# KLASIFIKASI TUTUPAN LAHAN PERKOTAAN MENGGUNAKAN *NAIVE BAYES* BERBASIS *FORWARD SELECTION*

**M. Salim***
*) Tenaga Pengajar pada Sekolah Tinggi Manajemen Informatika & Komputer ( STMIK ) Ichsan Gorontalo
E-mail: Mohsalim87@yahoo.com

***Abstract*** *: Urban growth as one of the economic symptoms related to the process of urbanization and population displacement in a major way from the countryside to urban areas has fueled the city's growth issues. These developments will bring up a number of problems when faced with the reality of the limited City area. Urban land cover data that has many attributes with 9 types of target classification using the best attributes of search techniques by applying the forward algorithm selection and naive bayes which merit independently in target and requires only a small amount of training data to determine the required parameter estimation in classification process where accuracy 87,04% better compared to testing using random forest algorithm-based forward selection at the level of 72,72% accuracy. so it can be inferred with previous studies using random forest algorithm with 84,42% accuracy. but better with naive bayes algorithm-based forward selection with increased accuracy percentage level of 2.62%.*

***Keywords****: Dataset Urban land cover, forward selection, naïve bayes, random forest.*

## PENDAHULUAN

### A. Latar Belakang

Permasalahan perkotaan hasil kajian Rossi-Hausberg (2004) menunjukkan bahwa pertumbuhan perkotaan sebagai salah satu gejala ekonomi berkaitan dengan proses urbanisasi. Urbanisasi dapat menghambat pertumbuhan ekonomi dan kegiatan ekonomi yang diukur dengan populasi, serta output pendapatan yang merupakan pola konsentrasi variabel ekonomi dan demografi membentuk beberapa gejala ekonomi perkotaan.

Perpindahan penduduk secara besar-besaran dari pedesaan ke perkotaan telah memicu berbagai ragam masalah pertumbuhan di kota karena para pencari kerja menginginkan gaji tinggi dan fasilitas yang lebih baik. Banyak kota-kota besar di dunia seperti Amerika, Prancis dan China yang sangat produktif, namun ada beberapa masalah yang dihadapi misalnya kemacetan, tingkat kejahatan lebih tinggi dan polusi udara. Selain itu, timbul berbagai macam kasus seperti taman

yang merupakan paru-paru kota diubah fungsinya menjadi kawasan komersial seperti pompa bensin, supermarket atau *department store* yang mengakibatkan timbulnya berbagai masalah lingkungan. Dampak yang ditimbulkan sangat menyedihkan, mulai dari ketidaknyamanan penduduk akibat kurangnya sarana dan prasarana lingkungan, kesengsaraan masyarakat akibat banjir, sampai masalah sosial karena benturan berbagai kepentingan pemanfaatan lahan.

Pemantauan lingkungan, konservasi, perencanaan pelaksanaan tata ruang atau pengelolaan sumber daya alam yang berorientasi ekosistem dan pengembangan operasional merupakan hal penting dalam studi global dan regional keanekaragaman hayati, konservasi alam dan konsistensi penggunaan lahan sebagai area monitoring yang mampu mengamati, menganalisa, menyajikan serta membuat model keputusan tutupan lahan dengan penginderaan jauh. Penginderaan jarak jauh menggunakan satelit dapat digunakan untuk pemetaan perkotaan dan penggalian informasi tutupan lahan dari data resolusi pixel gambar yang tinggi karena tingkat variabilitas spektral dalam kelas tutupan lahan disebabkan oleh bayangan, sudut matahari yang menyebabkan akurasi klasifikasi rendah. Variabilitas disebabkan karena pixel biasanya hanya mewakili sebagian kecil target (misalnya atap gedung, jalan dan pohon). Informasi spektral dari segmen terkecil mungkin berguna untuk menargetkan setiap pohon, sedangkan ukuran dan bentuk informasi yang lebih besar berguna untuk memisahkan bangunan dari beton dan permukaan yang mirip dengan atap bangunan karena masalah utama dengan pendekatan klasifikasi disemua jenis tutupan lahan.

Penggalian informasi tutupan lahan perkotaan dari data analisis harus mempertimbangkan resolusi spasial dan resolusi spektral. Kualitas resolusi spasial yang meliputi ukuran pixel yang lebih kecil lebih baik dari pada ukuran pixel yang lebih tinggi pada resolusi spektral. Hal ini menjadi penyebab mengapa foto udara secara tradisional menjadi sumber utama untuk perencanaan dan manajemen perkotaan. Tujuan penginderaan jauh adalah untuk meningkatkan visibilitas fitur tersentral, terutama bentuk fisik perkotaan, spektral heterogenitas per-pixel untuk meningkatkan identifikasi tutupan lahan perkotaan.

Klasifikasi lahan merupakan upaya mengelompokkan berbagai jenis tutupan lahan atau penggunaan lahan ke dalam suatu kesamaan obyek. Klasifikasi tutupan lahan penelitian sebelumnya oleh Brian A. Johnson menjelaskan tentang informasi tutupan lahan dan metode berbasis obyek. Masalah yang dibahas adalah sulitnya memilih parameter segmentasi citra dan menentukan ukuran skala segmentasi jenis tutupan lahan, sehingga hanya satu skala segmentasi digunakan untuk klasifikasi tutupan lahan perkotaan. Segmentasi citra dengan skala yang berbeda diciptakan untuk *scene* perkotaan dengan menggunakan algoritma klasifikasi *Support Vektor Machines* (SVM) yang memberikan probabilitas

segmen dalam kelas tutupan lahan dan probabilitas yang lebih tinggi untuk kelas yang ditugaskan dengan hasil akurasi 82,1%.

Tingkat variabilitas spektral antar kelas kesamaan berbagai jenis tutupan lahan mengarah ke akurasi klasifikasi rendah bila menggunakan data berbasis pixel. Ketika segmen dan pixel yang besar digunakan akurasi klasifikasi menurun karena kurang segmentasi fitur kecil (misalnya bangunan dikelilingi oleh beton atau aspal , pohon dikelilingi oleh rumput dan bayangan). Metode klasifikasi berbasis objek dapat meningkatkan akurasi klasifikasi dengan memasukkan spektral (misalnya mean, standar deviasi) dan non - spektral (misalnya tekstur, ukuran, bentuk) sebagai informasi segmen gambar pixel super.

Oleh Brian Johnson dan Zhixiao xie tentang pendekatan multi-skala yang digunakan untuk mengklasifikasi tutupan lahan dalam gambar resolusi tinggi dari daerah perkotaan. Segmen gambar spektral, tekstur, ukuran dan bentuk informasi dari super objek dari jenis yang sama dan data set informasi super objek digunakan sebagai input data tambahan untuk klasifikasi citra dan mendapatkan tingkat akurasi 84,42 %.

Segmen gambar diklasifikasi ke dalam kelas tutupan lahan (pohon, rumput, tanah, beton, aspal, gedung, mobil, kolam renang dan bayangan), menunjukkan korespondensi yang relatif baik. Namun, karena kemiripan spektral antara bangunan dan lahan**,** kadangkala terjadi kesalahan klasifikasi misalnya lapangan bisbol memiliki bentuk yang mirip dengan lapangan sepak bola dan bangunan di wilayah studi karena bangunan dikelilingi oleh rumput yang sama dan halaman rumah yang hijau. Hal ini terjadi karena setelah segmentasi beberapa segmen yang berisi pixel bangunan juga berisi pixel tutupan lahan spektral dan non spektral menyebabkan bentuk segmen tidak akurat.

Tutupan lahan dari data resolusi tinggi menjadi sulit ketika metode klasifikasi berbasis pixel gambar tradisional digunakan karena tingginya tingkat variabilitas spektral dalam kelas tutupan lahan (misalnya: bayangan, sudut matahari, kesenjangan pohon) yang menyebabkan akurasi klasifikasi menurun. Sehingga kelas spektral dan variabilitas pixel tunggal biasanya hanya mewakili sebagian kecil objek atau target klasifikasi dalam tutupan lahan perkotaan.

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pembangunan model sebagai *prototype* untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan / klasifikasi / prediksi pada suatu objek data lain agar diketahui kelas mana objek data tersebut dalam model yang sudah disimpannya. Beberapa pendekatan algoritma klasifikasi, misalnya *decision trees, support vector machines, multilayer perceptron*, dilakukan oleh peneliti sebelumnya untuk klasifikasi tutupan lahan misalnya Annemarie Schneider, Luca

Demarchi, et al, dan DU Peijun, et al. Terjadi *overlap* terutama ketika kelas-kelas dan kriteria yang digunakan jumlahnya sangat banyak. Hal tersebut juga dapat menyebabkan meningkatnya waktu pengambilan keputusan dan jumlah *memory* yang diperlukan. *Trial* and *error* digunakan untuk memilih arsitektur yang terbaik. Pengakumulasian jumlah *error* dari setiap level dalam sebuah pohon keputusan yang besar.

Algoritma *Naive bayes* merupakan teknik prediksi berbasis probabilitas sederhana yang berdasar penerapan teorema bayes (atau aturan bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, *naïve bayes*, menggunakan "model fitur independen" dalam bayes (terutama *naïve bayes*). Maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter rata – rata dan variansi dari variabel yang dibutuhkan untuk klasifikasi. Namun, kelemahan *naïve bayes* ketika probabilitas kondisionalnya adalah nol, maka probabilitas prediksi akan bernilai nol juga dan tidak bisa mengasumsikan variable bebas. Hanya bisa digunakan untuk persoalan klasifikasi dengan *supervised learning* dan data-data kategorikal. Memerlukan pengetahuan awal untuk dapat mengambil suatu keputusan. Tingkat keberhasilan metode sangat tergantung pada pengetahuan awal yang diberikan.

Untuk membuat keputusan yang baik, harus menggunakan data yang baik pula (lengkap, benar, konsisten, terintegrasi) dengan cara mengoptimalkan akurasi klasifikasi *naïve bayes* sehingga dibutuhkan teknik *preprocessing* dalam memilih bagian dari atribut yang relevan dan memastikan data yang akan diolah pada tutupan lahan terhindar dari data kekurangan nilai atribut, data masih mengandung *error* dan *outliers*, data mengandung diskrepansi dalam *code* dan nama atau singkatnya datanya tidak konsisten. Data *preprocessing* menerangkan tipe-tipe proses yang melaksanakan data mentah untuk mempersiapkan proses prosedur dengan tujuan untuk mentransformasi data ke dalam suatu format yang lebih mudah dan efektif mendapatkan hasil yang lebih akurat, mengurangi waktu komputasi, membuat nilai data menjadi lebih kecil tanpa merubah informasi didalamnya. Terdapat beberapa metode yang berbeda digunakan untuk *preprocessing* (misalnya *Sampling, Trasnformation, Denoising, Normalization, Feature selection*) untuk penelitian tutupan lahan memilih *feature selection* yang berfungsi mengeksekusi fitur sebagai masukan untuk langkah-langkah mendapatkan akurasi yang terbaik. *Feature selection* dirancang untuk memilih fitur yang berkontribusi pada tingginya nilai akurasi klasifikasi. Walaupun semua algoritma *fetuare selection* umumnya bekerja untuk mencari nilai atribut terbaik, namun cara memilih atribut subset untuk evaluasi berbeda-beda dan untuk penelitian ini

mengunakan algoritma *forward selection* sebagai strategi memilih nilai parameter yang sesuai untuk meningkatkan akurasi.

Metode *forward selection* akan dimulai dengan seleksi variable independen yang masuk ke dalam model sesuai dengan kriteria atau prosedurnya dengan menggunakan salah satu metode pemodelan (pembangunan model linier) untuk menemukan kombinasi peubah yang "terbaik" dari suatu gugus peubah. Selain itu, *Forward selection* dapat berarti memasukkan variabel bebas yang memiliki korelasi yang paling erat dengan variabel tak bebasnya (variabel yang paling potensial untuk memiliki hubungan linier dengan Y). Kemudian secara bertahap memasukkan variabel bebas yang potensial berikutnya dan nanti akan terhenti sampai tidak ada lagi variabel bebas yang potensial untuk meningkatkan akurasi klasifikasi.

Klasifikasi tutupan lahan dari data resolusi tinggi sulit karena variabilitas spektral bayangan, air, pohon, bangunan dan sudut matahari sehingga akurasi klasifikasi menurun. Tujuan penelitian untuk meningkatkan akurasi klasifikasi dan dapat dengan mudah mengetahui batas-batas klasifikasi dengan baik penginderaan jarak jauh tutupan lahan dengan menerapkan algoritma *naïve bayes* berbasis *forward selection* melebihi akurasi dari penelitian sebelumnya karena algoritma *naïve bayes* berbasis probalistik sederhana yang berdasar pada penerapan teorema bayes dengan asumsi independensi yang kuat dan *forwad selection* melakukan seleksi variabel independen yang masuk ke dalam model dengan variabel yang mempunyai korelasi tertinggi dan *significant* dengan variabel *dependent* terus menerus sampai tidak ada lagi variabel *independent* yang *significant* sehingga pada saat *experiment* akurasi klasifikasi tutupan lahan akan meningkat dan memberikan manfaat terhadap pengembangan metode algoritma dan masukan kepada pemerintah tentang pemodelan evaluasi perencanaan pemanfaatan lahan.

## B. Rumusan Masalah

Dari uraian latar belakang masalah penelitian tentang tutupan lahan perkotaan, maka rumusan masalahnya adalah.

1. Urbanisasi yang terjadi setiap tahunnya meningkat karena pencari kerja di perkotaan dengan tujuan gaji tinggi dan perkembangan pembangunan telah merubah fungsi kawasan hijau menjadi tempat kegiatan komersial seperti pompa bensin, supermarket dan pengaliran sungai dibangun bangunan untuk pemukiman dan aktivitas lainnya. Sehingga dampak yang ditimbulkan beragam, mulai dari ketidaknyamanan penduduk akibat kurangnya sarana dan prasarana lingkungan, kesengsaraan masyarakat akibat banjir, sampai masalah sosial karena benturan berbagai kepentingan pemanfaatan lahan.

2. Tingkat spektral variabilitas antar kelas kesamaan berbagai jenis tutupan lahan mengarah ke akurasi klasifikasi rendah ketika segmen yang lebih besar digunakan sebagai unit dasar untuk klasifikasi. Penelitian tentang klasifikasi tutupan lahan perkotaan mendapatkan akurasi 84,42% yang telah dilakukan oleh Brian Johnson, Zhixiao Xie, dengan pendekatan gambar resolusi tinggi berbasis objek dengan mengunakan *algoritma random forest*.

### C. Tujuan Penelitian

Dari uraian rumusan masalah penelitian tutupan lahan hasil pengujian dan pelatihan.

1. Memberikan hasil pemodelan evaluasi perencanaan pemanfaatan lahan perkotaan dan kajian hubungan antara kenampakan fisik lahan dengan kondisi sosial yang lebih baik pada tutupan lahan perkotaan.
2. Untuk segmen spasial gambar resolusi tinggi dan tingkat spektral memperoleh batas-batas klasifikasi yang lebih baik dengan menerapkan algoritma *naïve bayes* berbasis *forward selection* sehingga dapat meningkatkan akurasi klasifikasi dan pembanding dengan *algoritma random forest* penelitian sebelumnya.

### D. Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut:

1. Memberikan rekomendasi atau masukan dalam pengambilan kebijakan terhadap permasalahan kenampakan fisik lahan perkotaan sehingga pola pemukiman perkotaan memberikan konsep yang nyaman terhadap ekosistem lingkungan di dalamnya.
2. Memberikan manfaat terhadap pengembangan metode algoritma khusunya algoritma *naïve bayes* dan *forward selection* dalam klasifikasi tutupan lahan yang lebih akurat.

## TINJAUAN PUSTAKA

### A. Penelitian Terkait

Analisis perencanaan kota berkelanjutan dan model pendekatan perkotaan yang semakin membutuhkan data klasifikasi tutupan lahan dari data penginderaan jarak jauh. Sebagai contoh, aplikasi penginderaan jarak jauh memperkirakan populasi berdasarkan jumlah tempat tinggal dari jenis perumahan yang berbeda di lingkungan perkotaan, biasanya membutuhkan ukuran pixel mulai dari 0,25 sampai 5 m menjadi identifikasi jenis struktur individu. Secara umum, setiap spektral inframerah pada rentang skala spasial harus menyediakan perbedaan

antara objek yang menarik di lingkungan sekitarnya (misalnya jalan, jalan masuk, trotoar, pohon, semak, rumput, dan kolam renang). Kebanyakan ilmuwan penginderaan jauh akan setuju bahwa resolusi radiometrik yang tinggi atau jumlah bit (misalnya 8 bit vs 16 bit) tidak akan terasa meningkatkan informasi tentang benda-benda kecil dan fitur di sekitarnya dalam data gambar beresolusi tinggi dari penginderaan jarak jauh.

Spektral dan informasi non - spektral (misalnya ukuran dan bentuk) super objek berguna untuk tujuan klasifikasi citra. Sebagai contoh, informasi spektral dari segmen terkecil mungkin berguna untuk menargetkan setiap pohon, sedangkan ukuran dan bentuk informasi super objek yang lebih besar mungkin berguna untuk memisahkan bangunan dari beton dan permukaan lainnya yang mirip dengan atap bangunan. Karena masalah utama dengan pendekatan klasifikasi skala tunggal pada semua jenis tutupan lahan tersegmentasi baik pada satu skala, keuntungan teoritis pendekatan multi skala adalah bahwa: (i) termasuk timbangan beberapa segmentasi untuk klasifikasi membuatnya lebih memungkinkan salah satu dari mereka sesuai dan baik untuk setiap jenis tutupan lahan, (ii) penggunaan variabel multi skala memberikan informasi tentang bagaimana masing-masing jenis tutupan lahan berperilaku pada banyak skala bukan hanya pada satu skala. Keuntungan lain dari pendekatan adalah bahwa hal itu kurang bergantung pada pengetahuan ahli dan kurang subjektif dari pada metode klasifikasi multiskala tradisional yang memerlukan segmen untuk diselidiki pada setiap skala segmentasi dalam menentukan skala terbaik dan mengklasifikasikan setiap jenis tutupan lahan.

## B. Landasan Teori

### 1. Algoritma *Random Forest*

Klasifikasi algoritma *random forest* adalah teknik pembelajaran *machine learning* yang menghasilkan klasifikasi dalam bentuk *forest* (hutan) dari *decision trees*. Memiliki banyak *tree*, dan setiap *tree* ditanam dengan cara yang sama. *Tree* dengan variabel x akan ditanam sejauh mungkin dengan *tree* dengan variabel y. Dan dalam perkembangannya, sejalan dengan bertambahnya data set, maka *tree* pun ikut berkembang. Penempatan *tree* yang saling berjauhan membuat apabila terdapat *tree* disekitar *tree* berarti pohon tersebut merupakan perkembangan dari *tree* x. Beberapa fungsi *learning* yang dihasilkan *random forest* digunakan strategi *ensemble "bagging"* untuk mengatasi masalah *overfitting* apabila dihadapkan data set yang kecil. Pada makalah ini *Ensemble* digunakan untuk melakukan *resample* data dengan mengklasifikasi ulang data *outlayer* sehingga merubah struktur data set yang asli. Tahapan penyusunan dan pendugaan menggunakan RF adalah:

a. Lakukan penarikan contoh acak berukuran **n** dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan *bootstrap*.
b. Dengan menggunakan contoh *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih **m** peubah penjelas secara acak, dimana **m** << **p**. Pemilah terbaik dipilih dari **m** peubah penjelas tersebut. Tahapan ini adalah tahapan *random feature selection*.
c. Ulangi langkah 1 dan 2 sebanyak **k** kali, sehingga terbentuk sebuah hutan yang terdiri atas **k** pohon.

Respons suatu pengamatan diprediksi dengan menggabungkan (*aggregating*) hasil prediksi **k** pohon. Pada masalah klasifikasi dilakukan berdasarkan *majority vote* (suara terbanyak).

### 2. Algoritma Naïve Bayes

*Naïve bayes* adalah algoritma pembelajaran induktif yang efektif dan efisien untuk *machine learning* dan datamining. Algoritma ini menggunakan probabilitas Bayesian untuk memprediksi masa depan berdasarkan pengalaman masa lalu. Bayes Teorema adalah teorema dalam statistik yang digunakan untuk menghitung probabilitas hipotesis. Sementara Bayes *classifier* optimal menghitung kemungkinan kelas dari masing-masing kelompok atribut yang ada dan menentukan mana yang kelas paling optimal.

Berdasarkan aturan Bayes, peluang E dalam bentuk satu set nilai atribut atribut (x1, x2, x3, x4 ... xn) di kelas dapat dirumuskan sebagai berikut:

$$p(\mathrm{C} \mid \mathrm{E}) = \frac{P(E \mid c)P(c)}{P(E)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Penjelasan dari formula tersebut adalah sebagai berikut:

| Paremeter | Keterangan |
|---|---|
| P(C\|E) | Probabilitas akhir bersyarat (*conditional probability*) suatu hipotesis C terjadi jika diberikan Bukti (*evidence*) E terjadi. |
| P(E\|C) | Probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis C |
| P(C) | Probabilitas awal (*priori*) hipotesis C terjadi tanpa memandang bukti apapun |
| P(E) | Probabilitas awal (*priori*) bukti E terjadi tanpa memandang hipotesis/bukti yang lain |

### 3. *Forward Selection*

*Forward selection* merupakan salah satu metode pemodelan untuk menemukan kombinasi peubah yang terbaik dari suatu gugus peubah. Dalam prosedur *forward selection,* sekalinya variabel masuk kedalam persamaan maka tidak bisa dihilangkan. Selain itu *forward selection* dapat berarti memasukkan

variabel bebas yang memiliki korelasi yang paling erat dengan variabel tak bebasnya (variabel yang paling potensial untuk memiliki hubungan linier dengan Y). kemudian secara bertahap memasukkan variabel bebas yang potensial berikutnya dan nanti akan terhenti sampai tidak ada lagi variabel bebas yang potensial.

Kelebihan *forward selection* :

a. Metode *forward selecion* merupakan alternatif untuk mengurangi kemungkinan adanya multikolinearitas dalam model yang dihasilkan.
b. Prosedur ini tidak selalu mengarahkan ke model yang terbaik, mengingat kita hanya mempertimbangkan sebuah subset kecil dari semua model–model yang mungkin. Sehingga resiko melewatkan atau kehilangan model terbaik akan bertambah seiring dengan penambahan jumlah variabel bebas.

Kelemahan *forward selection* :

a. Lama dalam perhitungan, karena harus menghitung satu – satu dari peubah yang ada, dari peubah yang memiliki F tersebar.
b. Dalam metode ini, ada kemungkinan untuk memasukkan lebih banyak variabel yang tidak begitu signifikan ke dalam model dibanding metode *backward* dan *stepwise*, karena MSE yang dihasilkan *forward* akan lebih kecil dan menyebabkan nilai Fobs besar.
c. Prosedur ini tidak selalu mengarahkan ke model yang terbaik, mengingat kita hanya mempertimbangkan sebuah subset kecil dari semua model–model yang mungkin. Sehingga resiko melewatkan atau kehilangan model terbaik akan bertambah seiring dengan penambahan jumlah variabel bebas.

Langkah – langkah *forward selection* :

a. Mulai dengan tidak ada *predictor* variabel (model hanya berisi konstanta).
b. Untuk semua *predictor* variabel tidak dalam model, pilih satu variabel dengan nilai p-value terkecil dan kurang dari taraf nyata $\alpha$.
c. Ulangi langkah b, hingga tidak terdapat *predictor* variabel yang dapat ditambahkan ke dalam model.

### 4. K-Fold Cross Validation

Merupakan pengujian standar yang dilakukan untuk memprediksi *error rate*. Setiap kelas pada data set harus diwakili dalam proporsi yang tepat antara data *training* dan data *testing.* Data dibagi secara acak pada masing-masing kelas dengan perbandingan yang sama. Untuk mengurangi bias yang disebabkan oleh sampel tertentu, seluruh proses *training* dan pengujian diulangi beberapa kali dengan sampel yang berbeda. Tingkat kesalahan pada iterasi yang berbeda akan dihitung rata-ratanya untuk menghasilkan *error rate* secara keseluruhan.

Pada setiap iterasi satu subset digunakan untuk pengujian sedangkan subset sisanya digunakan untuk pelatihan. Data awal dibagi menjadi *k* subset secara acak dengan ukuran subset yang hampir sama dengan mempertahankan perbandingan antar kelas. Pada iterasi pertama, subset satu menjadi data pengujian sedangkan subset lainnya menjadi data pelatihan. Pada iterasi kedua, subset kedua digunakan, sebagai data pengujian dan subset lainnya sebagai data pelatihan, dan seterusnya hingga seluruh subset digunakan sebagai data pengujian.

**5. Confusion Matrix**

Evaluasi model klasifikasi didasarkan pada pengujian untuk memperkirakan obyek yang benar dan salah, urutan pengujian ditabulasikan dalam *confusion matrix* dimana kelas yang diprediksi ditampilkan dibagian atas matriks dan kelas yang diamati disisi kiri. Setiap sel berisi angka yang menunjukkan berapa banyak kasus yang sebenarn ya dari kelas yang diamati untuk diprediksi.

**Tabel 2.** *Confusion matrix*

| | | Predicted Class | |
|---|---|---|---|
| | | Class=Yes/Positive | Class=No/Negative |
| Oberved Class | Class= Yes/Positive | A | B |
| | | (TP=True Positive) | (FN=False Negative) |
| | Class=No/Negative | C | D |
| | | (FP=False Positive) | (TN=True Negative) |

Keterangan :

TP = Prediksi positif yang positif
FN = Prediksi positif yang negatif
FP = Prediksi negatif yang positif
TN = Prediksi negatif yang negatif

*Precision* didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. *Precision* merupakan probabilitas bahwa sebuah item yang dipilih adalah relevan. Dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan itu. *Precision* dihitung dengan rumus:

$$\text{Precision} = \left( \frac{A}{A+B} \right) = \frac{TP}{TP+FP} \quad \text{.....} \quad (2)$$

Sedangkan *Recall* didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. *Recall* merupakan probabilitas bahwa suatu item yang relevan akan dipilih. *Recall* dapat dihitung dengan jumlah rekomendasi yang relevan yang dipilih oleh *user* dibagi dengan jumlah semua rekomendasi yang relevan baik dipilih maupun rekomendasi yang tidak terpilih. *Recall* dapat dihitung dengan rumus:

$$\mathbf{Recall} = \left( \frac{A}{A+C} \right) = \frac{TP}{TP+FN} \text{ ……..…….…… (3)}$$

*Accuracy* adalah representasi dari penggabungan antara *Precision* dan *Recall* dapat dihitung menggunakan rumus dibawah ini :

$$Accuracy = \left( \frac{A+D}{A+B+C+D} \right) = \frac{TP+TN}{TP+TN+FP+FN} \text{ …..……..……...(4)}$$

Merupakan tingkat akurasi terhadap sistem dalam memberikan rekomendasi yang diinginkan. Sistem akan dianggap baik jika memiliki tingkat akurasi yang tinggi.

## 6. Klasifikasi tutupan lahan perkotaan menggunakan *naïve bayes* berbasis *forward selection*

Permasalahan klasifikasi tutupan lahan perkotaan dari data resolusi tinggi pada penelitian sebelumnya terletak pada spektral variabilitas antar kelas kesamaan berbagai jenis tutupan lahan mengarah ke akurasi klasifikasi rendah. Dalam penelitian tutupan lahan mendapatkan akurasi klasifikasi 84,42% dengan berbasis objek dipengaruhi oleh kualitas gambar dan klasifikasi gambar pixel sebagai unit dasar untuk klasifikasi super objek tutupan lahan dan klasifikasi tutupan lahan belum memberikan batas-batas klasifikasi antar objek dengan objek yang lainnya kejelasan yang baik terhadap target klasifikasi tutupan lahan perkotaan.

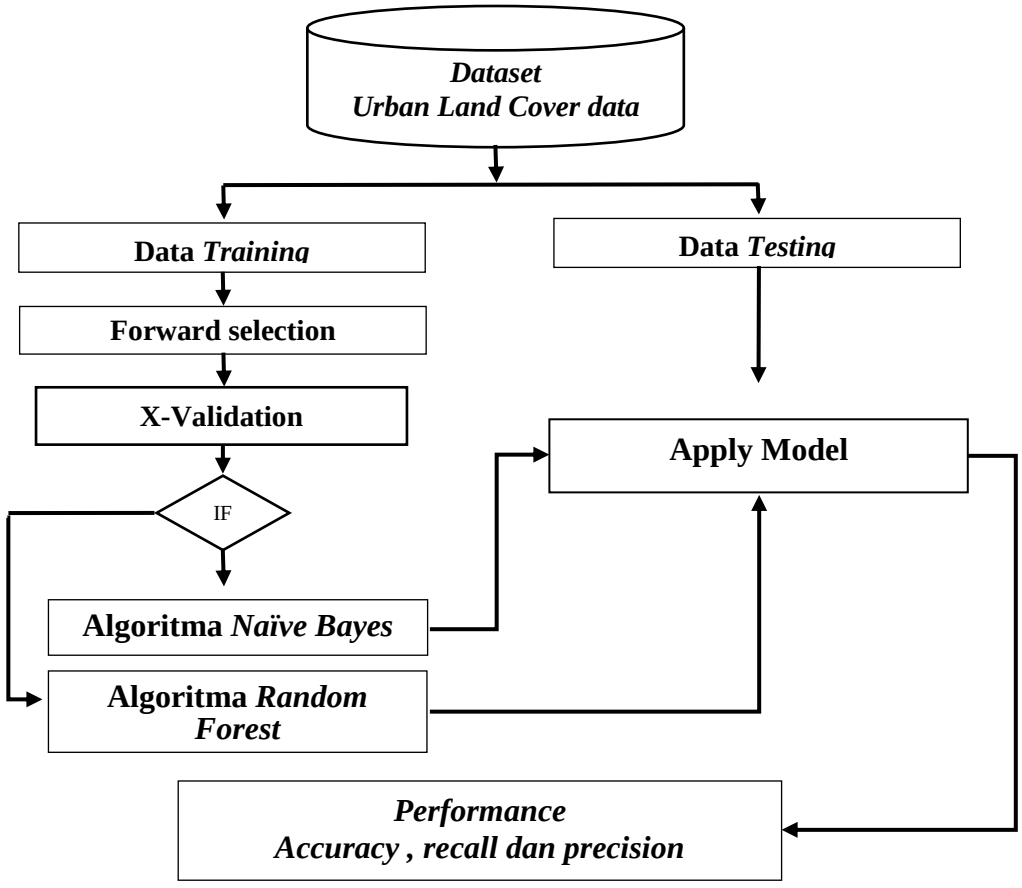
Dengan menerapkan *preprocessing* maka memberikan hasil yang lebih akurat, pengurangan waktu komputasi untuk *large scale problem*, membantu nilai data menjadi kecil tanpa merubah informasi yang dikandungnya dan metode *Feature selection* dengan algoritma *forward selection* sebagai suatu pemodelan untuk menemukan peubah yang tebaik dari suatu gugus peubah. Dalam prosedur *forward selection* dapat memasukkan variabel bebas yang memiliki korelasi yang paling erat dengan variabel tidak bebasnya dan secara bertahap memasukkan variabel bebas yang potensial berikutnya dan nanti akan berhenti sampai tidak ada lagi variabel yang potensial. Maka setelah pemilihan pemodelan dengan *forward selection* dengan memilih variabel yang potensial untuk mengoptimalkan akurasi yang lebih akurat tehadap proses selanjutnya.

Data pelatihan mengunakan pengujian *x-validation* sebagai model untuk menghindari *overlapping* pada data testing sehingga bisa memvalidasi model pelatihan valid atau tidak dengan menggunakan *10-fold cross-validation* maka akan memberikan hasil terbaik sehingga proses pada algoritma *naïve bayes* yang memiliki probabilitas dan teorema Bayesian dengan asumsi bahwa setiap variabel X bersifat bebas (*independence*). Dengan kata lain, *naïve bayesian classifier* mengansumsikan bahwa keberadaan sebuah atribut (variabel) tidak ada kaitannya dengan keberadaan atribut (variabel) yang lain. Baik untuk titik noise yang di isolasi, misalkan titik yang dirata-ratakan ketika mengestimasi peluang bersyarat

data. Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata-rata dan variansi dari variabel) yang dibutuhkan untuk klasifikasi.Menangani nilai yang hilang dengan mengabaikan data latih selama perhitungan estimasi peluang dan kokoh terhadap atribut yang tidak relevan, cepat dan efisiensi waktu proses. Maka penelitian klasifikasi tutupan lahan perkotaan menggunakan *naïve bayes* berbasis *forward selection* pasti melebihi akurasi dari penelitian sebelumnya dan memberikan batas-batas klasifikasi tutupan lahan yang lebih jelas dan akurat. Untuk memastikan apakah kinerja klasifikasi penelitian lebih baik dari penelitian sebelumnya dan berapa tingkat presentase peningkatan yang didapatkan maka peneliti menggunakan *confusion matrix* sebagai dasar untuk perhitungan *Accuracy*, *Recall*, *Precision.*

## METODE PENELITIAN

Pada bagian ini dijelaskan tentang langkah-langkah eksperimen meliputi cara pemilihan arsitektur yang tepat dari model atau metode yang diusulkan sehingga didapatkan hasil yang dapat membuktikan bahwa metode yang digunakan adalah tepat, eksperimen dilakukan dengan langkah sebagai berikut :




**Gambar 1.** Diagram alir metode

## HASIL DAN PEMBAHASAN

### A. Hasil Pengumpulan Data

Hasil pengumpulan data diambil dari penelitian sebelumnya UCI repository mengenai *urban land cover* dataset tahun 27 Maret 2014 [4],[5]. Data *urban landcover* terbagi dua yaitu data *training* dan data *testing*. Data *training* berjumlah 168 *records* dan *testing* berjumlah 507 *records* dengan 9 jenis tutupan lahan dan atribut berjumlah 148 akan tetapi atribut yang di proses hanya 147 karena 1 atribut merupakan variabel spesial.

### B. Hasil Naïve Bayes Berbasis Forward Selection Hasil

Percobaan dengan algoritma naïve bayes berbasis forward selection. Memberikan 14 atribut yang terbaik yang bernilai satu dan 133 atribut dibuang yang bernilai nol. Atributnya adalah Mean_G, Mean_R, SD_NIR, Assym, NDVI, NDVI_40, Dens_60, Round_100, Bright_100, Rect_100, NDVI_100, Compact_120, SD_R_120 dan 168 record data yang mempunyai nilai optimal dalam mempengaruhi akurasi klasifikasi tutupan lahan perkotaan.

Tabel 2. Hasil *naïve bayes* berbasis *forward selection*

| **No** | ***Attribute*** | ***Weight*** |
|---|---|---|
| **1** | *Mean_G* | 1 |
| **2** | *Mean_R* | 1 |
| **3** | *SD_NIR* | 1 |
| **4** | *Assym* | 1 |
| **5** | *NDVI* | 1 |
| **6** | *NDVI_40* | 1 |
| **7** | *Dens_60* | 1 |
| **8** | *Round_100* | 1 |
| **9** | *Bright_100* | 1 |
| **10** | *Rect_100* | 1 |
| **11** | *NDVI_100* | 1 |
| **12** | *Compact_120* | 1 |
| **13** | *SD_R_120* | 1 |
| **14** | *Assym_140* | 1 |
| **15** | *BrdIndx* | 0 |
| **16** | *Area* | 0 |
| **17** | *Round* | 0 |
| **18** | ... | … |
| **147** | *GLCM3_140* | 0 |

### C. Hasil *Random Forest* Berbasis *Forward Selection*

Hasil percobaan dengan algoritma *random forest* berbasis *forward selection* memberikan 2 atribut yang terbaik yang bernilai satu dan 145 atribut dibuang yang bernilai nol. Atributnya adalah NDVI, Mean_R_40 dan 168 record data yang mempunyai nilai optimal dalam mempengaruhi akurasi klasifikasi tutupan lahan perkotaan.

**Tabel 3.** Hasil *random forest* berbasis *forward selection*

| No | *Attribute* | *Weight* |
|---|---|---|
| **1** | *NDVI* | 1 |
| **2** | *Mean_R_40* | 1 |
| **3** | *BrdIndx* | 0 |
| **4** | *Area* | 0 |
| **5** | *Round* | 0 |
| **6** | Bright | 0 |
| **…** | … | … |
| **147** | *GLCM3_140* | 0 |

## PEMBAHASAN

### A. Implementasi perhitugan naïve bayes

Proses perhitungan selanjutnya menggunakan rumus densitas gauss (distribusi normal) karena data training bertipe data kontinyu. Untuk mengetahui apakah hasil klasifikasi benar atau salah, dimana data training mempunyai 9 target jenis kelas sesuai dengan records yang di testing. Hasil dari perhitungan densitas gauss yang telah selesai kemudian dibandingkan. Nilai yang dipilih adalah nilai yang tertinggi pada kelas target klasifikasi tutupan lahan perkotaan.

$$N(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Π = 3,14159
e = 2,71828
μ = *Mean*, menyatakan rata-rata dari seluruh atribut
σ = *Standar deviasi* dari nilai-nilai distribusi variabel
x = Nilai dari distribusi variabel

Adapun tahapan untuk penjabaran rumus densitas gauss dengan menggunakan bantuan perhitungan excel sebagai berikut.

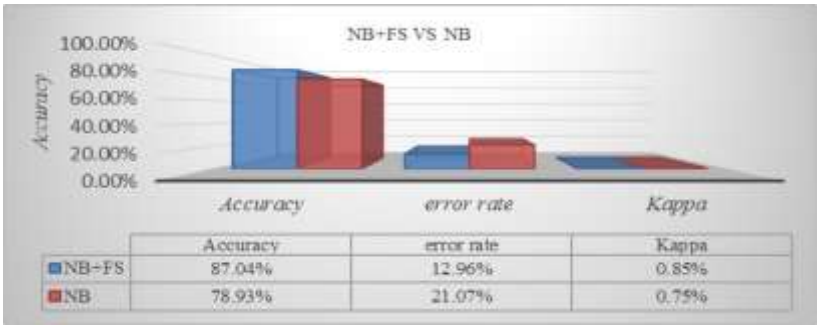
1. *Mean*, menggunakan fungsi ***average*** untuk menghitung nilai rata-rata dari sekumpulan data.
   *=AVERAGE(VALUE1;VALUE2;……)*

2. *Standard deviasi*, menggunakan fungsi ***stdev*** untuk memperkirakan standar devias berdasarkan pada suatu sampel tertentu
   *=STDEV(NUMBER1;NUMBER2;….)*
3. Rumus densitas gauss dengan menggunakan fungsi ***normdist*** untuk menghasilkan distribusi kumulatif normal untuk rata-rata dan standar deviasi tertentu.
   *=NORMDIST(X;MEAN;STANDR_DEV;CUMULATIVE)*

## B. Perbandingan Algoritma

### 1. Naïve bayes berbasis forward selection dengan naïve bayes

Untuk hasil proses perhitungan algoritma naïve bayes berbasis forward selection pada data *training* dan *testing urban land cover* dengan tingkat akurasi 87,04%, kappa 0,85% dan *error rate* 12,96% sedangkan untuk algoritma naïve bayes akurasinya 78,93%, kappa 0,75% dan *error rate* 21,07% . Sehingga untuk penelitian dengan kasus jumlah atribut yang banyak pada kasus *urban land cover* lebih baik dengan menggunakan *algoritma forward selection*.

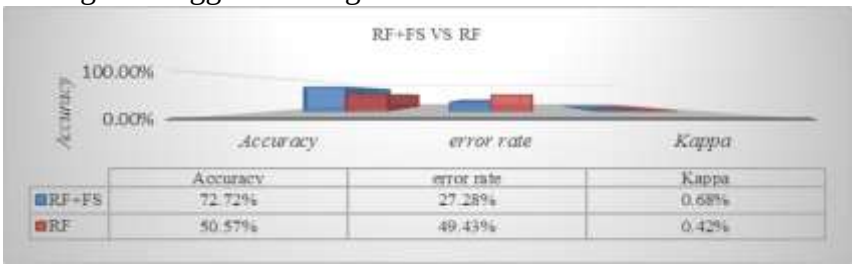



| | Accuracy | error rate | Kappa |
|---|---|---|---|
| NB+FS | 87.04% | 12.96% | 0.85% |
| NB | 78.93% | 21.07% | 0.75% |

**Gambar 2.** Diagram balok NB+FS VS NB

### 2. Random forest berbasis forward selection dengan random forest

Untuk hasil proses perhitungan algoritma random forest berbasis forward selection pada data training dan testing urban land cover dengan tingkat akurasi 72,72%, kappa 0,68% dan error rate 27,28% sedangkan untuk algoritma random forest akurasinya 50,57%, kappa 0,42% dan error rate 49,43% . Sehingga untuk penelitian dengan kasus jumlah atribut yang banyak pada kasus urban land cover lebih baik dengan menggunakan algoritma forward selection.

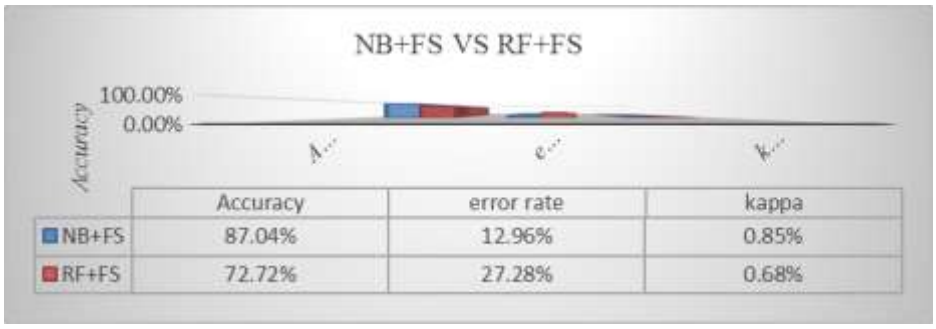



| | Accuracy | error rate | Kappa |
|---|---|---|---|
| RF+FS | 72.72% | 27.28% | 0.68% |
| RF | 50.57% | 49.43% | 0.42% |

**Gambar 3.** Diagram balok RF+FS VS RF

### 3. Naïve bayes berbasis forward selection VS random forest berbasis forward selection

Untuk hasil proses perhitungan algoritma naïve bayes berbasis forward selection dengan akurasi 87,04%, kappa 12,96% dan error rate 0,85% dan random forest berbasis forward selection akurasi 72,72%, kappa 27,28% error rate 0.68% pada data training dan testing urban land cover. Dapat di analisis bahwa semakin kecil nilai error rate tingkat akurasi semakin baik dan sebaliknya dengan kappa nilai yang tinggi akan memberikan akurasi yang lebih baik juga. Sehingga pada penelitian dengan studi kasus urban land cover lebih baik dengan algoritma naïve bayes berbasis forward selection. Indenpenden dalam memproses target klasifikasi dan hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata– rata dan variansi dari variabel) yang dibutuhkan untuk klasifikasi sehingga Cepat dan efisien. Ditambah dengan metode forward selection berfungsi mencari atribut terbaik untuk mengoptimalkan akurasi pada klasifikasi 9 jenis kelas tutupan lahan perkotaan.




**Gambar 4.** Diagram balok NB+FS VS RF+FS

## KESIMPULAN DAN SARAN

### A. Kesimpulan

Hasil percobaan dan pengujian terhadap data *urban land cover* berjumlah 675 *records* yang terbagi menjadi data *training* berjumlah 168 *records* dan data *testing* berjumlah 507 *records* dengan jumlah masing-masing variabel 148 akan tetapi yang di proses 147 karena 1 atribut variabel spesial. Data *training* digunakan untuk membentuk model klasifikasi setelah *forward selection* memproses atribut untuk meningkatkan akurasi klasifikasi dan dilanjutkan dengan pengujian *algoritma naiva bayes* atau *random forest* dengan melakukan pengujian terhadap model data *training* terhadap data *testing*.

Forward selection untuk algoritma naïve bayes sebanyak 14 atribut yaitu (Mean_G, Mean_R, SD_NIR, Assym, NDVI, NDVI_40, Dens_60, Round_100, Bright_100, Rect_100, NDVI_100, Compact_120, SD_R_120, Assym_140). Selanjutnya X-Validation membagi data dengan jumlah yang sama menjadi dua

sebanyak K kali yang berulang hanya satu kali dan dilanjutkan dengan pembentukan model klasifikasi dengan menggunakan algoritma naïve bayes dalam pengujian model klasifikasi pada data testing. Mendapatkan akurasi 87,04% lebih baik dari penelitian sebelumnya dengan akurasi 84,42% dan meningkat 2,62% dengan jumlah error rate 12,95%.

*Forward selection* untuk *algoritma random forest* sebanyak 2 atribut yaitu (*NDVI, Mean_R_40*). Selanjutnya X-Validation membagi data dengan jumlah yang sama menjadi dua sebanyak K kali yang berulang hanya satu kali dan dilanjutkan dengan pembentukan model klasifikasi dengan menggunakan *algoritma random forest* dan dilanjutkan dengan pengujian model klasifikasi pada data *testing*. Mendapatkan akurasi 72,72% dengan jumlah *error rate* 27,27%. Lebih akurat dengan pengujian naïve bayes berbasis forward selection 87,04%.

Dari hasil proses mendapatkan uji model terbaik dengan menggunakan algoritma *naïve bayes* berbasis *forward selection* 87,04% sebagai pembanding agoritma *random forest* 72,72 %. Memberikan hasil klasifikasi yang lebih akurat dan tepat antara kelas yang satu dengan yang lainnya, sehingga memberikan rekomendasi dalam pengambilan kebijakan terhadap perencanaan tutupan lahan perkotaan serta memberikan konsep yang nyaman terhadap ekosistem lingkungan didalamnya.

## B. Saran

Untuk penelitian selanjutnya disarankan agar menggunakan jumlah *records* data training lebih banyak dengan jumlah *records* data testing dan algoritma naïve bayes dengan independen dalam mengklasifikasi target, baik digunakan pada penelitian yang mempunyai target klasifikasi yang banyak.

## DAFTAR RUJUKAN

Alfonso Ibáñez, Concha Bielza, Pedro Larrañaga, *Cost Sensitive Selective Naive Bayes Classifiers For Predicting The Increase Of Then Index For Scientific* Journals, 2014.

Blaschke. T, *Object Based Image Analysis For Remote Sensing*, Austria : Z_Gis Centre For Geoinformatics And Department For Geography And Geology, University Of Salzburg, Hellbrunner Street 34, A-5020 Salzburg, 2010.

Brian A. Johnson, *High-Resolution Urban Land-Cover Classification Using A Competitive Multi-Scale Object-Based Approach*, Vol.4, No. 2 February 2013, 131-140.

Brian Van Essen, Chris Macaraeg, Maya Gokhale And Ryan Prenger, *Accelerating A Random Forest Classifier: Multi-Core, Gp-Gpu, Or Fpga*, Lawrence Livermore National Laboratory, Livermore, Ca 94550, 2012.

Eko Prasetyo, *Datamining Konsep Dan Aplikasi Menggunakan Matlab*, Pertama, Andi Offset, Cv Andi Offset, 2012, 45-59.

Hongsheng Zhang, Yuanzhi Zhang, Hui Lin, *Urban Land Cover Mapping Using Random Forest Combined With Optical And Sar Data*, The Chinese University Of Hong Kong, Shatin, New Territories, Hong Kong, 2012.

Ian H. Witten, Eibe Frank, Mark A. Hall, *Data Mining: Practical Machine Learning Tools And Techniques*.-3rd Ed. Library Of Congress Cataloging - In- Publication Data, 2011.

Intan Martina Md Ghani And Sabri Ahmad, *Comparison Methods Of Multiple Linear Regressions In Fish Landing*, Australian Journal Of Basic And Applied Sciences, 5(1): 25-30, 2011.

Klaus Desmet, Esteban Rossi-Hansberg, *Analyzing Urban Systems Have Megacities Become Too Large*, The World Bank Sustainable Development Network Urban And Disaster Risk Management Department May 2014.

Luca Demarchi, Frank Canters, Claude Cariou, Giorgio Licciardi, Jonathan Cheung-Wai Chan, *Assessing The Performance Of Two Unsupervised Dimensionality Reduction Techniques On Hyperspectral Apex Data For High Resolution Urban Land-Cover Mapping*, Isprs Journal Of Photogrammetry And Remote Sensing 87, 2014, 166-179.

Marcin Czajkowski, Marek Grze, Marek Kretowski, *Multi-Test Decision Tree And Its Application To Microarray Data Classification*, Artificial Intelligence In Medicine 61, 2014, 35–44.

Matthias Reifa, Faisal Shafait, *Efficient Feature Size Reduction Via Predictive Forward Selection*, Pattern Recognition 47, 2014, 1664-1673.

Nariswari Karina Dewi, Utami Dyah Syafitri, Soni Yadi Mulyadi, *Penerapan Metode Random Forest Dalam Driver Analiysis, Forum Statistika Dan Komputasi*, April 2011 P : 35-43.

Rong Jia, Li Gang, Chen Yi-Ping P., *Accoustic Feature Selection For Automatic Emotion Recognition From Speech, Information Processing And Management* 45 (2009) 315–328, 2009.

W. Myint., Seo, Gober., Patricia, Brazel., Anthony, Grossman-Clarke., Susanne, Weng., Qihao . *Per-Pixel Vs. Object-Based Classification Of Urban Land Cover Extraction Using Highspatial Resolution Imagery*, United States, Germany, 2011.

Wahyu S. J. Saputra, Arif Rahman Sujatmika, Agus Zainal Arifin, *Seleksi Fitur Menggunakan Random Forest Dan Neural Network, Electronic Engineering Polytechnic Institute Of Surabaya* (Eepis), Indonesia, October 26, 2011.