BAB 2
LANDASAN TEORI

2.1 Web Scraping

Web Scraping atau juga dikenal sebagai *web extraction* atau *harvesting*, merupakan sebuah teknik untuk menggali data dari *World Wide Web* (WWW) dan menyimpannya kedalam sebuah file atau database untuk nantinya dilakukan analisis [5]. *Web scraping* mengacu pada suatu proses yang bertujuan untuk mengekstraksi data dari suatu situs yang ada di internet. Informasi yang telah dikumpulkan kemudian diekspor ke dalam format yang dibutuhkan atau yang lebih berguna bagi pengguna. Format tersebut bisa disimpan dalam bentuk *spreadsheet* atau API.

2.2 Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual [3]. *Data Mining* dilakukan untuk mencari informasi yang menarik, berharga, trend dan juga pola atau korelasi yang ada pada suatu database / dataset yang besar sehingga data mining sering disebut sebagai bagian dari proses penggalian pengetahuan dalam database yang sering disebut dengan istilah *Knowledge Discovery in Database* (KDD).

Pada *data mining*, terdapat beberapa pengelompokan yang didasarkan pada tujua atau tugas dilakukannya data mining. Berikut merupakan beberapa kelompok data mining berdasarkan tugasnya yaitu [6] :

1. *Description*

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan memiliki sedikit pendukung dalam pemilihan presiden. Deskripsi dari pola dan kecendrungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecendrungan.

2. *Estimation*

Hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah data numerik dari pada ke data kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi.

3. *Prediction*

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

4. *Classification*

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

5. *Clustering*

Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidak miripan dengan record dalam kluster lain.

6. *Association*

Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul bersamaan dalam satu waktu.

Terdapat enam fase dalam melakukan data mining [6] Yaitu :

1. *Business/Research Understanding Phase*
    a. Tentukan secara jelas tujuan project / research dan persyaratan yang dibutuhkan secara keseluruhan.
    b. Ubah tujuan / goal dan batasan menjadi formula masalah data mining.
2. *Data Understanding Phase*
    a. Pertama-tama kumpulkan data
    b. Pelajari data tersebut untuk lebih memahami data tersebut sehingga mendapat wawasan awal.
    c. Evaluasi kualitas dari data tersebut

d. Lalu jika di inginkan, pilih subsets yang menarik yang kira-kira berisi pola yang dapat ditindak lanjuti

3. *Data Preparation Phase*
   a. Fase ini meliputi seluruh aspek dalam menyiapkan final data set, mulai dari data awal (initial data), mentah (raw), dan data kotor (dirty data).
   b. Pilih kasus atau variabel yang ingin di analisa, dan sesuai dengan analisis kita.
   c. Lakukan transformasi pada variabel-variable tertentu jika diperlukan
   d. Bersihkan (clean) raw data sehingga siap untuk diterapkan oleh modeling tools
4. *Modeling Phase*
   a. Pilih dan terapkan teknik pemodelan yang sesuai
   b. Sesuaikan pengaturan model untuk mendapatkan hasil yang optimal
   c. Terkadang beberapa teknik mungkin diterapkan pada permasalahan data mining yang sama.
   d. Ada kemungkinan untuk kembali ke fase sebelumnya untuk memenuhi kekurangan khusus teknik data mining tertentu
5. *Evaluation Phase*

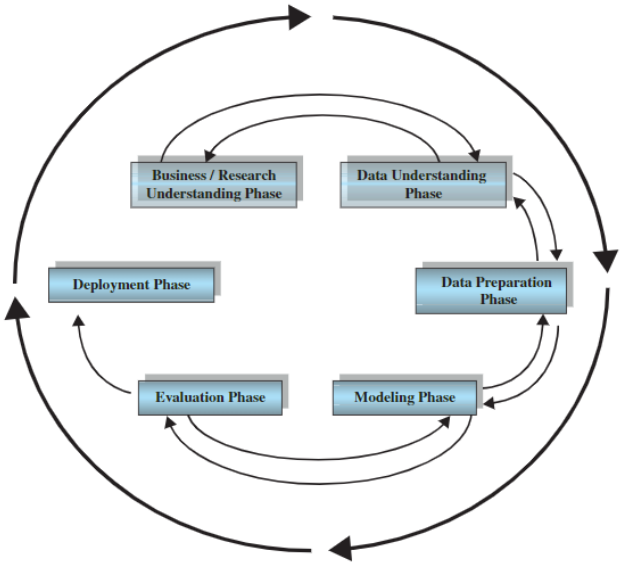
   a. Setelah mendapatkan satu atau lebih model dari fase modeling, model-model tersebut harus di evaluasi kualitas dan efektifitasnya, sebelum di terapkan di lapangan.

   b. Lalu tentukan juga apakah model tersebut benar dapat meraih tujuan yang sudah ditentukan sebelumnya

   c. Akhirnya, tentukan keputusan tentang penggunaan hasil data mining
6. *Deployment Phase*
   a. Model yang sudah dibuat tidak menentukan bahwa project sudah selesai, dibutuhkan pembuatan model tersebut.
   b. Contoh simple deployment : pembuatan laporan
   c. Contoh Deployment yang lebih kompleks : menerapkan proses data mining tersebut ke departemen lainnya

2.3 CRISP-DM

CRISP-DM disusun oleh tiga penggagas bernama Daimler Chrysler (Daimler-Benz), SPSS (ISL), NCR dan dipublikasikan pada tahun 1999. CRISP-DM merupakan singkatan dari *CRoss-Industry Standard Process for Data Mining*. CRISP-DM merupakan standarisasi proses data mining sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian [4].



Gambar 2. 1 Alur CRISP-DM [7].

Terdapat enam tahap siklus proses dari CRISP-DM, yaitu [7]:

1. *Business* understanding Phase

Fase awal ini berfokus pada memahami tujuan dan persyaratan projek (project requirement) dari perspektif bisnis, kemudian mengubah pengetahuan ini menjadi sebuah permasalahan pada data mining dan kemudian mengembangkan rencana awal yang dirancang untuk mencapai tujuan.

2. *Data Understanding Phase*

Fase pemahaman data dimulai dengan mengumpulkan data awal dan dilanjutkan dengan kegiatan untuk mengenali data, mengidentifikasi masalah kualitas data, untuk menemukan wawasan pertama terhadap data atau untuk mendeteksi subsets yang menarik untuk membentuk sebuah hipotesis.

3. *Data Preparation Phase*

Tahap persiapan data mencakup semua kegiatan untuk membangun kumpulan data akhir dari data mentah (raw data) awal.

4. *Modeling Phase*

Pada fase ini, teknik pemodelan akan dipilih dan diterapkan lalu mengkalibrasi parameternya agar mencapai nilai yang optimal.

5. *Evaluation Phase*

Pada tahap ini model yang diperoleh akan dievaluasi secara lebih menyeluruh dan langkah-langkah yang dijalankan untuk membangun model ditinjau kembali untuk memastikan model tersebut benar-benar memenuhi tujuan bisnis.

6. *Deployment Phase*

Jika tujuan dari model tersebut adalah untuk meningkatkan pengetahuan tentang data, pengetahuan baru yang diperoleh perlu diatur dan disajikan sedemikian rupa sehingga pengguna dapat memahami dan menggunakannya. Pada tahap ini sistem akan di-demokan dan dipresentasikan sehingga dapat digunakan oleh pengguna. Tahap deployment ini dapat berupa pembuatan laporan atau mengimplementasikan hasil proses data mining yang telah dilakukan.

2.4 Analisis Multivariat

Analisis multivariat (*multivariate analysis*) merupakan salah satu jenis analisis statistik yang digunakan untuk menganalisis data yang terdiri dari banyak variabel baik variabel bebas (*independent variables*) maupun banyak variabel tak bebas (*dependent variables*) [8]. Analisis multivariat digunakan untuk memecahkan masalah dimana lebih dari satu variabel terikat atau dependen dianalisis secara bersamaan dengan beberapa variabel bebas atau independen.

2.4.1 Klasifikasi Analisis Multivariat

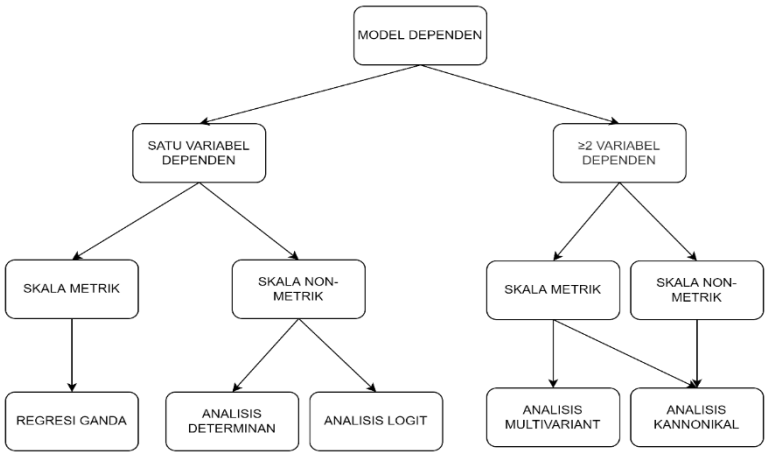
Analisis mutlivariat dapat diklasifikasikan menjadi dua bagian atau metode, yaitu analisis dependensi dan analisis interdependensi. Perbedaan teknik tersebut

didasarkan pada hubungan diantara variabel-variabel yang ada, hubungan tersebut dapat dibagi menjadi dua bagian besar, yaitu [9] :

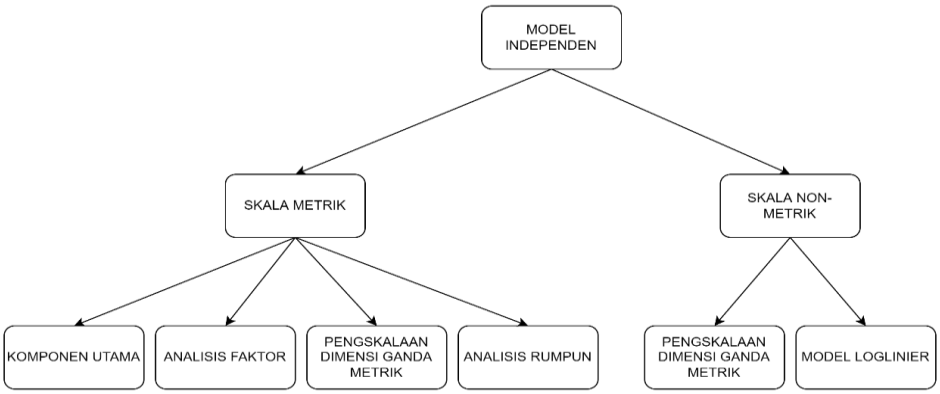
1. Variabel-variabel tidak saling bergantung satu dengan yang lain, yang disebut dengan interdependensi. Ciri penting interdependensi adalah tidak adanya variabel dependen dan variabel independen. Semua variabel bersifat independen.
2. Antar variabel ada saling ketergantungan, yang disebut dengan dependensi. Ciri penting dependensi adalah adanya dua jenis variabel yakni variabel dependen dan variabel independen.

2.4.2 Pembagian Metode Analisis Multivariat

Pemilihan metode pada analisis multivariat didasarkan pada jenis variabel dependen yang digunakan, pemilihan metode yang dimaksud yaitu [10] :



Gambar 2. 2 Pembagian metode analisis multivariat dependen



Gambar 2. 3 Pembagian metode analisis multivariat independen

2.5 Analisis Regresi

Analisis regresi merupakan salah satu metode dalam dunia statistik yang sering digunakan dalam ilmu pengetahuan terapan. Analisis regresi merupakan suatu teknik yang digunakan untuk menentukan hubungan sebab akibat pada satu variabel dengan variabel lainnya. Metode regresi merupakan sebuah metode statistik yang melakukan prediksi menggunakan pengembangan hubungan matematis antara variabel, yaitu variabel dependen (Y) dengan variabel independen (X) [11]. Dalam analisis regresi dikenal dua jenis variabel yaitu:

a. Variabel terikat atau dependen (Y) yaitu variabel yang dipengaruhi.

b. Variabel bebas atau independen (X) yaitu variabel yang mempengaruhi

Terdapat dua jenis regresi dalam proses analisis estimasi yaitu, regresi linear sederhana dan regresi linear berganda. Jika hanya terdapat satu variabel bebas atau independen maka analisis tersebut disebut analisis regresi sederhana, dan jika variabel bebas berjumlah lebih dari satu maka disebut analisis regresi berganda.

Pendekatan standar untuk mendapatkan nilai dugaan parameter dari model regresi linier adalah menggunakan Metode OLS (Ordinary Least *Square*). Metode OLS adalah metode yang digunakan untuk meminimalisir jumlah kuadrat kesalahan dengan mengestimasi suatu garis regresi. Tetapi OLS dianggap kurang tepat untuk menganalisis sejumlah data yang tidak simetris karena memiliki *outlier*. Karena jika data berbentuk tidak simetris, maka nilai rata-rata (*mean*) yang didapat menjadi sangat peka dengan adanya data outlier.

2.6 Data Outlier

Data outlier atau data pencilan merupakan data observasi yang memiliki nilai-nilai ekstrim dibanding dengan data lainnya. Data bernilai ekstrim yang dimaksud adalan data yang memiliki nilai jauh berbeda dengan nilai lain dalam kelompok atau kolom nya. Menurut [12] Outlier adalah pengamatan yang jauh dari pusat data yang mungkin berpengaruh besar terhadap koefisien regresi. Data outlier dapat dideteksi dengan mempetakan sebaran data menggunakan boxplot.

Data outlier akan sangat berpengaruh pada proses analisis regresi OLS karena akan menyebabkan hal-hal seperti :

1. Model yang dihasilkan akan membentuk residual yang besar
2. Varians pada data yang diteliti menjadi lebih besar
3. Interval pada data memiliki rentang yang lebar

2.7 Regresi Kuantil

Regresi Kuantil merupakan salah satu pendekatan metode regresi yang diperkenalkan oleh Koenker dan Basset di tahun 1978. Metode Regresi Kuantil dapat dikatakan sebagai perluasan dari model Regresi OLS. Regresi Kuantil sangat berguna jika distribusi data tidak homogen (*heterogenous*) dan tidak berbentuk standar seperti tidak simetris, terdapat ekor pada sebaran, atau truncated distribution [13] Regresi kuantil dilakukan dengan membagi atau memisahkan data menjadi dua bagian atau lebih ketika dicurigai terdapat perbedaan nilai estimator pada kuantil-kuantil tertentu [14]. Model umum dari persamaan regresi kuantil yang terdiri dari satu variabel terikat atau dependen (Y) dan dua atau lebih variabel bebas atau independen (X) yaitu :

$$Q(Y_i) = \beta_0(\theta) + X_{i1}\beta_1(\theta) + X_{i2}\beta_2(\theta) + \cdots + X_{ip}\beta_p(\theta) + \varepsilon_i(\theta) \qquad i = 1, \ldots, n$$

Dimana :

Y = Variabel Dependen

X = Variabel Independen

p = Jumlah Fitur

$\beta(\theta)$ = Koefisien regresi kuantil

Lalu model regresi kuantil diatas dapat disajikan dalam bentuk matriks yaitu :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ : \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{31} & X_{41} \\ 1 & X_{12} & X_{22} & X_{32} & X_{42} \\ 1 & X_{13} & X_{23} & X_{33} & X_{43} \\ : & : & : & : & : \\ 1 & X_{1n} & X_{2n} & X_{3n} & X_{4n} \end{bmatrix} \begin{bmatrix} \beta_0(\theta) \\ \beta_1(\theta) \\ \beta_2(\theta) \\ : \\ \beta_n(\theta) \end{bmatrix} + \begin{bmatrix} \varepsilon_1(\theta) \\ \varepsilon_2(\theta) \\ \varepsilon_3(\theta) \\ : \\ \varepsilon_n(\theta) \end{bmatrix}$$

Mencari nilai beta pada nilai kuantil tertentu dilakukan dengan mengurangi nilai *Median Absolute Deviation* (MAD)

$$MAD = \frac{1}{n}\sum_{i=1}^{n} \rho_\theta(\beta_0(\theta) + X_{i1}\beta_1(\theta) + X_{i2}\beta_2(\theta) + \cdots + X_{ip}\beta_p(\theta))$$

Fungsi ρ merupakan fungsi pengecekan (*check function*) untuk memberikan bobot pada error tergantung pada kuantil yang telah ditentukan dan tanda error secara keseluruhan. Secara matematis, rumus ρ yaitu :

$$\rho_\theta(u) = \theta max(u, 0) + (1 - \theta)\, max\,(-u, 0)$$

Dalam rumus tersebut, u adalah nilai error pada satu titik data dan fungsi max akan mengembalikan nilai terbesar dalam tanda kurung. Sehingga jika nilai error nya positif maka fungsi pengecekan akan mengkalikan nilai error dengan θ, dan jika nilai error nya negatif maka fungsi pengecekan akan mengkalikan nilai error dengan $(1 - \theta)$.

Contohnya jika peneliti menginginkan nilai median pada kuantil ke-10 (10%) maka peneliti menginginkan 90% dari error bernilai positif dan 10% bernilai negatif. Sehingga untuk mendapatkan nilai MAD terkecil perlu menambahkan beban ke nilai error. Pada kasus kuantil ke-10, maka penambahan beban sebesar 0.9 kedalam beban negatif dan 0.1 kedalam beban positif. Sehingga masalah regresi kuantil dapat dipecahkan dengan menggunakan metode Least Absolute Deviation (LAD) sebagai berikut :

Minimalkan $: \theta \sum | d_{1i} | + (1 + \theta) \sum | d_{2i}$

Dengan Kendala $: X \beta + d_1 - d_2 = Y$

$X \beta + d_1 - d_2 > 0$

Solusi dari permasalahan regresi kuantil tersebut tidak dapat diperoleh secara analitik namun dapat dipecahkan dengan melakukan tahapan iterasi dengan metode simpleks.

2.8 Metode Simpleks

Metode simpleks merupakan salah satu cara untuk menyelesaikan masalah optimasi linear dengan melakukan iterasi atau pengulangan pada pengujian titik-titik sudut sampai dengan menemukan penyelesaian yang optimal. Metode ini berjalan selangkah demi selangkah yang dimulai dari suatu titik ekstrim pada daerah fisibel sampai dengan menuju ke titik ekstrim optimum. Metode simpleks merupakan prosedur yang dilakukan secara berulang, yang dimana cara yang sama akan digunakan pada setiap pengujian titik sudut hingga mencapai penyelesaiian yang optimal, penyelesaian yang optimal yaitu penyelesaiian yang memenuhi seluruh masalah atau kendala dan menghasilkan nilai tujuan. Didalam model pengoptimasian linear, titik sudut merupakan perpotongan antara paling sedikit dua garis kendala.

Berikut merupakan istilah-istilah yang terdapat pada metode simpleks, yaitu [15]:

1. **Variabel Slack**

Variabel slack adalah variabel basis yang berfungsi untuk menampung sisa kapasitas pada kendala yang berupa pembatas.

2. **Variabel Surplus**

   Variabel surplus adalah variabel yang berfungsi untuk menampung kelebihan nilai ruas kiri pada kendala yang berupa syarat. Dalam kasus regresi kuantil menggunakan metode LAD, variabel surplus adalah deviasi bawah yang diboboti dengan

3. **Variabel Artificial**

   Variabel artificial adalah variabel yang bernilai positif, berfungsi untuk memulai penyelesaian dan harus dijadikan nol pada solusi akhir. Variabel ini digunakan untuk setiap persamaan yang tidak memiliki variabel basis. Dalam kasus regresi kuantil menggunakan metode LAD variabel artificial adalah deviasi atas yang diboboti dengan

4. **Variabel Basis dan Nonbasis**

   Variabel basis (basic variabel) dan variabel nonbasis (nonbasic variabel) merupakan dua terminologi penting yang akan selalu digunakan di dalam algoritme simpleks. Variabel basis adalah variabel yang bernilai positif, dan variabel nonbasis adalah variabel yang bernilai nol.

Metode simpleks memerlukan suatu tabel yang biasa disebut dengan tabulasi simpleks bertujuan untuk menentukan apakah variable keputusan pada titik ekstrim telah menghasilkan nilai tujuan yang optimal. Tabulasi simpleks yang digunakan pada kasus regresi kuantil dapat dilihat pada Tabel

Tabel 2. 1 Tabulasi Simpleks

C_j			0	0	...	0	θ	...	θ	(1- θ)	...	(1- θ)
C_b	Vb	Wb	X_1	X_2	...	X_n	d_1	...	d_{1n}	d_{2i}	...	d_{2n}
d_{11}^+	X_1	b_1	a_{ij}									
d_{2n}^+	X_2	b_2										
...	...	...										
d_{1n}^+	X_n	b_n										

$Zj = \sum d_i a_{ij}$										
C_j-Z_j										

Dimana :

a. Baris C_j diisi dengan nilai koefisien fungsi tujuan, dalam kasus regresi kuantil diisi dengan nilai error yang telah ditentukan

b. Kolom CB diisi dengan nilai koefisien pada variabel yang menjadi basis

c. Kolom VB diisi dengan nama variabel yang menjadi basis.

d. Kolom WB diisi dengan nilai ruas kanan dari kendala.

Langkah-langkah penyelesaian regresi kuantil menggunakan metode simpleks adalah sebagai berikut :

1. Mengubah masalah optimasi lienar menjadi ke bentuk standar, kendala-kendala dan fungsi tujuan kemudian diubah ke dalam bentuk persamaan. Contoh : $Z = 21x_1 + 32x_2$ diubah menjadi $Z - 21x_1 - 32x_2$
2. Memilih atau menentukan kolom kunci. Untuk masalah maksimasi atau maksimum maka dipilih zj – cj terbesar, sedangkan untuk masalah minimasi atau minimum dipilih zj-cj terkecil.
3. Menentukan baris kunci dengan mimilih rasio nilai terkecil antara b_i (ruas kiri) dengan baris pada kolom kunci. Dimana $Rasio = \frac{b_i}{a_{ij}} > 0$
4. Mengubah nilai baris kunci dengan cara membaginya dengan angka kunci sehingga nilainya menjadi 1
5. Mengubah nilai-nilai selain baris kunci dengan melakukan operasi baris dasar (OBD) nilai elemen yang termasuk pada kolom kunci kemudian dijadikan nol (0).
6. Proses iterasi pada masalah maksimasi atau maksimum akan berhenti apabila semua nilai pada baris $zj - cj \leq 0$ yang menandakan bahwa solusi sudah optimum sedangkan untuk masalah minimasi atau minium proses iterasi akan berhenti apabila baris $zj - cj \geq 0$.

2.9 Goodness of fit

Goodness of fit atau uji kebaikan model pada regresi kuantil dimotivasi oleh nilai koefisien determinasi (R^2) pada regresi klasik. Goodness of fit pada regresi kuantil dapat dilakukan dengan menghitung nilai Quantile Verification Skill Score (QVSS) yang setara dengan nilai R^2 hanya saja QVSS hanya menghitung nilai kebaikan model pada kuantil yang telah ditetapkan.

$$QVSS = 1 - \frac{|\hat{y} - y|j}{|\hat{y} - y|}$$

Dimana :

$|\hat{y} - y|j$ = Sisaan dari model penuh

$|\hat{y} - y|$ = Sisaan dari model hanya intersep

Nilai dari QVSS berkisar diantara 0 – 1 saja dimana jika nilai QVSS = 1 menandakan bahwa variable bebas dapat menjelaskan variabel terikat secara penuh (100%) dan bernilai QVSS = 0 menandakan bahwa variabel bebas tidak dapat menjelaskan variabel terikat.