Analisis komparatif teknik penambangan data prediktif

Artikel *di* International Journal of Rapid Manufacturing · Januari 2009
doi: 10.1504/IJRAPIDM.2009.029380

Analisis Komparatif Teknik Penambangan Data Prediktif

Xueping Li *, Godswill Chukwugozie Nsofor, Lagu Laigang

Departemen Teknik Industri dan Informatika,
Universitas Tennessee,
Knoxville, TN 37996, Amerika Serikat
Telepon: +1-865-974-7648
Faks: +1-865-974-0588
Surel: Xueping.Li@utk.edu, cnsofor@utk.edu , SongLG@utk.edu
*Penulis korespondensi

Abstrak: Tidak sepele untuk memilih teknik prediksi yang sesuai dari berbagai teknik yang ada untuk kumpulan data, karena evaluasi kompetitif teknik (mengantongi, meningkatkan, menumpuk, dan meta-learning) dapat memakan waktu. Makalah ini membandingkan lima teknik penambangan data prediktif pada empat kumpulan data unik yang memiliki kombinasi karakteristik berikut: beberapa variabel prediktor, banyak variabel prediktor, variabel yang sangat kolinier, variabel yang sangat berlebihan dan keberadaan pencilan. Teknik penambangan data yang berbeda, termasuk regresi linier berganda (MLR), regresi komponen utama (PCR), regresi punggungan, kuadrat terkecil parsial (PLS) dan kuadrat terkecil parsial non-linier (NLPLS), diterapkan pada masing-masing himpunan data. Perbandingan didasarkan pada kriteria yang berbeda: R-square, R-square disesuaikan, mean square error (MSE), mean absolute error (MAE), koefisien efisiensi, nomor kondisi (CN) dan jumlah variabel fitur yang termasuk dalam model. Keuntungan dan kerugian dari teknik dibahas dan diringkas.

Kata kunci: Data Mining, Analisis Statistik, Penemuan Pengetahuan

Referensi untuk makalah ini harus dibuat sebagai berikut: Nsofor, G.., Li, X., Song, L. (2009) 'Analisis Komparatif Teknik Penambangan Data Prediktif', *IJRAPIDM*, Vol. 1, No. 2, hlm. 150-172.

Xueping Li adalah Asisten Profesor Teknik Industri dan Informasi dan Direktur Laboratorium Sistem Teknik Informasi Cerdas (IIESL) di University of Tennessee - Knoxville. Beliau meraih gelar Ph.D. dari Arizona State University. Bidang penelitiannya meliputi jaminan informasi, penjadwalan, penambangan web, manajemen rantai pasokan, lean manufacturing, dan jaringan sensor. Dia adalah anggota IIE, IEEE dan INFORMS.

Godswill Chukwugozie Nsofor memegang gelar MS dari University of Tennessee, Knoxville. Bidang minat penelitiannya meliputi penambangan data, manajemen kualitas, dan keandalan sistem. Dia telah aktif terlibat dalam penelitian yang disponsori dan proyek-proyek industri.

Laigang Song adalah Ph.D. kandidat di Departemen Teknik Industri dan Informasi di University of Tennessee, Knoxville. Bidang minat penelitiannya meliputi penambangan web dan intelijen web, simulasi, dan jaminan informasi.

1. Pendahuluan

Dalam beberapa tahun terakhir, data mining telah menjadi salah satu alat yang paling berharga untuk memanipulasi data dan membangun pola informasi yang berguna untuk pengambilan keputusan. Dengan terobosan dalam teknologi pengumpulan data dan munculnya Internet, orang dibanjiri oleh data (Lyman, 2003; Han dan Kamber, 2006). Pertumbuhan data dan basis data yang luar biasa ini telah melahirkan kebutuhan mendesak akan teknik dan alat baru yang dapat secara cerdas dan otomatis mengubah data menjadi informasi dan pengetahuan yang berguna. Kebutuhan termasuk peringkasan otomatis

data, ekstraksi "esensi" informasi yang disimpan, penemuan pola dalam data mentah, dan akhirnya prediksi masa depan dari sampel sebelumnya, yaitu data mining prediktif (Usama, 1996; Indurkhya, 1998; Berson, 1999; Witten dan Frank, 2000). Berbagai teknik data mining prediktif telah dikembangkan. Namun, tidak sepele untuk memilih teknik prediksi yang tepat untuk dataset tertentu karena evaluasi kompetitif teknik (mengantongi, meningkatkan, menumpuk, dan meta-learning) dapat memakan waktu. Makalah ini membandingkan lima teknik penambangan data prediktif pada empat kumpulan data unik dan memberikan pedoman untuk mengidentifikasi teknik terbaik untuk kumpulan data tertentu dan menggunakannya secara langsung alih-alih metode coba-coba biasa dalam analisis penambangan data prediktif.

Data mining adalah proses analitik untuk mencari pola yang konsisten atau hubungan sistematis antara variabel dan kemudian memvalidasi temuan dengan menerapkan pola yang terdeteksi ke dataset baru (Berry, 2000; Giudici, 2003; Mitra dan Acharya, 2006). Akar penambangan data berasal dari tiga bidang: statistik klasik, kecerdasan buatan (AI), dan pembelajaran mesin (Han dan Kamber, 2006). Pregibon (1997) menggambarkan data mining sebagai perpaduan statistik, kecerdasan buatan, dan penelitian basis data, dan menyatakan

bahwa itu bukan bidang yang menarik bagi banyak orang sampai saat ini. Data mining terutama dapat dibagi menjadi dua tugas (Usama, 1996): tugas prediktif dan tugas deskriptif. Penambangan data prediktif adalah jenis yang paling umum dan memiliki aplikasi paling banyak untuk bisnis. Nugget informasi yang tak terduga dapat mengarah ke pasar baru, cara baru untuk menjangkau pelanggan dan cara-cara baru dalam melakukan bisnis (Betts, 2003).

Banyak kerangka kerja telah diusulkan untuk membangun model penambangan data (Jolliffe, 1986; Chapman, 2000; Lessmann dan Vob, 2009). Proyek penambangan data yang kompleks memerlukan upaya terkoordinasi dari berbagai ahli, pemangku kepentingan, atau departemen di seluruh organisasi. Oleh karena itu, perlu untuk mengidentifikasi kerangka kerja yang dapat berfungsi sebagai cetak biru untuk pengumpulan data, analisis, diseminasi hasil dan implementasi organisasi. RENYAH

DMAIC dan SEMMA adalah tiga kerangka kerja yang paling umum. CRISP adalah singkatan dari *Cross-Industrial Standard Process* for data mining, yang diusulkan oleh konsorsium perusahaan Eropa pada pertengahan 1990-an (Chapman, 2000). DMAIC menunjukkan proses *Define-Measure-Analyze-Improve-Control* , metodologi six-sigma untuk

menghilangkan cacat, limbah, atau masalah kontrol kualitas dari semua jenis di bidang manufaktur, pemberian layanan, manajemen dan kegiatan bisnis lainnya (Pyzdek, 2003). SEMMA adalah untuk *Sample-Explore-Modify-Model-Assess*, yang merupakan kerangka kerja lain yang mirip dengan six-sigma dan diusulkan oleh SAS Institute (Jolliffe, 1986).

Makalah ini mengambil jalan yang berbeda untuk membantu pemodelan penambangan data, terutama dalam domain penambangan data prediktif. Kami membandingkan lima teknik penambangan data prediktif pada empat kumpulan data unik yang memiliki kombinasi

karakteristik berikut: beberapa variabel prediktor, variabel yang sangat kolinier, variabel yang sangat berlebihan dan keberadaan pencilan.

Pengukuran kinerja atau kriteria evaluasi model yang berbeda digunakan untuk mengidentifikasi teknik terbaik untuk kumpulan data dengan karakteristik khusus. Teknik penambangan data prediktif ini memiliki aplikasi luas dalam layanan keuangan, telekomunikasi, ritel, perawatan kesehatan, farmasi, dan banyak bidang lainnya.

Sisa makalah disusun sebagai berikut. Bagian 2 secara singkat memperkenalkan teknik persiapan data. Lima teknik penambangan data prediktif umum dibahas di bagian 3. Kumpulan data dan kriteria yang dipilih untuk perbandingan model diberikan di bagian 4. Pada bagian 5, kami menerapkan teknik penambangan data prediktif untuk setiap dataset, membandingkan kinerja, dan mendiskusikan keunggulan masing-masing teknik dibandingkan yang lain. Bagian 6 menyimpulkan penelitian ini.

2. Persiapan Data

Masalah kualitas data yang buruk biasanya dihadapi dalam proses memperoleh data dan dapat menghambat tujuan penambangan data prediktif. Langkah tepat yang diambil dalam akuisisi dan penanganan data akan membantu pemodel dalam mendapatkan hasil yang andal dan prediksi yang lebih baik. Akuisisi dan penanganan data memiliki begitu banyak langkah dan merupakan topik besar tersendiri; Tetapi untuk tujuan pekerjaan ini, hanya topik-topik yang relevan dengan penambangan data prediktif yang disebutkan secara singkat. Secara khusus, kita akan membahas penyaringan dan penghalusan data, analisis komponen utama (PCA), dan analisis koefisien korelasi (CCA).

2.1. Pemfilteran dan Penghalusan Data

Terkadang selama preprocessing data, mungkin ada kebutuhan untuk memperlancar data untuk dihilangkan

outlier dan kebisingan. Ini sangat tergantung, bagaimanapun, pada definisi pemodel tentang "kebisingan." Untuk memperlancar himpunan data, pemfilteran digunakan. Filter adalah perangkat yang secara selektif melewati beberapa nilai data dan menahan beberapa tergantung pada batasan pemodel (Pyle, 1999). Ada beberapa cara untuk memfilter data.

- Moving average: Metode ini digunakan untuk penyaringan tujuan umum untuk frekuensi tinggi dan rendah (Pyle, 1999; Gencay, 2002; Olaf, 2002).
- Pemfilteran median: Teknik ini biasanya digunakan untuk kumpulan data deret waktu untuk menghapus pencilan atau titik data yang buruk. Ini adalah metode penyaringan nonlinier dan cenderung mempertahankan fitur data (Olaf, 2002; Kassama, 1985).
- Peak-valley mean (PVM): Dibutuhkan rata-rata puncak dan lembah terakhir sebagai perkiraan bentuk gelombang yang mendasarinya. Puncak adalah nilai yang lebih tinggi dari nilai sebelumnya dan berikutnya dan lembah adalah nilai yang lebih rendah dari yang terakhir dan yang berikutnya dalam seri (Pyle, 1999; Olaf, 2002).
- Normalisasi / standardisasi: Metode ini mengubah nilai instance dengan cara yang spesifik dan jelas untuk mengekspos konten informasi (Pyle, 1999; Olaf, 2002). Nilai yang diukur dapat diskalakan ke rentang dari -1 hingga +1. Metode ini mencakup teknik normalisasi desimal dan standar deviasi. Dalam makalah ini, yang terakhir digunakan. Itu membuat kumpulan data memiliki rata-rata kolom nol dan varians kolom satu. Setiap catatan memiliki kesempatan yang sama untuk muncul dalam model.

$$MCi = xi$$

Pemusatan rata-rata kolom; *MCi* memiliki kolom rata-rata nol.

$$MCi\ SC_i$$

Penskalaan kolom *SCi* memiliki rata-rata kolom nol dan varians 1.

- Memperbaiki nilai yang hilang dan kosong: Sebagian besar algoritma penambangan data tidak dapat menangani nilai yang hilang dan kosong dengan benar. Nilai-nilai ini diharapkan akan dihapus sebelum proses penambangan data. Ada banyak cara untuk memperbaiki nilai yang hilang dan kosong. Salah satu caranya adalah dengan mengganti nilai yang hilang dengan nilai rata-rata.

2.2. Analisis Komponen Utama (PCA)

PCA (Jolliffe, 1986) adalah metode parametrik tanpa pengawasan yang mengurangi dan mengklasifikasikan jumlah variabel dengan mengekstraksi variabel dengan persentase varians yang lebih tinggi dalam data (disebut komponen utama, PC) tanpa kehilangan informasi yang signifikan. PCA mengubah satu set variabel berkorelasi menjadi satu set variabel baru yang tidak berkorelasi dan memungkinkan analis untuk menggunakan sejumlah variabel yang berkurang, meskipun dengan beberapa kehilangan informasi. Karena kebisingan biasanya lebih lemah dari pola, pengurangan dimensi dapat menghilangkan banyak kebisingan.

PCA hanya sesuai dalam kasus-kasus bahwa variabel diukur dalam unit yang sama atau setidaknya dalam unit yang sebanding, dan variansnya kira-kira berukuran sama. Jika variabel tidak diukur dalam unit yang sebanding, mereka harus distandarisasi atau dinormalisasi sebelum analisis PCA. Standardisasi akan memberikan semua variabel bobot yang sama dan menghilangkan pengaruh satu variabel atas yang lain. Untuk hampir semua situasi analisis data, PCA dapat direkomendasikan sebagai langkah pertama (Johnson, 1998). Dalam proses

melakukan ini, variabel baru (faktor) yang disebut komponen utama dapat dibentuk dalam urutan kepentingan yang menurun, sehingga (1) mereka tidak berkorelasi dan ortogonal, (2) komponen utama pertama menyumbang sebanyak mungkin variabilitas dalam data, dan (3) setiap komponen berikutnya menyumbang sebanyak mungkin variabilitas yang tersisa. PCA dihitung menggunakan dekomposisi nilai tunggal (SVD) (Jolliffe, 1986), yang merupakan metode yang menguraikan matriks *X* menjadi matriks kesatuan *U*, dan matriks diagonal S yang memiliki ukuran yang sama dengan *X*, dan matriks persegi V lainnya yang memiliki ukuran jumlah kolom *X*.

$$X \; U \; S \; V \square \; \square \; \square^{T}$$

Di mana U adalah matriks ortonormal (*m m*□), *S* adalah matriks diagonal (*m n*□), *n* adalah pangkat *X* dan diagonal dikenal sebagai nilai tunggal dan menurun secara monoton. Ketika nilai-nilai tunggal ini dikuadratkan, mereka mewakili nilai-nilai eigen.

V = Matriks ortonormal (*n n*□) dari vektor eigen, yang disebut vektor pemuatan atau komponen utama:

$$z \; U \; S \square \; \square$$

atau

$$z \; X \; V \square \; \square$$

di mana ***Z*** adalah *matriks m n*□ yang disebut matriks skor, ***X*** adalah *matriks m n*□ dari data asli, dan ***V*** adalah *matriks transformasi n n*□ yang disebut matriks pemuatan. *m* adalah dimensi ruang asli, *n* adalah dimensi ruang PC yang berkurang, dan *m* adalah jumlah pengamatan di kedua ruang.

2.3. Analisis Koefisien Korelasi (CCA)

CCA (Cohen, 2003) menilai ketergantungan linier antara dua variabel acak. CCA sama dengan kovarians dibagi dengan kovarians terbesar yang mungkin dan memiliki kisaran dari -1 hingga +1. Koefisien korelasi negatif berarti hubungan itu tidak langsung, atau, ketika seseorang naik, yang lain cenderung turun. Koefisien korelasi positif menunjukkan hubungan proporsional langsung. Koefisien korelasi dapat ditunjukkan dengan persamaan hubungan kovarians:

Jika matriks kovarians diberikan oleh

$$\begin{bmatrix} \sigma_x^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_y^2 \end{bmatrix}$$

Koefisien korelasi adalah:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Fungsi koefisien korelasi mengembalikan matriks dengan bentuk berikut:

$$Corrcoef(x, y) = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix}$$

Koefisien korelasi yang kurang dari 0,3 menunjukkan korelasi yang sangat kecil. Koefisien korelasi yang lebih besar dari 0,3 tetapi kurang dari 0,7 dikatakan cukup berkorelasi. Koefisien korelasi yang lebih besar dari 0,7 berarti hubungan linier yang kuat. Koefisien korelasi dari setiap sinyal konstan (bahkan dengan noise) dengan sinyal lain biasanya kecil. Untuk mendapatkan estimasi koefisien korelasi yang baik, terutama untuk kumpulan data dengan besaran yang bervariasi, data harus diskalakan atau dinormalisasi terlebih dahulu. Jika tidak, itu akan lebih mementingkan input dengan besaran yang lebih besar. **3. Teknik Penambangan Data Prediktif**

3.1. Beberapa Teknik Regresi Linear (MLR)

Model regresi linier berganda memetakan sekelompok variabel prediktor *x* ke variabel respons *y* (Berk, 1977). Persamaan pemetaan dalam bentuk:

$$y = w1x1 + w2x2 + w3x3 + ... + WPXP$$

di mana ***wi*** adalah koefisien regresi. Ini juga dapat direpresentasikan dalam formasi matriks, dalam hal ini *b* setara dengan intersep pada sumbu *y*:

$$y = Xw + b = [X\ 1] * \begin{bmatrix} w \\ b \end{bmatrix}$$

Kita dapat menyelesaikan hal di atas untuk matriks berat optimal, ***dengan*** berat atau kemiringan. Matriks bobot ini optimal ketika jumlah kesalahan kuadrat (SSE) minimal. Di bawah ini adalah perkiraan "*e*",

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (y - Xw)^2$$

di mana ada *n* parameter dan $\hat{y}$ adalah prediksi *y*.

Satu asumsi yang dibuat di sini adalah bahwa istilah kesalahan adalah ortogonal (independen) dan Gaussian (memiliki rata-rata nol dan varians yang diketahui; asumsi lain telah dinyatakan sebelumnya).

Kolinearitas menyebabkan model menjadi sakit. Kolinearitas adalah situasi di mana variabel berkorelasi dan jumlah kondisi (CN) sangat tinggi. Angka kondisi melayani tujuan yang sama dengan faktor inflasi varians (VIF), toleransi atau indeks kondisi (CI) (Berk, 1977; Draper 1981).

Tujuan akhir dari setiap teknik prediksi adalah untuk meminimalkan kombinasi kesalahan dan kompleksitas. Pepatah yang dikenal luas adalah bahwa semakin sederhana modelnya semakin baik. Oleh karena itu, teknik prediksi yang baik mengurangi dimensi data, mengurangi kesalahan prediksi, dan memberikan garis regresi halus. Penghalusan mengurangi bobot parameter regresi sebanyak mungkin.

3.2. Regresi Komponen Utama (PCR)

PCR menggunakan analisis komponen utama (Jolliffe, 1986; Xie, 1997) dibahas dalam bagian 2.2. PCR terdiri dari tiga langkah: perhitungan PC, pemilihan PC yang relevan dalam model prediksi, dan regresi linier berganda. Dua langkah pertama digunakan untuk mengurus kolinearitas dalam data dan untuk mengurangi dimensi matriks.

Dalam regresi komponen utama, kami mengatasi masalah dengan data kolinier dan melakukan regresi dengan mengurangi serangkaian input independen dan tidak berkorelasi.

3.3. Pemodelan Regresi Ridge (RR)

Teknik regresi punggungan mengecilkan koefisien regresi dengan menjatuhkan penalti pada ukurannya (Trevor, 2002; Malinowski, 1977). Penambahan produk alfa kuadrat dan matriks identitas disebut regularisasi, dan alfa adalah parameter regularisasi atau koefisien punggungan:

$$w = (X\,XT + \alpha^{2}\,Saya)^{-1X}\,Yt$$

Parameter ini α mengontrol trade-off antara kelancaran solusi dan kesesuaian data. Teknik ridge disebut teknik smoothing karena ditandai dengan mengurangi bobot, pada gilirannya mengurangi angka kondisi. Persamaan ridge untuk pengurangan bilangan kondisi diberikan di bawah ini.

Tanpa koefisien regularisasi "alpha", kondisi number = $\frac{S_{max}^2}{S_{min}^2}$; tetapi dengan alpha, dan *S2menit* adalah nilai tunggal maksimal dan minimal. Nomor kondisi adalah $\frac{S_{Maksimal}^2 + \alpha}{S_{min}^2 + \alpha}$, mana S_{Maks}

Ini juga sangat mirip dengan teknik regresi komponen utama karena memilih jumlah PC yang relevan. Parameter regularisasi terkait dengan nilai tunggal. Nilai α optimal sedikit lebih kecil dari komponen utama paling tidak signifikan yang akan masuk ke dalam model (nilai tunggal paling tidak signifikan).

Operasi regularisasi juga terkait dengan berat oleh

$$b = \sum_{i=1}^{n} \frac{\beta_i * v_i}{\alpha_i},$$

di mana $\beta_i = u_i^T Y$, di mana *ui* dan *vi* didasarkan pada matriks dekomposisi nilai tunggal *U* dan *V*.

Bobot kecil memberikan solusi yang halus. Jika α_i lebih besar dari α, maka regularisasi tidak banyak berpengaruh pada solusi kuadrat terkecil akhir. Jika α_i kurang dari α, maka istilah yang sesuai dalam solusi dapat dinyatakan sebagai

$$\frac{\alpha_i v_i}{\alpha_i + \alpha} * \beta_i v_i,$$

Dan istilah ini mendekati *0* karena α_i cenderung *0*. Membuat alpha (koefisien regularisasi) lebih besar membantu mengurangi bobot koefisien regresi. Hasil ini adalah salah satu manfaat regresi punggungan.

3.4. Pemodelan Kuadrat Terkecil Sebagian (PLS)

PLS melibatkan transformasi data input (x) ke variabel baru atau skor (t) dan data output (y) ke skor baru (u) menjadikannya faktor yang tidak berkorelasi dan menghilangkan kolinearitas antara variabel input dan output (Malinowski, 1977). Pemetaan linier (**b**) dilakukan antara vektor skor $\boldsymbol{t}$ dan $\boldsymbol{u}$ seperti yang ditunjukkan pada Gambar 1. Vektor skor adalah nilai data pada vektor pemuatan $\boldsymbol{p}$ dan $\boldsymbol{q}$. Selanjutnya, analisis seperti komponen prinsip dilakukan pada skor baru untuk membuat vektor pemuatan ($\boldsymbol{p}$ dan $\boldsymbol{q}$).

<<Masukkan Gambar 1 di sekitar sini>>

Desain inferensial PLS diilustrasikan pada Gambar 1. Berbeda dengan PCA, PLS berfokus pada menjelaskan matriks korelasi antara input dan output tetapi PCA berkutat pada menjelaskan varians dari dua variabel. PCA adalah teknik tanpa pengawasan dan PLS diawasi. Ini karena PLS berkaitan dengan korelasi antara input (x) dan output (y) sedangkan PCA hanya berkaitan dengan korelasi antara variabel input x.

Seperti yang dapat dilihat pada Gambar 1, b akan mewakili bagian pemetaan linier antara skor t dan u. Poin bagus dari PLS adalah mengeluarkan jumlah maksimum kovarians yang dijelaskan dengan jumlah komponen minimum. Jumlah faktor laten untuk memodelkan model regresi dipilih menggunakan faktor eigen tereduksi. Faktor eigen setara dengan nilai tunggal atau variasi yang dijelaskan dalam pemilihan PC dan

biasanya disebut Malinowski mengurangi nilai eigen. Ketika nilai eigen yang dikurangi pada dasarnya sama, mereka hanya memperhitungkan kebisingan.

3.5. Kuadrat Terkecil Sebagian Non Linear (NLPLS)

NLPLS (Frank, 1990; Bakshi, 1999) pada dasarnya sama dengan PLS. Ini melibatkan transformasi data input (*x*) ke variabel baru atau skor (*t*) dan data y ke skor baru (*u*), menjadikannya faktor yang tidak berkorelasi dan menghilangkan kolinearitas antara variabel input dan output. Desain inferensial NLPLS ditunjukkan secara diagram pada Gambar 2. Ini sama dengan proses yang dijelaskan di atas, dengan perbedaan utama adalah bahwa dalam metode PLS linier, hubungan batin dimodelkan menggunakan regresi linier sederhana sedangkan hubungan batin dalam NLPLS dimodelkan menggunakan jaringan saraf.

<<Masukkan Gambar 2 di sekitar sini>>

Untuk setiap set vektor skor yang dipertahankan dalam model, jaringan saraf Single Input Single Output (SISO) diperlukan. Jaringan SISO ini biasanya hanya berisi beberapa neuron yang disusun dalam arsitektur dua lapis. Jumlah jaringan saraf SISO yang diperlukan untuk unit NLPLS inferensial tertentu sama dengan jumlah komponen yang dipertahankan dalam model dan secara signifikan kurang dari jumlah parameter yang termasuk dalam model.

4. Himpunan Data dan Kriteria untuk Perbandingan Model

4.1. Himpunan Data yang Dipilih

Empat dataset digunakan dalam penelitian ini: dataset Boston Housing (lib.stat.cmu.edu/datasets), dataset Airliner, dataset COL, dan dataset simulasi (Hines, 2005). Kami melakukan analisis awal pada dataset untuk mendapatkan wawasan tentang mereka. Masing-masing dari empat set data ini memiliki properti unik. Data Perumahan Boston memiliki tiga belas variabel input yang tidak kolinier satu sama lain dan ada 506 titik data. Beberapa variabelnya kategoris. Fitur-fitur ini, diberi nomor sesuai dengan nomor kolom, meliputi:

1. Tingkat kejahatan per kapita menurut kota (CRIM)

2. Proporsi lahan perumahan dikategorikan untuk lot lebih dari 25.000 sq. ft. (ZN).
3. Proporsi hektar bisnis non-ritel per kota (INDUS).
4. Variabel dummy Sungai Charles (1 jika traktat membatasi sungai; 0 sebaliknya) (CHAS).
5. Konsentrasi Oksida Nitrat (bagian per 10 juta) (NOX).
6. Jumlah rata-rata kamar per hunian (RM).
7. Proporsi unit yang ditempati pemilik yang dibangun sebelum tahun 1940 (AGE).
8. Jarak tertimbang ke lima pusat ketenagakerjaan Boston (DIS).
9. Indeks aksesibilitas ke jalan raya radial (RAD).
10. Pajak properti nilai penuh per $ 10.000 (PAJAK).
11. Rasio murid-guru menurut kota (PTRATIO).
12. $1000*(Bk - 0.63)^2$ di mana Bk adalah proporsi penduduk Afrika-Amerika menurut kota (B).
13. Status populasi (LSTAT) yang lebih rendah.
14. Nilai rata-rata rumah yang ditempati pemilik di $ 1000 (Mval).

Dataset Airliner memiliki 19 variabel (18 variabel input dan 1 variabel output) dan 836 titik data. Dataset COL hanya memiliki tujuh variabel input dan variabel respons dengan 9559 titik data. Variabel-variabel ini memiliki korelasi yang hampir sempurna satu sama lain dan dengan variabel respons. Dataset yang disimulasikan memiliki 44 variabel dan 5.000 titik data. Preprocessing data pada kumpulan data ini membantu mengungkapkan properti data ini dan karenanya dalam pembagian himpunan data menjadi set validasi pelatihan dan pengujian, titik data dipangkas menjadi blok 200 sebelum menetapkan blok ganjil ke set pelatihan dan blok genap sebagai set uji. Dari diagnosis ini, dataset Airliner menunjukkan beberapa korelasi antara variabel dan beberapa kolinearitas tetapi tidak sekuat dataset COL. Dataset simulasi memiliki sejumlah besar variabel input dan kebanyakan dari mereka tidak membantu prediksi.

4.2. Kriteria Perbandingan Model

Banyak kriteria dapat digunakan untuk mengevaluasi kemampuan prediksi dari teknik penambangan data yang berbeda. Dalam makalah ini, kriteria berikut akan digunakan untuk membandingkan teknik: R-square, R-square adjusted, mean square error (MSE), mean absolute error (MAE), koefisien efisiensi, condition number (CN), dan jumlah fitur yang termasuk dalam model.

- R-square (R2 atau R-sq) mengukur persentase variabilitas dalam matriks data yang diberikan yang dicatat oleh model yang dibangun (nilai dari 0 hingga 1).
- R-square adjusted (R2adj) memberikan estimasi R2 yang lebih baik karena tidak terlalu terpengaruh oleh pencilan. Sementara R-sq meningkat ketika fitur (variabel input)

  ditambahkan, R2adj hanya meningkat jika fitur yang ditambahkan memiliki informasi tambahan yang ditambahkan ke model. Nilai R2adj berkisar antara 0 hingga 1.
- Kesalahan kuadrat rata-rata (MSE): MSE prediksi adalah rata-rata kuadrat perbedaan antara nilai variabel dependen yang diamati dan nilai variabel dependen yang diprediksi oleh model. Ini adalah rata-rata perbedaan kuadrat antara nilai yang diamati dan yang diprediksi atau kuadrat rata-rata residu. MSE dapat mengungkapkan seberapa baik model ini dalam hal kemampuannya untuk memprediksi kapan set data baru diberikan. Nilai rendah selalu diinginkan. Pencilan dapat membuat kuantitas ini lebih besar dari yang sebenarnya. MSE memberikan informasi yang setara dengan R-square yang disesuaikan

  (R2adj).
- Mean absolute error (MAE): Pengukuran ini adalah penjumlahan dari nilai absolut kesalahan (perbedaan antara yang diamati dan prediksi). MAE memiliki keunggulan

dibandingkan UMK karena menangani estimasi berlebihan karena pencilan. Menggunakan MSE, kumpulan data yang memiliki banyak pencilan akan membengkak ketika mereka dikuadratkan, dan ini mempengaruhi angka yang dihasilkan bahkan ketika akar kuadrat dihitung.

- Modified Coefficient of Efficiency (E-mod.): Ini telah digunakan di banyak bidang ilmu untuk mengevaluasi kinerja model (Legates, 1999; Nash, 1970; Willmott, 1985). Menurut Nash et al. (1970), koefisien efisiensi dapat didefinisikan sebagai

$$CE = 1 - \frac{\sum_{i=1}^{n} (O_i - X_i)^2}{\sum_{i=1}^{n} (O_i - \bar{X})^2} = 1 - \frac{UMK}{\text{Varians dari Diamati}}.$$

Rasio kesalahan kuadrat rata-rata terhadap varians data yang diamati dikurangi dari kesatuan. Ini berkisar dari -1 hingga +1, di mana -1 menunjukkan model yang sangat buruk, karena rata-rata yang diamati adalah prediktor yang lebih baik daripada variabel yang diprediksi. Nilai nol akan menunjukkan bahwa rata-rata yang diamati sama baiknya dengan model yang diprediksi.

- Nomor kondisi (CN)/berat koefisien regresi: ukuran yang menunjukkan stabilitas model. Angka kondisi tinggi (>100) menunjukkan bahwa masalahnya tidak terkondisi dan karenanya tidak dapat memberikan hasil yang konsisten atau stabil.
- Jumlah variabel atau fitur yang termasuk dalam model (N): Teknik penambangan data prediktif yang baik menyumbang sebagian besar informasi yang tersedia. Ini membangun model yang memberikan perwakilan informasi yang paling mungkin dari sistem yang diprediksi dengan MSE yang paling tidak mungkin. Namun, ketika lebih banyak fitur ditambahkan, kesalahan kuadrat rata-rata cenderung meningkat.

Penambahan lebih banyak informasi yang ditambahkan meningkatkan kemungkinan menambahkan informasi yang tidak relevan ke dalam sistem. Model DM yang baik memilih fitur terbaik yang akan menjelaskan informasi terbanyak.

5. Hasil dan Perbandingan

Gambar 3 menunjukkan diagram metodologi yang digunakan dalam makalah ini. Keempat set data pertama kali diperkenalkan, serta diagnosis awal yang dilakukan pada setiap set data untuk mendapatkan wawasan tentang properti mereka. Pemeriksaan hubungan dilakukan dengan memplot input di atas output dari kumpulan data mentah. Data diproses sebelumnya dengan menskalakan atau menstandarkannya (persiapan data) untuk mengurangi tingkat dispersi antara variabel dalam kumpulan data. Koefisien korelasi dari masing-masing berbagai kumpulan data dihitung untuk memverifikasi lebih lanjut tentang hubungan antara variabel input dan variabel output. Ini diikuti dengan menemukan dekomposisi nilai tunggal dari kumpulan data, mengubahnya menjadi komponen utama. Ini juga akan membantu dalam memeriksa hubungan antara variabel dalam setiap kumpulan data.

<<Masukkan Gambar 3 di sekitar sini>>

Pada tahap ini, kumpulan data dibagi menjadi dua bagian yang sama, menetapkan titik data angka ganjil sebagai "set pelatihan" dan titik data angka genap sebagai "set data validasi pengujian." Sekarang data kereta untuk setiap kumpulan data digunakan untuk pembuatan model. Untuk setiap set data kereta, teknik penambangan data prediktif digunakan untuk membangun model, dan berbagai metode teknik itu digunakan. Tidak tersedianya kumpulan

data kehidupan nyata yang berbeda tetapi serupa telah membatasi penelitian ini untuk hanya menggunakan kumpulan data uji untuk validasi model. Ini bukan masalah serius karena pekerjaan ini terbatas pada perbandingan model dan tidak terutama berkaitan dengan hasil setelah penyebaran model.

Akhirnya, semua metode dari semua teknik dibandingkan. Teknik atau algoritma terbaik memberikan prediksi terbaik untuk jenis kumpulan data tertentu.

5.1. Analisis Data Perumahan Boston

5.1.1. Model Regresi Linier Berganda

Dalam model MLR, tiga metode dipertimbangkan: regresi model penuh, regresi bertahap, dan pemilihan variabel berdasarkan korelasinya dengan variabel respons menggunakan matriks koefisien korelasi dari semua variabel.

a) Regresi model penuh, (semua tiga belas variabel). Model lengkap menggunakan semua variabel respons untuk memprediksi *output yang* diamati *yi*.

$$yi = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \text{, di mana } k = 1, 2,..., 13.$$

b) Model berbasis korelasi. Matriks koefisien korelasi digunakan untuk memilih variabel yang paling baik berkorelasi dengan variabel output.

c) Regresi bertahap. Model regresi bertahap yang dibangun hanya memberikan hasil prediksi kumpulan data pelatihan. Variabel yang berbeda secara signifikan dari nol membuat model, dan variabel yang sama digunakan untuk membangun regresi linier berganda.

5.1.2. Model Regresi Komponen Utama

Ringkasan model yang dibangun PCR diberikan pada Tabel 1. Dapat diamati bahwa ada kontroversi mengenai model mana yang lebih baik, antara model dengan 11 PC (= variasi >90%) dan

yang dibangun dengan 10 PC. Menggunakan MSE, R2adj., CN, dan mempertimbangkan kesederhanaan model, model dengan sepuluh PC keluar lebih dari itu dengan sebelas PC. Menggunakan MAE dan koefisien efisiensi yang dimodifikasi, bagaimanapun, model yang dibangun dengan sebelas PC terlihat lebih baik. Jadi, dari dua opsi tersebut, model dengan sepuluh PC akan memerintahkannya dengan sebelas PC.

5.1.3. Regresi Punggungan (RR)

Tabel 1 menunjukkan hasil regresi punggungan. Model terbaik adalah yang dibangun dari data standar. Di antaranya, solusi terbaik datang dari model yang dibangun dengan parameter regularisasi (nilai alfa optimal) sebesar 4,0949. Ini memberikan hasil yang sangat stabil dengan MSE yang relatif baik, koefisien efisiensi yang dimodifikasi dengan baik dan angka kondisi yang baik. Model dengan nilai alpha 1 juga baik tetapi stabilitas hasilnya dibandingkan dengan nilai alpha

4,0949 tidak terlalu bagus.

<<Masukkan Tabel 1 di sekitar sini>>

5.1.4. Model Kuadrat Terkecil Sebagian (PLS)

Dari Tabel Ringkasan 1, dapat diamati bahwa model dengan sembilan faktor mengungguli setiap model lain di PLS. Model dengan semua faktor (tiga belas faktor) lebih baik hanya dengan R2adj dan dilakukan sama dengan model dengan sembilan faktor dalam hal MSE.

Dalam hal setiap kriteria lain kecuali MAE, model dengan sembilan faktor berkinerja lebih baik. Model terbaik dalam hal MAE adalah model dengan tiga faktor dan nomor kondisinya sangat baik di 7,2. 5.1.5. Model Kuadrat Terkecil Parsial Nonlinier (NLPLS)

Hasil NLPLS ditunjukkan pada Tabel 1. Semua parameter yang diukur lebih baik di NLPLS. Ada kemungkinan bahwa NLPLS memetakan juga nonlinieritas ke dalam model. MSE adalah 17.9547, yang paling sedikit di antara semua model.

5.2. Analisis Kumpulan Data COL

5.2.1. Beberapa Model Regresi Linier (MLR)

Tiga model dibangun, seperti yang dilakukan sebelumnya dalam analisis kumpulan data Perumahan Boston. Tabel 2 memberikan hasil MLR pada data COL. Ini menunjukkan bahwa semua metode memberikan hasil yang sama. Untuk kumpulan data yang berkorelasi hampir sempurna, penggunaan variabel bertahap atau berkorelasi tidak membuat banyak perbedaan. Gambar 4 menunjukkan bahwa prediksi terlihat sempurna, tetapi ada masalah serius collinearity. Angka kondisinya terlalu tinggi, dan ini akan membuat model sangat tidak stabil.

<< Sisipkan Tabel 7 di sekitar sini>>

<<Masukkan Gambar 4 di sekitar sini>>

5.2.2. Regresi Komponen Utama (PCR)

Dari Tabel 2, dengan MSE, model lengkap mengungguli model lain tetapi model lengkap memiliki masalah serius dengan jumlah kondisi yang sangat tinggi. Oleh karena itu, modelnya sangat tidak stabil. Model yang memberikan hasil terbaik dengan konsistensi yang wajar adalah

model berbasis korelasi. Jumlah kondisi di bawah 100 tetapi MSE tinggi dibandingkan dengan model penuh. Model yang dibangun dengan variabel yang skornya berkorelasi dengan output adalah yang terbaik dalam PCR pada data COL. Bahkan dengan PC yang lebih sedikit, itu mengungguli model yang dibangun dengan tiga PC.

5.2.3. Regresi Ridge pada Data COL

Tabel 2 menunjukkan bahwa regresi punggungan dengan nilai alfa optimal 9 menonjol dalam model punggungan. Kumpulan data sangat kolinier dan karenanya sangat tidak terkondisi. Hanya nilai alfa yang akan membuat kompromi dengan MSE saat smoothing akan memberikan hasil yang cukup konsisten dan stabil. Dengan nilai alfa 3,6, solusinya tampak baik, tetapi angka kondisinya masih sangat tinggi (1965,6); Oleh karena itu, model ini sangat tidak stabil karena kumpulan data sangat tidak terkondisi. Dua model punggungan pertama adalah bukti bahwa regresi punggungan berkinerja lebih baik ketika data distandarisasi sebelum dianalisis. Oleh karena itu dengan data yang tidak terstandarisasi, smoothing (menggunakan parameter regularisasi) memberikan hasil yang tidak masuk akal.

5.2.4. Kuadrat Terkecil Sebagian (PLS) pada data COL

Seperti dapat dilihat pada Tabel 2, model terbaik adalah model nilai Eigen optimal (empat faktor). Solusi dari faktor optimal (4 faktor) dan model yang dibangun dengan semua faktor tampak hampir sama dalam hal R.Sq. R^{2adj}, MAE dan koefisien efisiensi yang dimodifikasi. Jumlah kondisi mereka CN di atas 2.000. Model yang dibangun dengan hanya dua faktor memiliki angka kondisi yang baik (49), dan R.Sq dan R^{2adj} tidak buruk, tetapi MSE relatif tinggi (57,8).

5.2.5. Non-linear partial least squares (NLPLS) pada data COL

Hasil dari tiga model NLPLS diberikan pada Tabel 2

. Model dengan hanya dua faktor mengungguli satu dengan empat faktor. Faktor laten optimal 5 (C) memberi MSE 22,7752 dan MAE 3,4819. Solusinya tidak stabil dengan NLPLS dan karena itu tidak dapat diandalkan. Diamati bahwa ketika data dilatih ulang, hasil optimal baru muncul. Ini diulang berkali-kali, dan hasil optimal yang berbeda diperoleh setiap kali.

5.3. Regresi Linier Berganda pada Data Pesawat

Ringkasan hasil MLR ditunjukkan pada Tabel 3. Model lengkap jelas mengungguli sisanya di MLR. Dapat diamati bahwa semua model yang menggunakan MLR memiliki angka kondisi yang sangat tinggi; Oleh karena itu, solusi dari model ini sangat tidak stabil dan karenanya tidak realistis.

5.3.1. Regresi Komponen Utama pada Data Pesawat

Tabel 3 merangkum hasil lima model PCR. Dalam hal MSE, model lengkap dan model dengan 13 PC adalah yang terbaik di grup. Model lengkap memberikan MSE paling sedikit, tetapi angka kondisi tinggi menunjukkan sangat tidak stabil dan tidak memberikan solusi unik. Model dengan sepuluh PC memberikan MSE, MAE, R2adj, dan E-mod yang relatif baik, dan nomor kondisinya juga sangat baik. Oleh karena itu, ini adalah pilihan pertama dalam PCR, diikuti oleh PC yang berkorelasi dengan 1 hingga 4 PC.

<<Masukkan Tabel 3 di sekitar sini>>

5.3.2. Regresi ridge pada data Airliner

Model terbaik, seperti yang terlihat pada Tabel 3, adalah model yang dibangun dengan nilai alfa optimal 6,65. 5.3.3. Partial Least Squares (PLS) pada data Pesawat

Ringkasan hasil PLS pada data Airliner ditunjukkan pada Tabel 3. Model dengan semua faktor berkinerja buruk dalam hal jumlah kondisi dan kesederhanaan model tetapi paling baik didasarkan pada setiap kriteria pengukuran model lainnya. Di sisi lain, model dengan tiga faktor memiliki angka kondisi yang baik tetapi UMK yang buruk dibandingkan dengan UMK dengan jumlah faktor optimal.

5.3.4. Kuadrat Terkecil Parsial Non-Linear pada Data Pesawat

Model NLPLS menggunakan fungsi pelatihan jaringan saraf untuk belajar dari data pelatihan. Menggunakan faktor laten14, 15 untuk membangun model, kesalahan absolut rata-rata menjadi 1,2992 dan 1,4228. Hal ini ditunjukkan pada Tabel 3. UMK 4,0712 dan 4,9818, namun angka kondisinya tinggi.

5.4. Analisis Dataset Simulasi

5.4.1. Regresi Linier Berganda pada Kumpulan Data Simulasi

Tabel 4 menunjukkan hasil MLR pada kumpulan data yang disimulasikan. Model lengkap adalah yang terbaik dalam hal semua kriteria pengukuran lainnya kecuali untuk nomor kondisi. Jumlah kondisi ketiga model itu cukup tinggi. UMK tidak berbeda secara signifikan. Plot output data uji yang diprediksi untuk tiga model MLR pada data simulasi menunjukkan prediksi yang baik.

<<Masukkan Tabel 4 di sekitar sini>>

5.4.2. Regresi Komponen Utama pada Kumpulan Data Simulasi

Dari Tabel 4, Model 3, 4, 5, dan 6a baik karena mereka memiliki MSE yang baik, tetapi Model 5 memiliki nomor kondisi yang sangat tinggi (1951,7). Model 3 adalah yang terbaik dengan 26 PC (PC yang memiliki hingga 90% dari informasi yang dijelaskan). Model ini memiliki MSE terkecil dan nomor kondisinya di bawah 100. Model 3 lebih baik daripada Model 4 karena terlihat lebih sederhana dengan 26 PC, dibandingkan dengan 29 PC di Model 4. Model berbasis korelasi sangat dekat dengan keduanya dan juga bagus karena, dengan hampir setengah jumlah PC Model 3 dan 4, ia memiliki MSE dan nomor kondisi yang masuk akal.

5.4.3. Regresi Ridge pada kumpulan data Simulasi

Tabel 4 adalah ringkasan regresi ridge pada kumpulan data Simulasi. Solusi terbaik, dari Tabel 4, adalah model yang dibangun dengan nilai alpha optimal 18,44. Dalam model ini, ada sedikit kompromi antara parameter smoothing dan bias, MSE. Jumlah kondisi berkurang dari 10.000 menjadi 87. Model yang dibangun dengan nilai alfa 3 terlihat sangat mirip dengan model dengan nilai alfa nol. Dapat disimpulkan bahwa pada nilai alfa ini, penghalusan belum dimulai. Dapat juga diperhatikan bahwa α 0 dan 3,06 tampak sama dengan MLR model lengkap dan PCR model penuh dengan perbedaan hanya pada angka kondisi.

5.4.4. Kuadrat Terkecil Sebagian pada Kumpulan Data Simulasi

Dari Tabel 4, solusi terbaik menggunakan PLS adalah model yang dibangun dengan jumlah faktor optimal (8) dari metode iteratif (generalisasi). Ini memiliki UMK terbaik dibandingkan dengan yang lain, dan jumlah kondisinya di bawah 100.

5.4.5. Non-linear partial least squares (NLPLS) pada kumpulan data simulasi

Tabel 4 adalah ringkasan NLPLS pada kumpulan data yang disimulasikan. Model NLPLS terbaik adalah model faktor optimal yang dibangun dengan delapan faktor; itu memiliki MSE yang lebih baik daripada model lengkap dan memiliki nomor kondisi yang baik.

5.5. Ringkasan perbandingan

Perbandingan teknik penambangan data prediktif di atas dirangkum dalam Tabel 5 hingga Tabel 8, yang dapat berfungsi sebagai pedoman bagi para peneliti dan praktisi untuk memilih model yang sesuai untuk kumpulan data yang berbeda.

<<Masukkan Tabel 5 ke Tabel 8 di sekitar sini>>

6. Kesimpulan

Makalah ini membandingkan lima teknik penambangan data prediktif populer, termasuk regresi linier berganda (MLR), regresi komponen utama (PCR), regresi punggungan, kuadrat terkecil parsial (PLS) dan model kuadrat terkecil parsial nonlinier (NLPLS), pada empat set data unik: data perumahan Boston klasik, data COL, data Maskapai, dan data simulasi. Perbandingan model didasarkan pada kriteria yang berbeda: R-square, R-square disesuaikan, mean square error (MSE), mean absolute error (MAE), koefisien efisiensi, nomor kondisi (CN), dan jumlah variabel fitur yang termasuk dalam model.

PLS umumnya berkinerja lebih baik daripada empat teknik lainnya ketika membangun model linier. Ini berkaitan dengan kolinearitas dalam data COL dan memberikan model paling sederhana yang membuat prediksi terbaik. PLS juga mengurangi dimensi data. Studi ini menunjukkan bahwa teknik yang diawasi menunjukkan kemampuan prediksi yang lebih baik

daripada teknik yang tidak diawasi. Dapat dilihat bahwa dalam MLR dan PCR, model berbasis korelasi yang merupakan teknik yang diawasi dilakukan cukup baik daripada kebanyakan model di mana variabel dan PC dipilih secara acak untuk membangun model. Variabel yang menambahkan informasi berharga ke model prediksi adalah yang memiliki korelasi dengan output yang diprediksi.

Regresi Ridge juga berkinerja sangat baik dengan data Airliner yang tidak terkondisi. Ini mengurangi jumlah kondisi matriks data dari 137.000.000 menjadi hanya 62 dengan sedikit kompromi pada MSE (bias). Ini juga berkinerja lebih baik daripada kebanyakan teknik lain pada data COL: jumlah kondisi ditarik turun dari 4.000.000 menjadi hanya 383, dan hanya dikalahkan oleh PLS dengan dua faktor.

Beberapa masalah penambangan data prediktif adalah tipe non-linear. Untuk masalah prediksi (atau peramalan) yang sangat kompleks, algoritma non-linear atau perpaduan teknik linier dan non-linear sangat membantu. Upaya harus diarahkan untuk membangun model gabungan. Penting untuk mengevaluasi kekuatan teknik berikut: jaringan saraf (NN), mendukung regresi vektor (SVR), pohon regresi, regresi kernel, SVR kernel, dan sebagainya.

Ucapan Terima Kasih

Pekerjaan ini didukung sebagian oleh hibah start-up dari University of Tennessee. Para penulis ingin mengucapkan terima kasih kepada Wesley J. Hines di University of Tennessee untuk menyediakan dataset dan masukannya untuk pekerjaan penelitian kami.

Referensi:

Bakshi R. Bhavik, dan Utomo Utojo (1999) "Kerangka Kerja Umum untuk Penyatuan metode pemodelan saraf, Kemometrik, dan Statistik", *Analytica Chimica Acta*, 384, hlm. 227-274.

Berk, Kenneth, N. (1997) "Toleransi dan Kondisi dalam Perhitungan Regresi", *Journa dari American Statistical Association*, 72: 360, Desember, hlm. 865-866.

Berry, M. J. A., dan G. S. Linoff (2003) *Menguasai Data Mining*. New York: Wiley.

Berson, A., K. Thearling dan J. Stephen (1999) *Membangun Aplikasi Data Mining untuk CRM*, AS, McGraw-Hill.

Betts, M. (2003) "Almanak: Hot Tech", *ComputerWorld* 52, 17 November.

Chapman, P. et al. (2000) "CRISP-DM 1.0: Panduan Penambangan Data Langkah demi Langkah", *Konsorsium CRISP-DM*, http://www.crisp-dm.org.

Cohen, Jacob, et al. (2003) *Menerapkan Analisis Regresi / Korelasi Berganda untuk Ilmu Perilaku*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Draper, N., dan H. Smith (1981) *Analisis Regresi Terapan*, edisi ke-2. New York: John Wiley, hlm. 307-312.

Frank, I. E. (1990) "Model PLS Nonlinier", *Jurnal Kemometrik dan Sistem Laboratorium Cerdas*, 8, hlm. 109-119.

Gencay, Ramazan, and F. Selcuk (2002) *Pengantar Wavelet dan lainnya Metode Penyaringan di Bidang Keuangan dan Ekonomi*. San Diego, CA: Elsevier.

Giudici, P. (2003) *Penambangan Data Terapan: Metode Statistik untuk Bisnis dan Industri*. West Sussex, Inggris: John Wiley and Sons.

Han, J., dan M. Kamber (2006) Data Mining: Konsep dan Teknik, Edisi 2, New York: Morgan Kaufman.

Hines, JW (2005), Komunikasi Pribadi (*Pemantauan Tingkat Lanjut dan Teknik Diagnostik,* NE 579), Universitas Tennessee, Knoxville.

Indurkhya Nitin (1998) Predictive Data Mining: panduan praktis, New York: Morgan Kaufman.

Johnson, D.E. (1998) *Metode Multivariat Terapan untuk Analisis Data*, Pacific Grove CA: Brooks / Cole.

Jolliffe, I.T. (1986) *Analisis Komponen Utama*, New York: Springer-Verlag.

Kassams Lee Yong (1985) "Generalized Median Filtering dan Teknik Nonlinear Terkait ", *Transaksi IEEE pada Pemrosesan Sinyal*, 33: 3, hlm. 672-683.

Koh, Chye Hian, dan Kee Chan Low (2004) "Going Concern Prediction Using Data Mining Techniques", *Managerial Auditing Journal*, 19:3.

Legates, David R. dan Gregory J. McCabe (1999) "Mengevaluasi Penggunaan Goodness of Fit Measures dalam Validasi Model Hidrologi dan Hidroklimatik", *Penelitian Sumber Daya Air,* 35, hlm. 233-241.

Lessmann Stefan dan Vob Stefan (2009), "model referensi untuk data mining yang berpusat pada pelanggan dengan dukungan mesin vektor", *Jurnal Operasional Eropa Riset,* 199, 2, hlm. 520-530.

Lyman, P., dan Hal R. Varian (2003) "Berapa banyak penyimpanan yang cukup?", *Penyimpanan*, 1:4.

Malinowski, E. R. (1977) "Penentuan Jumlah Faktor dan Kesalahan Eksperimental dalam Matriks Data", *Anal. Chem*. 49, hlm. 612-617.

Mitra S. dan Acharya T. (2006) Data Mining: Multimedia, Soft Computing, dan Bioinformatika, Edisi 1, Wiley-Interscience.

Nash, J.E. dan J.V. Sutcliffe (1970) "Peramalan Aliran Sungai melalui Model Konseptual: Bagian 1-A Diskusi Prinsip", *J. Hidrologi*, 10, hlm. 282-290.

Olaf, Rem, dan M. Trautwein (2002) "Best Practices Report Experiences with Using the Mining Mart System", *Mining Mart Techreport.* No. D11.3.

Pregibon, D. (1997) "Data Mining", *Komputasi Statistik dan Grafik*, hlm. 7-8.

Pyle, Dorian (1999) · *Persiapan Data untuk Data-Mining.* San Francisco, Morgan Kaufmann.

Pyzdek, Thomas, (2003) *Buku Pegangan Six Sigma, Direvisi dan Diperluas.* Jakarta: McGraw-Bukit.

Tikhonov, A. N. (1963) "Solusi untuk Masalah yang Dirumuskan Secara Tidak Benar dan Metode Regularisasi", *Soviet Math* DokI 4, 1035-1038, Terjemahan Bahasa Inggris dari *DokI Akod Nauk* SSSR 151, hlm. 501-504.

Trevor, H., T. Robert, dan F. Jerome (2002) *Unsur-unsur Pembelajaran Statistik*, Baru York: Springer-Verlag.

Usama, M. Fayyad, et al. (1996) *Kemajuan dalam Penemuan Pengetahuan dan Data Mining.* Cambridge, Misa .: MIT Press.

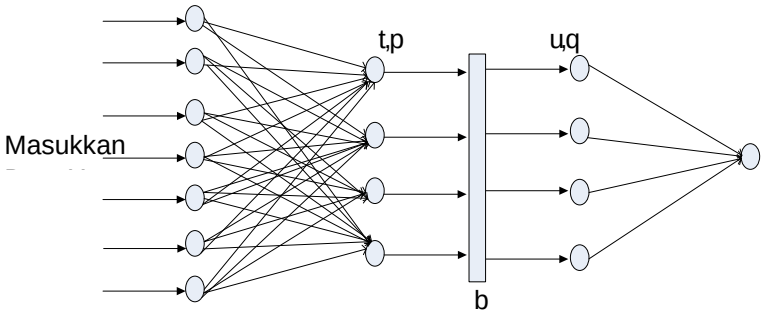
Usama, M. Fayyad (1996) "Data-Mining and Knowledge Discovery: Making Sense Out of Data", *Microsoft Research IEEE Expert,* 11:5., hlm. 20-25.

Willmott, CJ, SG Ackleson, RE Davis, JJ Feddema, KM Klink, DR Legates, J. O'Donnell, dan CM Rowe (1985) "Statistik untuk evaluasi dan perbandingan model", *J. Geophy. Penelitian* 90, hlm. 8995-9005.

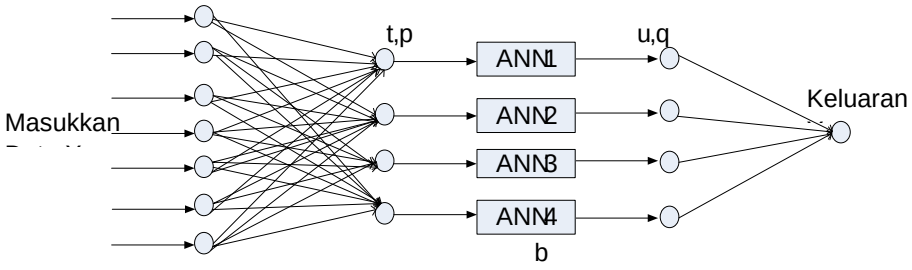
Witten, I. H. dan Frank, E. (2000) Penambangan data. New York: Morgan-Kaufmann.

Xie, Y.-L, dan J.H. Kaliva (1997) "Evaluasi Metode Pemilihan Komponen Utama untuk membentuk Model Prediksi Global dengan Regresi Komponen Utama", *Analytica Chemica Acta,* 348: 1, Agustus. hlm. 19-27.

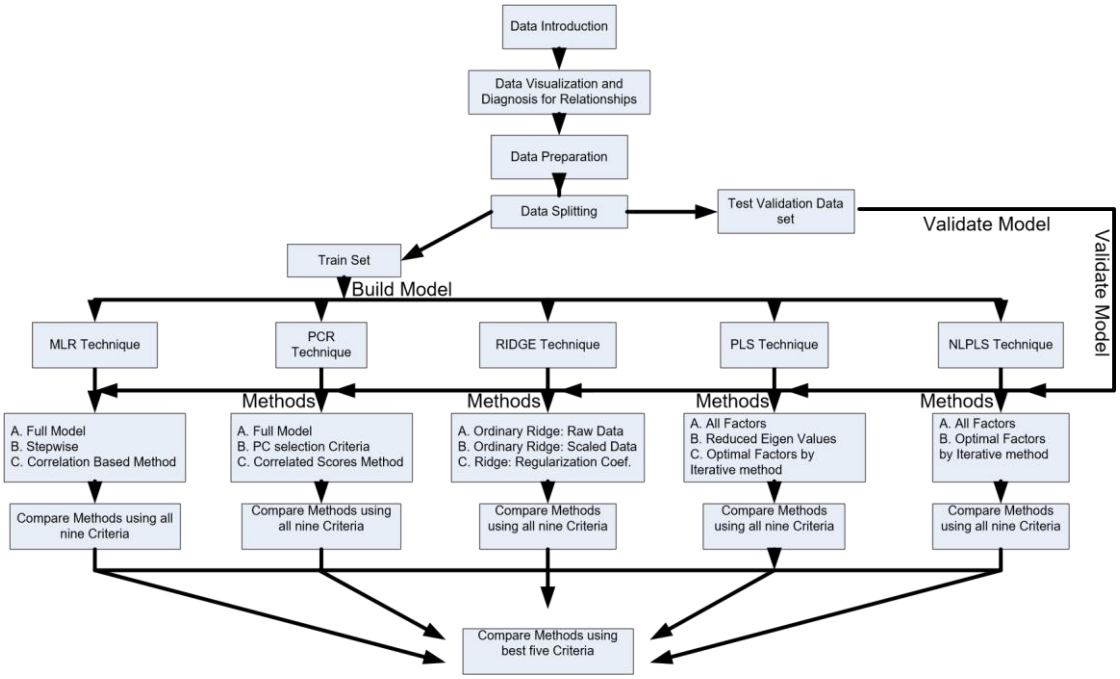
Daftar tokoh



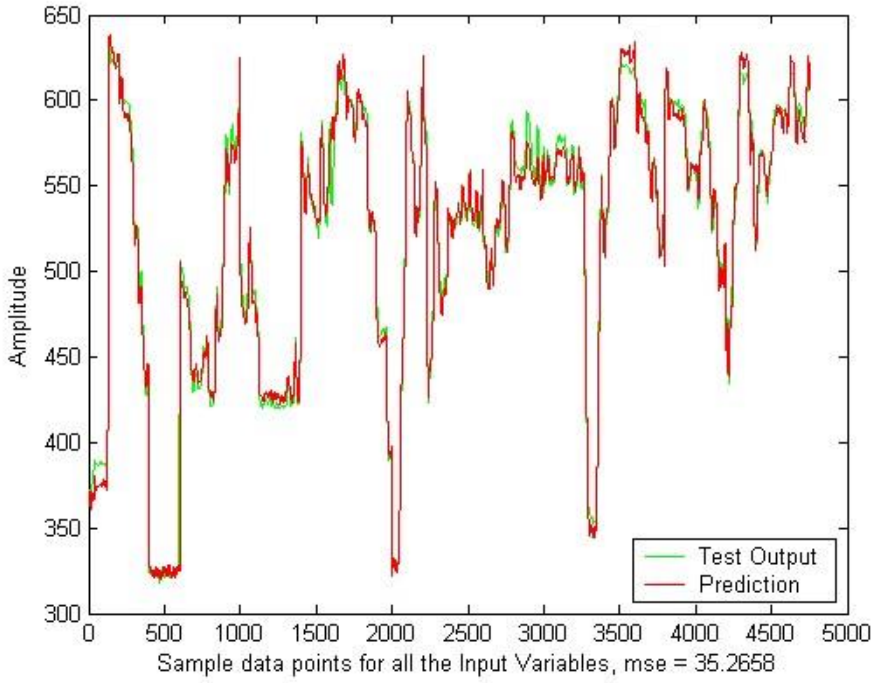
Gambar 1: Diagram skematik dari PLS Inferential Design



Gambar 2: Diagram skematik Desain Inferensial Non Linear Partial Least Squares.



Gambar 3: Diagram alir metodologi



Gambar 4: Output uji yang diprediksi diplot pada data uji output

Daftar Tabel

Tabel 1: Ringkasan analisis dataset perumahan Boston.

MLR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Model lengkap	0.7445	0.7306	21.1503	3.2500	0.4405	7,33e+7	13
Cor.Coeff.	0.7038	0.6915	24.5201	3.4430	0.3645	7,14e+7	11
Bertahap	0.6727	0.6968	24.5971	3.3989	0.3809	2.122+7	6
PCR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**# Jumlah PC**
Penuh	0.7445	0.7317	21.1503	3.250	0.4405	87.5639	13
=>90%	0.7160	0.7043	23.5042	3.3517	0.3948	31.9635	11
Lutut ke-2	0.7181	0.7077	23.3328	3.3714	0.3833	29.7520	10
Lutut ke-1	0.6943	0.6906	25.3053	3.4919	0.3432	7.1561	4
Kor.PC (1-3)	0.6716	0.6690	27.1818	3.5908	0.3393	7.1561	3
Kor. PC (1-5,12)	0.7280	0.7225	22.5133	3.3975	0.3919	36.3902	6
RR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Data Mentah	0.000	0.0000	340100	460	0.000	7.3 E+7	13
Data berskala α=0	0.7160	0.7043	23.5042	3.3517	0.3948	31.9635	13
Data berskala,α=1	0.7444	0.7305	21.1576	3.2429	0.4396	82.8704	13
Data berskala, α = 4,0949	0.7387	0.7257	21.6261	3.2255	0.4166	45.1361	13
PLS	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**# Jumlah Faktor**
Red.eig 1	0.7212	0.7201	23.0754	3.3019	0.3952	<7	2
Red.eig 2	0.7330	0.7309	22.0992	3.2433	0.4204	7.2	3
Min.eig.	0.7400	0.7359	21.5159	3.2928	0.4360	<36	5
Optimal	0.7446	0.7307	21.1395	3.2498	0.4408	<29	9
Semua faktor	0.7445	0.7317	21.1503	3.2500	0.4405	87.5639	13
NLPLS	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**Faktor**
Tanggal 1	0.7831	0.7805	17.9547	2.9545	0.5121	-	4
Ke-2	0.7921	0.7853	17.9547	2.9120	0.5180	-	9
ke-3	0.7925	0.7866	17.1734	2.9182	0.5216	-	8

Tabel 2: Ringkasan analisis dataset COL.

MLR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
	0.9944	0.994	35.2658	4.7274	0.9266	4,24E+06	7
PCR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**Pcs.**
Lutut	0.9899	0.9899	63.5893	6.1286	0.9053	49.1301	2
=>90%	0.9898	0.9898	64.4823	6.152	0.9053	2311.4	3

Cor.built	0.9899	0.9899	63.5893	6.1286	0.9053	49.1301	2
<1% keluar	0.9942	0.9942	36.6783	4.7118	0.927	2767.1	6
Semua	0.9944	0.9944	35.2658	4.7274	0.9266	8940.5	7
RR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Data mentah α = 0	0	0	3,30e+9	10,000	0	4,2e+6	7
data berskala α =0	0.9944	0.9944	35.2658	4.7274	0.9909	8,9e+3	7
α = 3.6	0.9948	0.9948	33.0698	4.6334	0.9279	1965.6	7
α = 9	0.9914	0.9914	54.5487	5.7354	0.9093	382.69	7
PLS	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Merah. Eig.Val	0.9908	0.9908	57.8274	5.8683	0.9094	49.13	2
Val optimal.	0.9946	0.9946	33.9342	4.651	0.9278	2311.4	4
Semua faktor	0.9944	0.9944	35.2658	4.7274	0.9266	8940.5	7
NLPLS	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
4 Lat.Faktor. (A)	0.9942	0.9942	36.8616	3.4720	0.9457	-	4
2 Faktor Latin (B)	0.9958	0.9958	26.7676	3.3850	0.9471	-	2
5 Faktor Latin (C)	0.9964	0.9964	22.7552	3.4819	0.9460	-	5

Tabel 3: Ringkasan analisis dataset maskapai.

MLR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Model lengkap	0.9954	0.9952	1.1840	0.8505	0.9307	1,37e+08	18
Korelasi	0.9917	0.9915	2.1177	1.0761	0.9129	2.81+07	12
Bertahap	0.9895	0.9892	2.7089	1.2584	0.8986	1,95e+12	8
PCR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**Pcs**
Lutut	0.984	0.9836	4.0969	1.5998	0.8707	32.0314	10
PC >1%	0.9928	0.9925	1.8508	1.0187	0.9174	361.892	13
Model Lengkap	0.9954	0.9952	1.184	0.8505	0.9307	1,19E+04	18
Kor. PC 1-3	0.9489	0.9487	13.051	2.8854	0.7739	2.7382	3
Kor. PC 1-4	0.97	0.9698	7.6625	2.1825	0.8262	5.2562	4
RR	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
α mentah = 0	0	0	2,1e+9	0	0	1,37E+08	18

Diskalakan, α=0	0.9954	0.9952	1.184	0.8505	0.9989	1,37E+08	18
α = 6.6494	0.9888	0.9883	2.874	1.3361	0.8893	61.8195	18
PLS	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Eig-val.	0.9797	0.9796	5.1856	1.8057	0.8547	2.7382	3
Optimal	0.9954	0.9952	1.184	0.8505	0.9307	1,19e+04	18
NLPLS	**R-persegi**	**R-sq-Adj**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
14 lat. Faktor	0.9841	0.9836	4.0712	1.2992	0.8959	1,933e+03	14
15 Lat. Faktor	0.9805	0.9798	4.9818	1.4228	0.8878	1,933e+03	15

Tabel 4: Ringkasan analisis dataset yang disimulasikan.

MLR	**R-persegi**	**R-sq-Adj.**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
PENUH	0.9065	0.9049	0.0604	0.1946	0.6993	9,885e+03	43
COR	0.8939	0.8932	0.0685	0.2063	0.6825	2,35e+03	17
Bertahap	0.8897	0.8876	0.0698	0.2081	0.6726	1,083e+03	9
PCR	**R-persegi**	**R-sq-Adj.**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Lutut ke-1 (1)	0.8102	0.8098	0.1226	0.2894	0.5079	12.0198	6
Lutut ke-2 (2)	0.8101	0.8095	0.1226	0.2894	0.5081	33.5263	10
=>90%(3)	0.9069	0.906	0.0601	0.1943	0.6998	82.8167	26
PC >1% (4)	0.907	0.906	0.0601	0.1942	0.6998	89.2839	29
Penuh (5)	0.9065	0.9049	0.0604	0.1946	0.6993	1951.7	43
Skor (6A)	0	0	0.6459	0.6817	-59.292	45.7169	2
Skor 6(b)	0.8463	0.8455	0.0993	0.2503	0.591	62.2957	14
RIDGE	**R-persegi**	**R-sq-Adj.**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Data mentah α = 0	0.8463	-0.381	0.8768	0.6794	0.4391	9885	43
Diskalakan, α = 0	0.9065	0.9049	0.0604	0.1946	0.7616	9885.1	43
α = 3.06	0.9067	0.9065	0.0603	0.1946	0.698	9885.1	43
αopt = 18,44	0.891	0.8908	0.0704	0.2135	0.6481	87.4899	43
α = 23.2857	0.8804	0.8801	0.0773	0.225	0.622	56.1658	43
α = 26.165	0.874	0.8718	0.0814	0.2315	0.6071	44.9507	43
PLS	**R-persegi**	**R-sq-Adj.**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**N**
Merah. Eig. (a)	0.8388	0.8386	0.1041	0.2639	0.5653	1.83	3

Merah. Eig. (b)	0.9045	0.9044	0.0617	0.197	0.6942	4.3104	5
Faktor pilihan	0.907	0.9067	0.0601	0.1942	0.6998	12.2608	8
Semua faktor	0.9065	0.9049	0.0604	0.1946	0.6993	1951.7	43
NLPLS	**R-persegi**	**R-sq-Adj.**	**UMK**	**MAE**	**Mod elektronik.**	**CN**	**Faktor**
1st Opt.factors	0.8928	0.8925	0.0692	0.208	0.6952	12.2608	8
Faktor Pilihan ke-2	0.8978	0.8969	0.0664	0.2042	0.6971	12.2608	9
Semua faktor	0.8748	0.8726	0.0809	0.2162	0.6778	-	43

Tabel 5: Model linier dibandingkan dengan kuadrat terkecil parsial non-linier.

	MODEL LINIER	**NLPLS**
1	Model hanya hubungan linier	Model hubungan linier dan non-linier
2	Secara komputasi lebih murah	Secara komputasi lebih mahal
3	Hanya bagus untuk model linier	Baik untuk model linier dan model yang mengandung sekitar 20% non-linearitas
4	Untuk beberapa data kolinear, berkinerja lebih baik	Dapat memberikan hasil yang tidak stabil (lihat analisis data COL)
5	Generalisasi yang baik untuk model linier.	Tidak dapat memberikan generalisasi yang baik untuk model linier

Tabel 6: Perbandingan MLR dengan teknik regresi PCR, PLS dan Ridge.

	MLR/PLS	**PCR, PLS, RR**
1	Tidak diperlukan standardisasi atau penskalaan	Standardisasi atau penskalaan diperlukan
2	Memberikan prediksi yang baik ketika variabel input benar-benar independen	Memprediksi lebih baik ketika variabel input tidak independen satu sama lain.
3	Baik ketika variabel input semua berguna dalam memprediksi respons	Lebih baik ketika ada kebutuhan untuk pengurangan variabel. Kecuali punggungan.
4	Murah secara komputasi	Mahal secara komputasi
5	Lebih mudah dipahami dan ditafsirkan	Lebih kompleks dalam solusinya
6	Sering kali menghasilkan koefisien regresi yang besar	Koefisien regresi jauh lebih rendah
7	Terkadang memberikan hasil yang tidak stabil	Sering kali memberikan hasil yang stabil tetapi dapat memberikan solusi yang tidak mewakili matriks yang dimodelkan
8	Tidak mengurus data yang tidak terkondisi atau data kolinier	Lebih baik dengan data yang tidak terkondisi atau kolinier
9	Tidak mengurus data Collinear	Menghilangkan kolinearitas
10	Tidak lebih baik ketika ada banyak variabel berlebihan dalam input.	Lebih baik untuk pengurangan dimensi atau pemilihan fitur
11	memaksimalkan korelasi kuadrat antara input dan output yang diproyeksikan	PLS memaksimalkan kovarians antara input dan output yang diproyeksikan, PCR memaksimalkan varians dari input yang diproyeksikan, Ridges bekerja sama seperti OLS tetapi menggunakan parameter regularisasi untuk mengurangi bobot regresi.
14	Tidak mudah mendeteksi keberadaan non-linearitas dalam model	Mudah dideteksi menggunakan skor dari PCA.

Tabel 7: PCR dibandingkan dengan PLS

	PCR	**PLS**
1	Hanya mempertimbangkan variabel input dalam transformasinya	Mempertimbangkan variabel input dan output dalam transformasinya
2	Teknik tanpa pengawasan	Teknik yang diawasi
3	Komputasi yang kurang kompleks	Komputasi yang lebih kompleks
4	Menangani prediksi data kolinier	Model prediksi yang lebih baik untuk kumpulan data kolinier
5	Memberikan model prediksi yang baik	Membuat model prediksi yang lebih baik

Tabel 8: PLS / PCR dibandingkan dengan RR

	PLS/PCR	**RR**
1	Mengubah data menjadi ruang ortogonal	Tidak mengubah data
2	Mengurus kolinearitas	Mengurus kolinearitas
3	Menghapus kolinearitas dengan mengubah data menjadi ruang ortogonal	Menghapus kolinearitas dengan menggunakan koefisien regularisasi
4	Berkinerja baik dengan masalah kolinier	Selalu bekerja dengan baik dengan masalah kolinier
5	Mudah dideteksi non-linearitas dalam model	Tidak mudah dideteksi.
6	Hasil tergantung pada jumlah PC, faktor	Menggunakan variabel penuh sepanjang waktu tetapi hasilnya tergantung pada parameter regularisasi
7	Berurusan dengan masalah regresi yang tidak terkondisi dengan menjatuhkan PC yang terkait dengan nilai eigen kecil	Meredam komponen minor
8	memotong nilai tunggal ketika ada kesenjangan yang jelas antara dua nilai eigen	Bekerja dengan baik ketika tidak ada kesenjangan yang jelas antara dua nilai eigen