Analisis *Cluster Single Linkage* Berdasarkan Potensi Desa Di Kabuapten Kutai KartannegaraTahun 2019

Single Linkage Cluster Analysis Based on Village Potential In Kutai Kartannegara Regency in 2019

Suyanto[1], Syarippudin[2], Wasono[2]
[1]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman
[2,3]Laboratorium Matematika Komputasi FMIPA Universitas Mulawarman
Email: shuyanto95@gmail.com

Abstract

Data mining is a step in the process of Knowledge Discovery in Database (KDD) which consists of the application of data analysis and the discovery of algorithms that produce certain enumerations of patterns in the data,Cluster Analysis is one of the methods in multivariate statistical analysis that is used to group objects into groups based on their characteristics, so the objects in one group have more homogeneous characteristics compared to objects in other groups. Single Linkage is a clustering process based on the closest distance between objects. If two objects are separated by a short distance, then the two objects will merge into one cluster. This study aims to obtain a cluster of village potential in Kutai Kartanegara Regency in 2019, based on the variable availability of educational facilities, the availability of health facilities, the availability of health workers, the availability of coin / card public telephones, the existence of lodging, the existence of market buildings, the existence of supermarkets, the existence of banks, the population obtaining credit facilities, the existence of other Non KUD cooperatives., Based on the results of the analysis, it can be seen that, Clusters formed in the grouping of potential villages / villages in Kutai Kartanegara Regency using a single linkage method are as many as 2 clusters.

Keywords: Data Mining, Cluster Analysis, Single Linkage, Village Potential, multivariate,

Pendahuluan

Data *mining* adalah suatu proses untuk mendapatkan informasi yang berguna dari gudang basis data berskala besar yang membantu dalam proses pengambilan keputusan. *Data mining* merupakan gabungan dari sejumlah disiplin ilmu, yang didefinisikan sebagai proses penemuan suatu pola baru dari kumpulan data berskala besar, yaitu *artificial intelligence* (kecerdasan buatan), *machine learning* (pembelajaran mesin), *statistics* (statistika) dan *database system* (teknologi basis data) (Prasetyo, 2012).

Analisis multivariat adalah analisis statistik yang digunakan untuk menganalisis data yang terdiri dari beberapa variabel dan variabel-variabel tersebut saling berkolerasi satu sama lain. Secara umum analisis multivariat dibagi menjadi dua, yaitu analisis dependensi dan analisis interdependensi. Dalam analisis dependensi karakteristik utama pada analisis ini adalah satu variabel atau lebih diidentifikasikan sebagai variabel tak bebas yang akan diprediksi atau diterangkan oleh variabel-variabel lain yang diketahui sebagai variabel bebas. Ciri dari analisis dependensi adalah adanya satu atau beberapa variabel yang berfungsi sebagai variabel tak bebas dan variabel bebas, seperti, analisis regresi linear berganda, analisis diskriminan, analisis logit, dan analisis korelasi kanonik. Dalam analisis interdependensi tidak terdapat satu pun variabel yang didefinisikan sebagai variabel bebas atau variabel tak bebas karena prosedur mencakup analisis yang dilakukan secara bersamaan untuk semua variabel dalam sekumpulan variabel yang diobservasi. Ciri dari analisis interdependensi adalah semua variabelnya bersifat independen. Berikut ini yang termasuk dalam analisis interdependensi adalah analisis faktor, analisis cluster dan multidimensional scaling (Sarwono, 2007).

Analisis Klaster merupakan salah satu metode dalam analisis statistika multivariat yang digunakan untuk mengelompokkan objek-objek ke dalam suatu kelompok berdasarkan karakteristik yang dimiliki, sehingga objek-objek dalam suatu kelompok memiliki ciri-ciri yang lebih homogen dibandingkan dengan objek dalam kelompok lain. Setiap unit pengamatan dalam satu kelompok akan memiliki ciri yang relatif sama sedangkan antar kelompok unit pengamatan memiliki sifat yang berbeda. Secara umum analisis klaster dibagi menjadi dua metode yaitu metode hierarki dan metode non-hierarki. Di dalam metode hierarki sendiri terdapat beberapa metode, metode-metode yang termasuk dalam metode hierarki diantaranya metode pautan tunggal (*single linkage*), metode pautan lengkap (*complete linkage*), dan metode pautan rata-rata (*average linkage*), sedangkan metode yang termasuk dalam metode non-hierarki diantaranya metode K-Means (Sarwono, 2007).

Analisis Klaster *Single linkage* adalah suatu metode statistik yang mengidentifikasi kelompok sampel berdasarkan karakteristik serupa, analisis klaster mengelompokan elemen mirip sebagai objek penelitian yang mempunyai homogenitas yang tinggi antar objek menjadi klaster yang berbeda dengan tingkat homogentias yang tinggi antar klaster. Pengclusteran ini didasarkan pada gugus variabel yang di pertimbangkan untuk diteliti(Han, J, 2006).

Pada tahun 1993, 1994 dan 1995 badan pusat statistik telah melakukan pengklasifikasian desa tertinggal. Hal tersebut dilakukan kembali pada tahun 2005 yang dilakukan dalam rangka penyaluran bantuan pemerintah sebagai kompensasi bahan bakar minyak yang dilakukan sampai saat ini. Data yang digunakan untuk penentuan desa tertinggal pada tahun 1993 adalah data PODES dengan faktor penentu menggunakan 25 variabel untuk daerah perkotaan dan 27 variabel untuk daerah pedesaan. Identifikasi status ketertinggalan desa pada tahun 1994 menggunakan 17 variabel untuk daerah perkotaan dan 18 variabel untuk daerah pedesaan. Sedangkan untuk tahun 2005 data yang digunakan adalah data potensi desa sehingga yang dihasilkan adalah model untuk penentu status ketertinggalan desa.(BPS, 2005)

Berdasarkan latar belakang tersebut maka peneliti tertarik malakukan pengelompokan terhadap potensi desa yang ada di Kabupaten Kutai Kartanegara, analisis yang digunakan adalah analisis klaster *single linkage* yang bertujuan untuk mengelompokan isi variabel (potensi-potensi yang ada di setiap desa di Kabupaten Kutai Kartanegarapada Tahun 2019). Pada analisis ini juga akan meneliti seluruh hubungan interdependensi, tidak ada pembedaan variabel bebas dan tidak bebas (*dependent and independent variables)*.

Tujuan dari penelitian ini adalah untuk mendapatkan Mendapatkan klaster berdasarkan potensi desa di Kabupaten Kutai Kartanegara tahun 2019 dengan menggunakan analisis klaster *Single linkage* serta 2. Mendapatkan validasi dari analisis klaster Single linkage berdasarkan potensi desa di Kabupaten Kutai Kartanegara tahun 2019

Prosedur Analisis Klaster

1. Merumuskan Masalah

Hal ini paling penting dalam masalah analisis klaster ialah pemilihan variabel-variabel yang akan dipergunakan untuk pengklasteran memasukan satu atau dua variasi variabel yang tidak relevan dengan masalah pengklasteran akan mendistorasi hasil pengklasteran yang kemungkinan besar sangat bermanfaat. Pada dasarnya variabel-variabel yang akan dipilih harus menguraikan kemiripan (*similiarly*) antara objek, yang memang benar – benar relevan dengan masalah yang dibahas.

2. Memilih Ukuran Jarak

Menurut (Supranto, 2004) jarak yang paling umum digunakan adalah jarak Euclidean. Ukuran jarak atau ketidaksamaan antar objek ke-i dengan objek ke-h disimbolkan dengan d_{ih}. Nilai d_{ih} diperoleh melalui perhitungan jarak kuadrat Euclidean sebagai berikut :

$$d_{ih} = \sum_{i=1}^{p} \sqrt{(x_{ij} - x_{hj})} \qquad (1)$$

dengan d_{ih} = jarak *kuadrat Euclidean* antar objek ke-*i* dengan objek ke-*h* p = jumlah variabel klaster

x_{ij} = nilai atau data dari objek ke-*i* pada variabel ke-*j* dan x_{hj} = nilai atau data dari objek ke-*h* pada variabel ke-*j*.

3. Memilih Prosedur Pengklasteran

Proses pembentukan *cluster* dapat dilakukan dengan dua cara, yaitu dengan metode hierarki dan non hierarki. Pada metode hierarki terdiri dari metode agglomerative dan metode devisif. Metode agglomerative sendiri terdiri dari 3 metode, yaitu metode linkage, metode variance, dan metode centroid, dimana linkage terdiri dari metode single linkage, complete linkage, dan average linkage. Sedangkan pada metode variance terdiri dari metode ward.

Metode single linkage adalah proses klastering yang didasarkan pada jarak terdekat antara objeknya. Jika dua objek terpisah oleh jarak pendek, maka kedua objek tersebut akan bergabung menjadi satu klaster dan demikian seterusnya. Pengukuran jarak metode ini dapat ditulis dalam rumus berikut:

$$d_{w(i,j)} = min\{d_{wi}.d_{wj}\} \qquad (2)$$

4. Standarisasi Data

Variabel yang memiliki nilai besar mempunyai pengaruh yang lebih besar dalam dalam melakukan prediksi klasifikasi daripada variabel dengan nilai kecil. Untuk mengatasi masalah tersebut, dapat digunakan teknik normalisasi variabel sehingga semua variabel akan berbeda dalam jangkauan yang sama. Cara menentukan nilai standarisasi adalah dengan menghitung nilai mean dan variansi dari masing-masing variabel dari masing-masing variabel.

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik} \qquad (3)$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_i)^2 \qquad (4)$$

$$\hat{x}_{ik} = \frac{x_{ik} - x_k}{\sigma_k} \qquad (5)$$

dimana,

N = Jumlah data

x_{ik} = data ke-*i* variabel ke-*k* dimana k = 1,2,…,*n*

$\bar{x}_i$ = rata-rata variabel ke-*k*
σ_k = simpangan baku variabel ke-*k*
σ_k^2 = variansi variabel ke-*k*
$\hat{x}_{ik}$ = normalisasi data ke-*i* variabel ke-*k*

5. Multikolinieritas

Analisis koefisien korelasi bertujuan untuk mempelajari apakah ada hubungan antara dua variabel. Koefisien korelasi antar variabel independen haruslah lemah (dibawah 0.8). Jika korelasi kuat, terjadilah problem multikolinearitas. Koefisien korelasi dirumuskan sebagai berikut :

$$r_{ab} = \frac{n\sum_{i=1}^{n} ab - (\sum_{i=1}^{n} a)(\sum_{i=1}^{n} b)}{\sqrt{\{n\sum_{i=1}^{n} a^2 - (\sum_{i=1}^{n} a)^2\}\{n\sum_{i=1}^{n} b^2 - (\sum_{i=1}^{n} b)^2\}}} \quad (6)$$

Dimana a dan b adalah variabel bebas (independen) pada model, sedangkan n adalah banyaknya sampel yang digunakan (Santoso, 2012).

6. Menentukan Banyaknya Klaster

Dalam analisis klaster adalah menentukan berapa banyaknya klaster. Karena tidak ada aturan yang baku untuk menentukan berapa banyak klaster , namun demikian ada beberapa pentunjuk yang bisa dipergunakan yaitu :

1. Pertimbangan teoritis, konseptual, praktis, mungkin bias diusulkan atau disarankan untuk menentukan berapa banyaknya klaster yang sebenarnya
2. Dalam pengklasteran hirarki, jarak dimana klaster yang digabung bias digunakan sebagai criteria

7. Menginterpretasi Dan Memprofil Klaster

Pada tahap interpretasi meliputi pengujian pada masing-masing klaster yang terbentuk untuk memberikan nama atau keterangan secara tepat sebagai gambaran sifat dari klaster tersebut, menjelaskan bagaimana mereka bisa berbeda secara relevan pada setiap dimensi.

8. Melakukan Validasi Klaster

Untuk menguji validasi klaster digunakan uji parsial F dengan tarag signifikansi α

H_0 : Variabel xj bukan variabel pembeda dalam pengklasteran.

H_1 : Variabel xj merupakan variabel pembeda dalam pengklasteran

Statistik uji

$$F = \frac{MSC}{MSE} \quad (6)$$

Tolak H_0 jika nilai F hitung > Fα.k-1, n-k

Hasil dan Pembahasan

Data yang digunakan dalam penelitian ini menggunakan data potensi desa pada 237 desa/kelurahan yang ada di Kabupaten Kutai Kartanegara pada tahun 2019 dan sebagai sampel dipilih 81 desa/kelurahan yang ada di Kabupaten Kutai Kartanegara pada tahun 2019 menggunakan *proportional stratified random sampling*. Berikut adalah variabel dari data penelitian

X_1= Ketersediaan sarana pendidikan Bangunan
X_2= Ketersediaan sarana kesehatan Jiwa
X_3= Ketersediaan tenaga kesehatan Jiwa
X_4= Ketersediaan telepon umum koin/kartu Unit
X_5= Keberadaan penginapan/motel/losmen/wisma Bangunan
X_2= Keberadaan bangunan pasar permanen/semi permanen Bangunan
X_7= Keberadaan supermarket/minimarket/pasar swalayan Bangunan
X_8= Keberadaan bank Unit
X_9= Penduduk memperoleh fasilitas perkreditan Jiwa
X_{10}= Keberadaan Koperasi Non KUD lainnya Jiwa

Standarisasi data

Perhitungan standarisasi data dilakukan dengan bantuan persamaan 3 persamaan 4 dan persamaan 5

$$\bar{x}_i = \frac{1}{81}\sum_{1=1}^{81} x_{i1} = 6{,}642$$

Nilai standar deviasi pada variabel X_1 dihitung sebagai berikut:

$$s_1 = \sqrt{\frac{1}{81-1}\sum_{i=1}^{81}(x\_i1 - \bar{x}\_1)^2} = 5{,}30$$

Nilai Standardisasi seluruh data pengamatan pada variabel X_1 dihitung dengan menggunakan Persamaan 5

$$\hat{x}_{1,1} = \frac{x_{1,1} - \bar{x}_1}{s_1} = \frac{3 - 6{,}6419}{5{,}30} = -0{,}6873$$

$$\hat{x}_{2,1} = \frac{x_{2,1} - \bar{x}_1}{s_1} = \frac{2 - 6{,}6419}{5{,}30} = -0{,}876$$

Perhitungan Standardisasi untuk setiap data dan hasilnya secara lengkap dapat dilihat pada Tabel 1.

Tabel 1 Hasil Perhitungan standarisasi data

No	X_1	X_2	X_3	…	X_{10}
1	-0,69	-0,31	-0,62	…	-0,36
2	-0,88	-0,67	-0,79	…	2,12
3	-0,31	-0,31	-0,45	…	-0,36
⋮	⋮	⋮	⋮	⋱	⋮
81	-1,06	-1,04	-0,62	…	-0,36

Pendeteksian Multikolinieritas

Pendeteksian adanya multikolinieritas salah satu caranya adalah dengan melihat nilai mutlak koefisen korelasi antar variabel dengan menggunakan persamaan 6.

Tabel 2 Pendeteksian Multikolinieritas

Var	X1	X2	X3	X4	…	X10
X1	1	0,77	0,68	0,48	…	0,15
X2	0,77	1	0,79	0,29	…	0,04
X3	0,68	0,79	1	0,16	…	0,09
X4	0,48	0,29	0,16	1	…	0,01
⋮	⋮	⋮	⋮	⋮	⋱	⋮
X10	0,15	0,04	0,09	0,01	…	1

Berdasarkan hasil Tabel 2 terlihat bahwa nilai mutlak dari koefisien korelasi antar variabel penelitian yang berbeda bernilai dibawah 0,8 yang artinya tidak ada multikolinieritas antar variabel dalam penelitian dan dapat dilanjutkan ke proses pengelompokan denga metode analisis klaster yaitu *single linkage*.

Perhitungan Jarak *Euclidean*

Perhitungan jarak Euclid antar variabel dapat dilakukan dengan menggunakan Persamaan 1. Perhitungan jarak Euclid diperoleh dengan menggunakan bantuan software-R dan salah satu contoh perhitungannya adalah sebagai berikut.

$$d_{i,j} = \sqrt{\sum_{k=1}^{p} (X\_ik - X\_jk)^2}$$

$$d_{1,2} = \sqrt{\sum_{k-1}^{p} (X\_1k - X\_2k)^2} = \sqrt{9{,}617918} = 3{,}101$$

Perhitungan Jarak untuk setiap data hasilnya secara lengkap dapat dilihat pada Tabel 3.

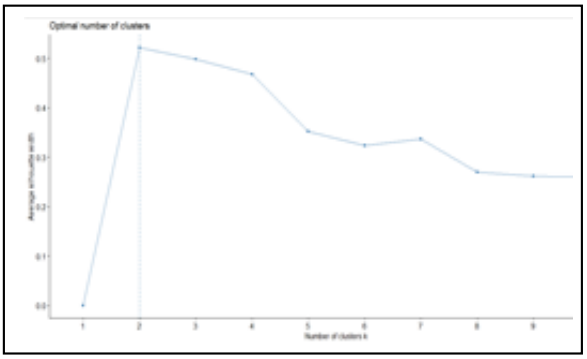
Tabel 3 Hasil Perhitungan Jarak Euclidean

No	1	2	3	4	…	79	80	81
1	0				…			
2	3,10	0			…			
3	1,87	4,47	0		…			
4	1,87	2,57	3,62	0	…			
⋮	⋮	⋮	⋮	⋮	0			
79	2,37	2,97	3,88	1,48	…	0		
80	3,79	4,99	4,13	4,22	…	4,17	0	
81	2,15	2,77	3,78	1,15	…	1,11	4,42	0

Berdasarkan hasil perhitungan jarak euclid antara desa Tani Bakti dan Salok Api Laut memiliki jarak terdekat sebesar 3,101276758. Hal ini menunjukan bahwa desa Tani Bakti dan Salok Api Laut memiliki kemiripan secara karakteristik potensi desa. Sedangkan antara desa Tani Bakti dan Ambarawang Laut memiliki jarak terdekat sebesar 1.871339392. Demikian pula untuk penafsiran objek lainnya.

Penentuan Jumlah Klaster optimal

Setelah melakukan perhitungan jarak dari masing-masing objek selanjutnya adalah mencari nilai K optimum atau banyaknya klaster.



Gambar 1 Nilai Otpimun Klaster

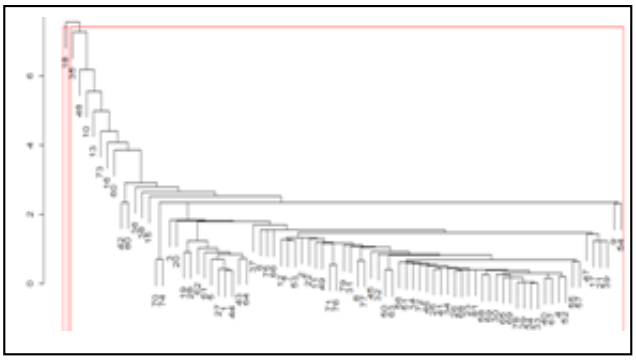
Dari Gambar 1 dapat dilihat bahwa nilai K optimumnya berada pada titik ke dua yang berati nilai K optimum yang dapat pada penelitian ini adalah dua klaster. Setelah peneliti mendapatkan nilai K yang optimum, senajutnya adalah melakukan klastering dengan menggunakan metode *single linkage*.

Analisis Klaster *Single Linkage*

Analisis klaster dengan menggunakan metode *Single linkage* perlu menentukan titik pusat awal untuk mengelompokkan objek penelitian. Pada penelitian ini, berasarkan hasil perhitunggan martiks jarak Euclid didapatkan pusat klaster awal yaitu pada kolom 24 baris 33 yang bernilai 0,028. Selanjutnya di lakukan analisis menggunakan persamaan 2

$$d_{w(i,j)} = min\{d_{wi} d_{wj}\}$$
$$d_{w(24,34)1} = \min\{d_{24,1} d_{33,1}\}$$
$$d_{w(24,33)1} = \min(1{,}970)(1{,}981) = 1{,}97$$
$$d_{w(24,33)2} = \min(2{,}678)(2{,}689) = 2{,}678$$
$$d_{w(24,33)80} = \min(4{,}290)(4{,}576) = 4{,}290$$
$$d_{w(24,33)81} = \min(0{,}434)(0{,}426) = 0{,}426$$

Perhitungan yang sama dilakukan pada pusat klaster ke-2 sampai dengan data ke-81 untuk semua variabel. Hasil data dapat dilihat pada Gambar 2. Berdasarkan Gambar 2 dapat dilihat dendogram hasil pembentukan klaster dari data potensi desa memperlihatkan data mana saja yang tergolong dalam dalam kalster satu dan klaster dua, dimana klaster satu beranggotakan 80 anggota dan klaster untuk klaster dua beranggotakan 1 anggota.



Gambar 2 Hasil analisis Klaster *Single Linkage* berdasarkan data yang digunakan

Validasi Klaster

Untuk melihat apakah variabel-variabel yang telah membentuk klaster tersebut merupakan variabel pembeda dalam pengklasteran dapat dilihat pada tabel anova.

Hipotesis :

H_0 : Variabel X bukan variabel pembeda dalam pengklasteran

H_1 : Variabel X merupakan variabel pembeda dalam pengklasteran

Kriteria pengujian

Tolak H0 jika nilai P-value < α (0,05)

Tabel 4 Anova

Mean	DB	SSE	MSE	P-*Value*
Antar Group	9	1,48	1,64	1
Dalam Group	800	800	1	
Total	800	800		

Dari tabel Anova dapat dilihat bahwa nilai *P-Value* yang didapat adalah sebesar 1 > 0,05, maka dapat disimpulkan bahwa tidak terdapat variabel pembeda dalam pengklasteran.

Kesimpulan

Berdasarkan hasil dari pegolahan data dan analisis yang telah dilakukan terhadap 81 data potensi desa Kabupaten Kutai Kartanegara tahun 2019 di peroleh kesimpulan sebagai berikut :

1. Klaster yang terbentuk pada pengelompokan potensi desa/kelurahan di Kabupaten Kutai Kartanegara dengan menggunakan metode single linkage adalah sebanyak 2 klaster yaitu klaster 1 dan klaster 2. Klaster 1 yang beranggotakan 80 desa/kelurahan dimana seluruh desa masuk pada klaster 1 kecuali desa nomor 18, dan klaster 2 yang beranggotakan 1 desa kelurahan yaitu desa no 18 Loa Kulu Kota.
2. Hasil pengujian Silhouette Coefficient dan ANOVA untuk validasi data hasil klastering potensi desa di Kabupaten Kutai Kartanegara dengan menggunakan metode single linkage adalah sebesar 2, yang berarti klaster optimal yang dapat dibentuk pada penelitian ini adalah 2 klaster dengan nilai P-value (1) > α (0,05) yang berarti tidak terdapat variabel pembeda dalam pengklateran.

Daftar Pustaka

Badan Pusat Statistik Kalimantan Timur, 2018. Website Badan Pusat Statistik Kalimantan Timur 2018. Samarinda

Edy. 2009. Manajemen Sumber Daya Manusia, Jakarta. Kencana Pernada Media Group

Gujarati, D. 2009. Ekonometrika Dasar. Erlanga. Bandung.

Han, J., & Kamber, M. (2006). Data Mining: Concept and Techniques. San Fransisco: Morgan Kauffman Publisher.

Johnson. 1998 R.A & Wichern, DW. (2002). Applied Multivariate Statistical Analysis Third Edition. New Jersey: Prentice Hall International.

Kuncoro, M. 2008. Metode Riset Untuk Bisnis dan Ekonometri. Erlanga. Jakarta.

Prasetyo. 2013 Data Mining: Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta: Andi Offset.

Rencher. 2002. Method Of Multivariate Analysis PT. Gramedia Pustaka. Jakarta.

Sarwono. 2007. Analisis Multivariat Arti dan Interpretasi. PT. Rineka Cipta. Jakarta

Supranto, J. 2004. Teknik Sampling Untuk Survey dan Eksperimen, Edisi Baru. Jakarta. PT. Rineka Cipta. Jakarta

Sugiyono. (2014). Metode Penelitian Kuantitaif Kualitatif dan R&D. Bandung: Alfabeta.

Sutopo, (1996) Pengertian, Tipologi Desa, Karakteristik Desa

Suyanto. (2017). Data Mining untuk Klasifikasi dan Klasterisasi Data. Bandung : Informatika.

Simamora, Bilson. 2005. Analisis Multivariat Pemasaran. PT. Gramedia Pustaka. Jakarta.

Sartono, B dkk. 2003. Analisis Peubah Ganda. Bogor. IPB