuk Prediksi [3].

Diantara ketiga metode tersebut di atas K-NN yang paling sering digunakan dalam hal peramalan atau prediksi, penentuan nilai parameter K yang digunakan pada K-NN akan sangat berpengaruh pada hasil RMSE yaitu 0,06 sedangkan akurasi metode yang dihasilkan sebesar 98,7 % untuk tipe data univariat [4]. Sementara Naïve Bayes mampu memprediksi kelancaran pembayaran dengan akurasi sebesar 71,97 % [5]. Sedangkan dengan menggunakan model SVM dalam prediksi harga Saham akurasi prediksinya sebesar 0.477 [6]. Adapun penelitian ini untuk mengetahui hasil akurasi tertinggi dari Algoritma Naïve Bayes, SVM dan K-NN untuk prediksi kelancaran pembayaran TV. Kabel, tujuannya adalah untuk memperoleh akurasi yang terbaik dari ketiga metode tersebut. Manfaat dari penelitian ini diharapkan dapat memberikan masukan baik untuk perkembangan ilmu pengetahuan ataupun teknologi, khusunya pada bidang ilmu komputer yaitu berupa uji coba metode sehingga bisa mengetahui metode yang terbaik dalam memprediksi kelancaran pembayaran.

2. Metode

2.1. Metode Pengumpulan Data

Penelitian ini menggunakan data multivariat yang di dapatkan langsung dari PT. Mimoza TV Kabel di Gorontalo, adapun attribut yang digunakan adalah Jenis Pemasangan, Status Rumah, Pekerjaan, Tanggungan, Status Perkawinan, Jenis Penghasilan, dan Penghasilan dengan jumlah record sebanyak 100.

2.2. Metode Analisa Data

Analisa data pada penelitian ini menggunakan tiga metode sebagai perbandingan yaitu Metode Naïve Bayes, Support Vector Machine dan K-NN.

2.2.1. Naïve Bayes

Naïve Bayes adalah sebuah metode dengan teknik prediksi probabilitas dengan berdasarkan pada penerapan teorema bayes dimana antara suatu fitur dengan fitur lain dalam suatu data itu tidak saling ketertaikatan, teknik metode ini merupakan salah satu bentuk sederhana untuk klasifikasi dengan persamaan dapat dilihat pada persamaan (1) [7].

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \qquad (1)$$

Dimana :
y = data kelas yang belum diketahui
x = hipotesis data y
P(x | y) = probabilitas hipotesis x berdasarkan kondisi y
P(x)= probabilitas hipotesis x
P(y | x)=probabilitas y berdasarkan kondisi pada hipotesis x
P(y)= probabilitas dari y

2.2.2. *Support Vector Machine* (SVM)

Parameter algoritma *Support Vector Machine* yang di gunakan adalah C (*cost*) dan Kernel, selanjutnya mencari parameter mana yang mempunyai nilai terbaik. Setelah itu membandingkan variabel mana yang mendapatkan hasil prediksi terbaik. Support Vector Machine merupakan salah satu teknik yang baru akan tetapi memiliki performa yang lebih baik dibandingkan dengan yang lain terutama dalam klasifikasi teks, dan pengenalan tulisan tangan [8].

Konsep pada SVM bermula pada masalah klasifikasi dari dua kelas latih positif dan negatif. Selain itu metode ini mencoba menemukan pemisah terbaik sehingga dapat memaksimalkan batasan antara

DOI: http://dx.doi.org/10.33096/ilkom.v11i1.408.11-16

dua kelas tersebut. Pada beberapa kasus yang sudah pernah dilakukan, data tidak bisa diklasifikasi menggunakan metode *linier* SVM, sehingga dikembangkan fungsi *kernel* untuk mengklasifikasikan data dalam bentuk non-*linier* [7]. *Sequential Training* memiliki *algoritma* yang lebih sederhana dan waktu yang diperlukan lebih cepat. Adapun *algoritme Sequential Training* adalah sebagai berikut:

1. Menginisialisasi dan parameter lain, misalnya λ = 0,5, γ = 0,01, *c* = 1, IterasiMax = 100, dan ε = 0,001. Kemudian menghitung *matriks Hessian* dapat dihitung dengan rumus persamaan (2).

$$D_{ij} = y_i y_j (K(x_i, x_j) + \lambda^2) \quad (2)$$

2. Mulai dari data ke *i* sampai *j*, hitung menggunakan persamaan (3)(4)(5)

a. $E_i = \sum_{j=1}^{n} a_j D_{ij}$ (3)

b. $\delta a_i = \min\{\max[\gamma(1 - E_i), -a_i], c - a_i\}$ (4)

c. $a_i = a_i + \delta a_i$ (5)

3. Langkah ke dua dilakukan secara terus menerus hingga kondisi iterasi maksimum tercapai. Setelah itu didapatkan nilai *support vector* (SV), SV = (*Treshold* SV). Nilai tersebut didapatkan dari beberapa percobaan, biasanya digunakan *Treshold >0.* Kemudian dilakukan proses testing untuk mendapatkan keputusan dimana fungsi keputusan dapat dihitung dengan persamaan (6)

$$f(x) = \sum_{i=1}^{m} a_i \, y_i K(x_i, x) + b \quad (6)$$

Dimana nilai b dengan persamaan (6)

$$-\frac{1}{2}[\sum_{i=1}^{m} a_i y_i K(x_i, x^+) + \sum_{i=1}^{m} a_i y_i K(x_i, x^-)] \quad (7)$$

2.2.3. K- *Nearest Neighbor* (K- NN)

Ketepatan Algoritma ini akan sangat dipengaruhi oleh adanya fitur yang tidak relevan, jika bobot fitur tersebut tidak setara atau relevan terhadap klasifikasi. Beberapa Riset yang menggunakan algoritma ini sebagian besar membahas bagaimana memilih serta memberi bobot terhadap fitur supaya performa algoritma menjadi lebih baik untuk klasifikasi [3]. Metode K-NN bekerja dengan mencari jarak terdekat antara data yang nantinya akan dievaluasi dengan K tetangga (neighbor) terdekat dalam data pelatihan. Sementara itu data training ditampilkan ke ruang yang berdimensi banyak, yang masing-masing dimensi menjelaskan fitur dari data[9]. Adapun tahap-tahap untuk menghitung K-NN adalah sebagai berikut [10]:

a. Menentukan Nilai K.
b. menghitung jarak antara dengan persamaan (8).

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (8)$$

Keterangan :
X = sampel data
Y = data uji
D = Jarak

c. Mengurutkan hasil jarak dan menetapkan tetangga terdekat berdasarkan jarak minimum terhadap -K
d. Gunakan mayoritas sederhana dari kelas tetangga terdekat sebagai nilai prediksi data baru

Sementara untuk mencari nilai prediksi k-NN dapat dihitung dengan persamaan (9).

$$Y = \frac{1}{K} \sum_{i=1}^{k} yi \quad (9)$$

DOI: http://dx.doi.org/10.33096/ilkom.v11i1.408.11-16

p-ISSN 2087-1716
e-ISSN 2548-7779

ILKOM Jurnal Ilmiah Volume 11 Nomor 1 April 2019
Terakreditasi peringkat 3 SK. No. 28/E/KPT/2019

Dimana:
Y = Perkiraan
K= jumlah tetanga terdekat
Yi = output tetangga terdekat

2.3. Model Komparasi

Penelitian ini akan menggunakan *Area Under Curve* (AUC) untuk mengukur performa akurasi yang dihasilkan dari algoritma. Algoritma yang memiliki nilai AUC diatas 0.6 merupakan algoritma yang mempunyai performa cukup efektif sementara itu model komparasi akan menggunakan test parametric dan non-parametric atau T-Test.

2.4. Evaluasi

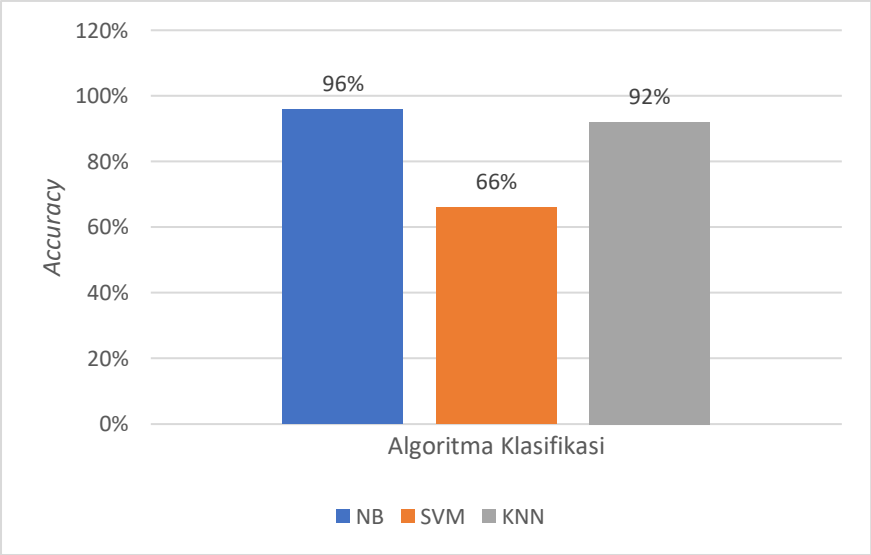
Dari hasil pengujian diatas, akan di evaluasi menggunakan rapid miner sehingga pada akhirnya akan dapat dilihat hasil pengujian dari ketiga model tersebut untuk dataset dengan variabel (Jenis Pemasangan, Status Rumah, Pekerjaan, Tanggungan, Status Perkawinan, Jenis Penghasilan dan Penghasilan) tingkat nilai akurasi trend prediksi yang lebih tinggi pada algoritma yang mana untuk dapat disimpulkan algoritma mana yang tingkat akurasi prediksinya tertinggi.

3. Hasil dan Pembahasan

Hasil dari pengujian yang telah dilakukan yaitu membandingkan dari ketiga metode tersebut mana yang lebih akurat adapun perbandingan performance dari masing-masing model algoritma dapat dilihat pada table 1, gambar 1 dan gambar 2.

Tabel 1. Hasil Perbandingan Metode

	Naïve Bayes	Support Vector Machine	K-Neural Network
Accuracy	96 %	66 %	92 %
AUC	0.99	0.786	0.971

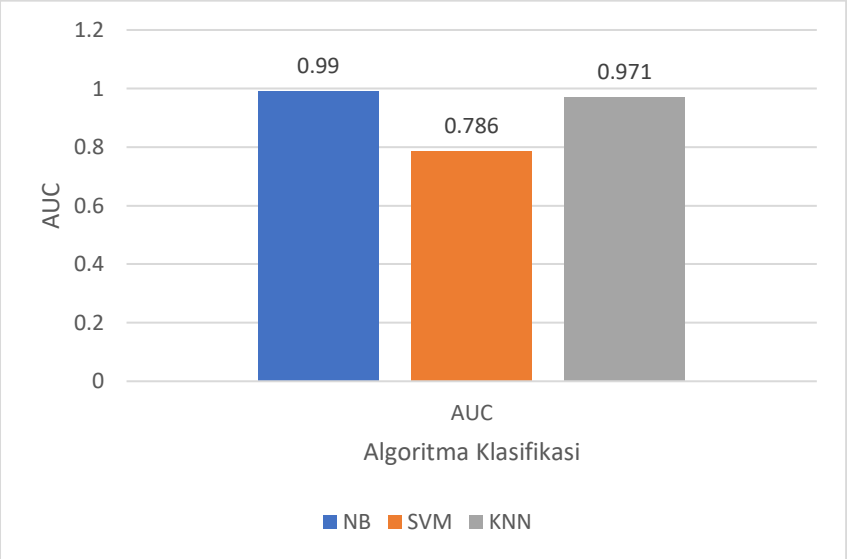


Gambar 1 Komparasi Accuracy Algoritma Klasifikasi

DOI: http://dx.doi.org/10.33096/ilkom.v11i1.408.11-16

p-ISSN 2087-1716
e-ISSN 2548-7779

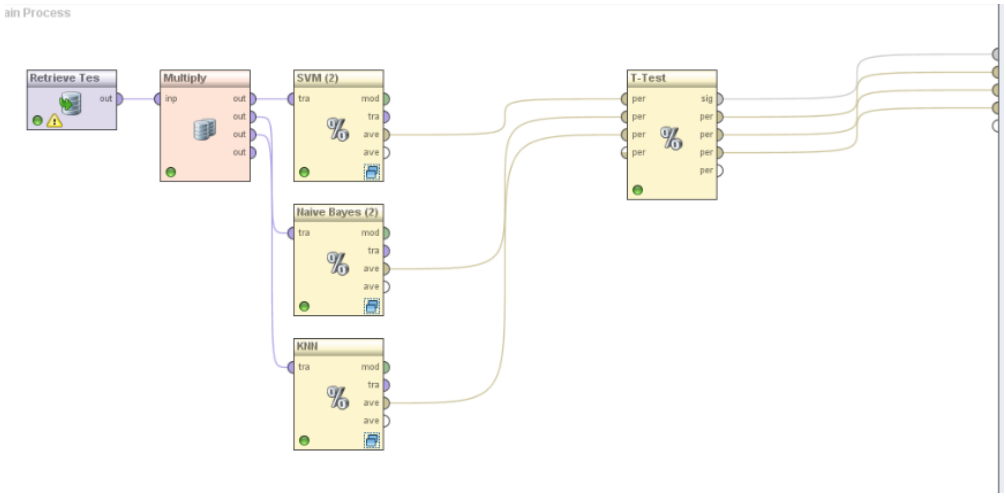
ILKOM Jurnal Ilmiah Volume 11 Nomor 1 April 2019
Terakreditasi peringkat 3 SK. No. 28/E/KPT/2019



Gambar 2 Komparasi AUC Algoritma Klasifikasi

Dapat diketahui bahwa algoritma *Naïve Bayes* unggul dari algoritma K-NN dan SVM untuk tipe data multivariat dengan nilai *accuracy* yaitu 96 %, sementara algoritma K-*Neural Network* di uji coba dengan beberapa nilai K yaitu antara 3, 5,7 dan 9, didapatkan *accuracy* tertinggi yaitu pada Nilai K = 3 yaitu 92% hasil ini sedikit berbeda dengan penelitian sebelumnya dimana metode K-NN sangat baik dalam prediksi akan tetapi apabila digunakan pada tipe data multivariat pada penelitian ini accuracy yang di hasilkan masih lebih baik dari Naïve bayes, sementara pada *Support Vector Machine* hanya mendapatkan *accuracy* sebesar 66 %. Sedangkan pada uji coba ROC *Curve* menunjukkan bahwa *Naïve Bayes* mencapai nilai AUC yang terbaik yaitu 0.99, kemudian K-*Neural Network* 0,971 dan *Support Vector Machine* 0,786.

Selanjutnya akan dilakukan pengujian dengan memanfaatkan uji statistik yaitu dengan menguji T-Test pada RapidMiner dengan desain model ditunjukan pada gambar 3.



Gambar 3 Skema Pengujian T-Test Pada Rapid Miner

Dalam pengujian T-test ini akan melihat hasil uji dari ketiga algoritma klasifikasi tersebut agar mendapatkan nilai yang terbaik, dimana dalam pengujian tersebut algoritma yang mendapatkan nilai terkecil ≤ 0.05 dinyatakan sebagai hasil uji terbaik. Adapun hasil T-tes dari algoritma *Naïve Bayes*, *Support Vector Machine* dan K-*Neural Network* dapat dilihat pada gambar 4.

DOI: http://dx.doi.org/10.33096/ilkom.v11i1.408.11-16

T-Test Significance

	0.660 +/- 0.049	0.930 +/- 0.064	0.880 +/- 0.075
0.660 +/- 0.049		**0.000**	**0.000**
0.930 +/- 0.064			0.155
0.880 +/- 0.075			

Gambar 4 Hasil T- Test Significance

Berdasarkan hasil T-test didapatkan bahwa Metode naïve bayes dan K-nn sangat dominan terhadap metode lain, dan hasil accuracy yang didapatkan pun sangat baik dibandingkan dengan metode SVM.

4. Kesimpulan dan Saran

Berdasarkan hasil komparasi antara algoritma klasfikasi data mining yaitu *Naïve Bayes*, K-*Neural Network* dan *Support Vector Machine* untuk prediksi kelancaran pembayaran dengan menggunakan data multivariat yaitu 100 data set nasabah PT. Mimoza TV Kabel yang ada di Kota Gorontalo, maka didapatkan hasil analisa bahwa metode *Naïve Bayes* merupakan algoritma yang akurat menghasilkan *accuracy* yaitu 96 %, dan metode ini juga sangat dominan terhadap metode lain. Berdasarkan nilai *Accuracy*, AUC dan T-tes ketiga metode ini masuk dalam kategori sangat baik dalam klasfikasi ataupun prediksi.

Saran untuk penelitian selanjutnya menabahkan beberapa metode klasifkasi lain untuk dikomprasi metode, dengan menambahkan jumlah atrribut ataupun komparasi dengan algoritma *feature seleksi*

Daftar Pustaka

[1] W. Zhang and F. Gao, “Procedia Engineering An Improvement to Naive Bayes for Text Classification,” vol. 15, pp. 2160–2164, 2011.
[2] A. S. Nugroho, A. B. Witarto, and D. Handoko, “Support Vector Machine,” 2003.
[3] A. Bode, “K-NEAREST NEIGHBOR DENGAN FEATURE SELECTION MENGGUNAKAN BACKWARD ELIMINATION UNTUK PREDIKSI HARGA KOMODITI KOPI ARABIKA,” vol. 9, pp. 188–195, 2017.
[4] M. E. Lasulika, “PREDIKSI HARGA KOMODITI JAGUNG MENGGUNAKAN K-NN DAN PARTICLE SWARM OPTIMAZATION,” vol. 9, pp. 233–238, 2017.
[5] M. Hasan, “Menggunakan Algoritma Naive Bayes Berbasis,” vol. 9, no. Desember, pp. 317–324, 2017.
[6] R. H. Kusumodestoni and S. Sarwido, “Komparasi Model Support Vector Machines (Svm) Dan Neural Network Untuk Mengetahui Tingkat Akurasi Prediksi Tertinggi Harga Saham,” *J. Inform. Upgris*, vol. 3, no. 1, 2017.
[7] S. Dewi, “Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan,” *Techno Nusa Mandiri*, vol. XIII, no. 1, pp. 60–66, 2016.
[8] D. B. A. Mezghani, S. Z. Boujelbene, and N. Ellouze, “Evaluation of SVM Kernels and Conventional Machine Learning Algorithms for Speaker Identification,” *Int. J.*, vol. 3, no. 3, pp. 23–34, 2010.
[9] S. B. Imandoust and M. Bolandraftar, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background,” vol. 3, no. 5, pp. 605–610, 2013.
[10] E. Prasetyo, “Fuzzy K-Nearest Neighbor in Every Class Untuk Klasifikasi Data,” *Semin. Nas. Tek. Inform. (SANTIKA 2012)*, no. Santika, pp. 57–60, 2012.

