BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Pesatnya perkembangan teknologi informasi dewasa ini khususnya dalam aplikasi-aplikasi database yang diiringi dengan meningkatnya kapabilitas media penyimpanan yang semakin besar telah memungkinkan terjadinya akumulasi data dalam jumlah besar. Komputerisasi diberbagai bidang dan penggunaan internet sebagai sarana sistem informasi global secara signifikan juga turut berperan dalam terjadinya akumulasi data dan informasi tersebut. Pertumbuhan yang begitu pesat dari akumulasi data yang tersimpan dalam suatu database akan menciptakan suatu kondisi *"rich of data but poor of information"* dan data yang tersimpan akan menjadi kuburan data apabila tumpukan data tersebut dibiarkan begitu saja sehingga tidak dapat digunakan untuk aplikasi yang berguna.

Didalam tumpukan data tersebut mungkin terdapat informasi-informasi tersembunyi yang sangat penting atau menjadi penting pada saat dibutuhkan yang dapat dijadikan dasar atau pedoman dalam pengambilan keputusan. Keputusan sering sekali dibuat tidak berdasarkan pada data-data yang ada yang tersimpan dalam tumpukan data tersebut melainkan hanya didasarkan intuisi sang pembuat keputusan. Hal ini dikarenakan tidak adanya sistem atau perangkat lunak yang dapat membantu dalam pencarian informasi yang tepat, cepat dan akurat, dilain pihak penggalian data untuk mendapatkan informasi yang dilakukan secara manual sangatlah tidak efektif dan memakan banyak waktu.

Universitas Widyatama (UTAMA) merupakan salah satu organisasi yang bergerak dalam bidang pendidikan yang memanfaatkan teknologi informasi dalam menjalankan proses bisnisnya. Dengan adanya pemanfaatan teknologi informasi di UTAMA maka akan terjadi akumulasi data dalam jumlah besar tiap tahunnya. Salah satu data yang mengalami peningkatan tiap tahunnya yaitu data nilai Ujian Saringan Masuk (USM) mahasiswa baru di UTAMA. Semakin lama data nilai USM ini akan menjadi kuburan data yang tidak memiliki suatu nilai maupun informasi yang dihasilkan dari data tersebut. Oleh karena itu diperlukan suatu teknik dan perangkat yang dapat membantu kita dalam mentransformasikan data

dalam jumlah besar tersebut menjadi suatu informasi yang berguna yaitu dengan penerapan *Data Mining* yang diaplikasikan dengan pembuatan perangkat lunak *data mining* atau *data mining engine*. Berdasarkan latar belakang masalah tersebut penulis tertarik untuk meneliti bidang ini dengan mengambil judul **"Penerapan *Data Mining* Untuk Menemukan Pola Antara Nilai Ujian Saringan Masuk (USM) Terhadap Indeks Prestasi (IP)".**

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah diatas, maka permasalahan yang bisa di ambil dalam penelitian ini yaitu :

1. Bagaimana mentransformasikan data nilai USM dalam jumlah besar menjadi suatu informasi yang berguna ?
2. Bagaimana membangun perangkat lunak yang bisa membantu mentransformasikan data nilai USM menjadi suatu informasi yang berguna ?

1.3 Maksud dan Tujuan Penelitian

Maksud dan Tujuan dari penelitian tugas akhir ini yaitu untuk membangun suatu perangkat lunak atau aplikasi data mining (*data mining engine*), yang akan digunakan untuk menggali dan menemukan pola-pola yang tersembunyi, yang tidak diketahui antara data nilai USM dengan IP. Sehingga informasi yang dihasilkan dapat digunakan sebagai dasar analisis dalam pengambilan keputusan dan memberikan nilai tambah bagi *database* yang telah dibangun. Selain itu, penelitian ini untuk mengetahui apakah nilai USM memiliki kerataan hubungan yang erat terhadap IP, yang sesuai dengan kriteria kerataan hubungan menurut Guilford.

1.4 Batasan Masalah

Untuk menghindari salah pengertian dalam penulisan tugas akhir ini dan untuk lebih memfokuskan terhadap permasalahan, maka dalam hal ini penulis perlu untuk menetapkan batasan-batasan terhadap permasalahan yang dikaji sebagai berikut :

1. Teknik yang digunakan dalam proses *data mining* adalah regresi linier sederhana dan tidak melibatkan atau menyertakan teknik lain.
2. Database yang digunakan dalam penelitian ini yaitu database relasional dengan menggunakan *Microsoft SQL Server 2000* sebagai *Database Management System* (DBMS) dan telah melalui proses *prepocessing* yang hanya menyertakan informasi-informasi yang dibutuhkan saja dengan membuang *variabel* yang tidak dibutuhkan.
3. Dalam penelitian ini pun akan mencari nilai rata-rata dari *Random Error* untuk variabel prediktor.
4. Aplikasi *data mining engine* yang akan dibangun hanya menampilkan data hasil proses *regresi* dalam bentuk grafik dan hasil perhitungan analisis.

1.5 Metode Penelitian

Pengumpulan data yang dilakukan menggunakan cara sebagai berikut:

1. Wawancara atau *interview* yaitu metode yang dilakukan dengan cara melakukan wawancara langsung dengan pihak yang berkepentingan terutama dengan pejabat unit Akademik Universitas Widyatama, tentang apa yang mereka harapkan dari penelitian tugas akhir ini.
2. Kajian pustaka, dilakukan dengan cara membaca buku-buku dan sumber lain yang menunjang serta membantu penyelesaian permasalahan.

Metode pembangunan software aplikasi ini menggunakan metode *waterfall* yang mempunyai tahap sebagai berikut :

1. Rekayasa Sistem, merupakan tahapan yang pertama kali dilakukan yaitu merumuskan sistem yang akan dibuat.
2. Analisis, merupakan suatu tahapan yang berkaitan dengan proses dan data yang diperlukan oleh sistem.
3. Perancangan, Pada tahapan perancangan ini diberikan gambaran umum yang jelas kepada pengguna dan rancang bangun yang lengkap tentang sistem yang akan dikembangkan.
4. Implementasi, suatu tahapan yang mengimplementasikan rancangan sistem ke dalam kode-kode dalam bahasa pemrograman yang diinginkan.

5. Pengujian, digunakan untuk memeriksa apakah software yang sudah dibuat tersebut sudah sesuai dengan perancangan
6. Pemeliharaan, setelah tahap-tahap tersebut diselesaikan maka perlu dilakukan tahap yang terakhir yaitu perawatan atau pemeliharaan software.

1.6 Sistematika Pembahasan

Adapun untuk sistematika pembahasan penelitian tugas akhir ini terdiri dari 6 bab, yaitu :

Bab satu pendahuluan, membahas latarbelakang masalah, identifikasi masalah, maksud dan tujuan penelitian tugas akhir, ruang lingkup, metode penelitian yang digunakan dan sitematika pembahasan.

Bab dua landasan teori, membahas tentang teori dari teknik yang akan digunakan penulis untuk analisis dalam data. Selain itu, akan menjelaskan teori-teori yang mendukung dalam penelitian tugas akhir ini.

Bab tiga analisis dan perancangan, berisi tentang objek penelitian, analisis yang didalamnya menjelaskan identifikasi masalah, penyebab masalah, dan hasil analisis. Selain itu, berisi mengenai deskripsi prinsip kerja sistem secara umum, fungsi-fungsi sistem, *Diagram Konteks, Data Flow Diagram, Entity Relationship Diagram*, Kamus Data, dan perancangan proses.

Bab empat implementasi, berisi tentang penjelasan aplikasi secara umum, algoritma global, bentuk tampilan struktur program, langkah-langkah menjalankan program.

Bab lima kesimpulan dan saran, berisi tentang kesimpulan dan saran secara menyeluruh dari uraian-uraian pokok sebelumnya.

BAB II
LANDASAN TEORI

2.1 Pengertian Data Mining

Data Mining merupakan salah satu cabang ilmu komputer yang relatif baru yang memiliki keterkaitan dengan *machine learning*, kecerdasan buatan (*artificial intelligence*), *statistic* dan *database*. Data Mining mengacu kepada ekstraksi atau penggalian pengetahuan dari suatu data dalam jumlah besar. Ada banyak pengertian data mining itu sendiri, diantaranya seperti penggalian pengetahuan dari database, ekstraksi pengetahuan (*knowledge extraction*), analisis data atau pola (*pattern analysis*), penggalian data dan lain sebagainya.

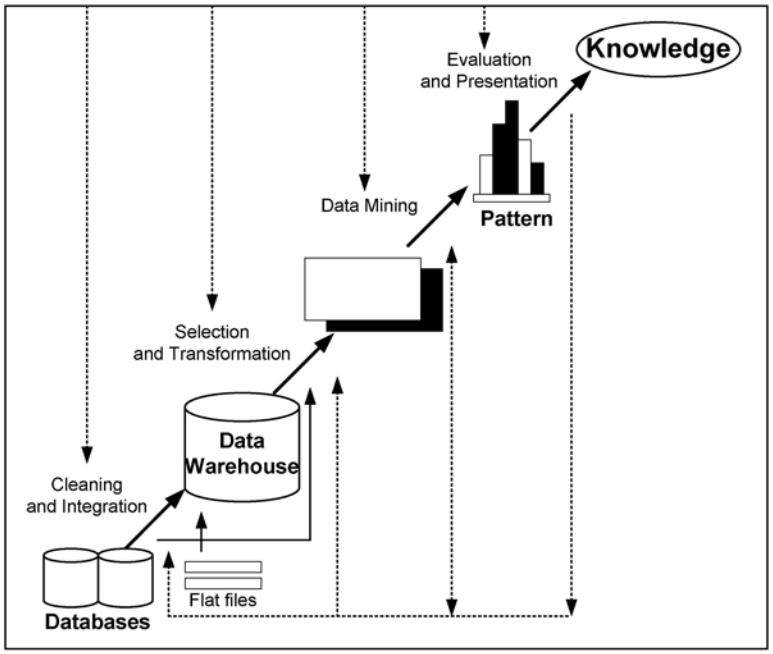
Definisi umum dari data mining itu sendiri adalah proses pencarian pola-pola yang menarik (*hidden patern*) berupa pengetahuan (*knowledge*) yang tidak diketahui sebelumnya dari suatu kumpulan data dimana data tersebut dapat berada dalam *database*, *data warehouse*, atau media penyimpanan informasi yang lain. Data mining seringkali diartikan dengan ”menulis banyak laporan dan query”, namun pada faktanya kegiatan data mining tidak melakukan pembuatan laporan dan query sama sekali. Data mining dilakukan dengan tool khusus, yang mengeksekusi operasi data mining yang telah didefinisikan berdasarkan model analisis. Data mining merupakan proses analisis terhadap data dengan penekanan menemukan informasi yang tersembunyi pada sejumlah besar data yang disimpan ketika menjalankan bisnis perusahaan.

Dalam aplikasinya, data mining sebenarnya merupakan bagian dari proses *Knowledge Discovery in Database* atau KDD, bukan sebagai teknologi yang utuh dan berdiri sendiri. Data mining merupakan suatu bagian langkah yang penting dalam proses KDD terutama berkaitan dengan ekstraksi dan penghitungan pola-pola dari data yang ditelaah, seperti ditunjukkan oleh gambar 2.1, langkah-langkah atau proses KDD itu sendiri terdiri dari :

1. Pembersihan data (*Data cleaning*), membuang noise dan data yang tidak konsisten.
2. Integrasi data (*Data integration*), menggabungan data yang berasal dari beberapa sumber.

3. Pemilihan data (*Data selection*), memilih data yang relevan atau sesuai dengan proses anilisis yang akan dilakukan.
4. Transformasi data (*Data transformation*), mengubah data menjadi bentuk yang sesuai untuk proses data mining.
5. Penggalian data (*Data mining*), merupakan proses terpenting dimana teknik data mining diaplikasikan untuk mengekstraksi pola-pola dari suatu data.
6. Evaluasi pola (*Pattern evaluation*), evaluasi pola yang ditemukan untuk menemukan pola yang bernilai atau menarik.
7. Presentasi pengetahuan (*Knowledge presentatition*), visualisasi dan teknik representasi pengetahuan digunakan untuk diperlihatkan kepada pengguna atau user.

Tahap-tahap tersebut bersifat interaktif dimana pengguna atau user terlibat langsung atau dengan perantaraan basis pengetahuan (*knowledge base*) yang terintegrasi didalam sistem. Pola-pola yang menarik disajikan kepada pengguna dan disimpan sebagai pengetahuan baru didalam basis pengetahuan. Dari tahapan diatas dapat diketahui bahwa data mining hanya merupakan satu bagain langkah dari keseluruhan proses KDD.



Gambar 2.1 Langkah-langkah Proses KDD

2.2 Masalah-masalah Dalam Data Mining

Banyak kasus dalam kehidupan sehari-hari yang memakai teknik-teknik data mining yang bisa kita pelajari. Contoh-contoh berikut ini memperlihatkan masalah-masalah dalam data mining :

1. Memprediksi apakah akan terjadi tornado berdasarkan informasi dari sebuah radar tentang kondisi angin dan kondisi atmosfir yang lain.
2. Dalam hal orang yang meminta hutang ke suatu bank. Haruskah suatu bank menyetujui hutang tersebut? Orang yang punya sejarah paling bagus biasanya tidak perlu hutang, dan orang yang mempunyai sejarah paling buruk biasanya tidak akan membayar hutang. Customer yang terbaik adalah yang ditangah-tengah.
3. Menemukan kelompok customer dan mempergunakan untuk target pemasaran dan re-organization.
4. Barang apa yang biasanya dibeli oleh customer supermarket ketika dia membeli diaper bayi? Bagaimana manajemen supermarket memberi respon setelah mengetahui pola pembelian customer?

Tentu saja masih banyak lagi contoh-contoh dari berbagai bidang yang bisa dimasukkan atau bisa diselesaikan dengan teknik-teknik data mining. Teknik-teknik belajar (*learning*) memegang peran kunci dalam masalah-masalah di atas. Masalah-masalah yang sesuai untuk diselesaikan dengan teknik data mining menurut Piatesky dan Shapiro (2006), yaitu :

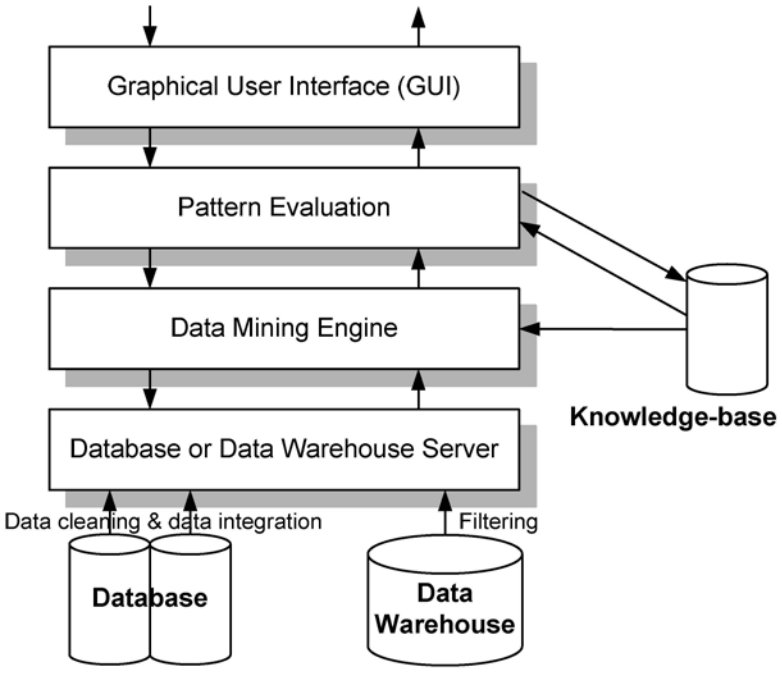
1. Memerlukan keputusan yang bersifat *knowledge-based*
2. Mempunyai lingkungan yang berubah
3. Metode yang ada sekarang bersifat sub-optimal
4. Tersedia data yang bisa diakses, cukup dan relevan
5. Memberikan keuntungan yang tinggi jika keputusan yang diambil tepat

2.3 Sistem Data Mining

Umumnya sistem data mining terdiri dari komponen-komponen berikut :

1. Database, data *warehouse*, terdiri dari satu atau beberapa database, data *warehouse*, atau data dalam bentuk lain. Pembersihan data dan integrasi data dilakukan terhadap data tersebut.

2. Database atau data *warehouse server*, bertanggung jawab terhadap pencarian data yang relevan sesuai dengan yang diinginkan pengguna atau user.
3. Basis Pengetahuan (*Knowledge Base*), merupakan basis pengetahuan yang digunakan sebagai panduan dalam pencarian pola.
4. Data mining engine, merupakan bagian penting dari sistem dan idealnya terdiri dari kumpulan modul-modul fungsi yang digunakan dalam proses karakterisasi (*characterization*), klasifikasi (*classification*) dan analisis kluster (*cluster analysis*).
5. Evaluasi Pola (*Pattern evaluation*), komponen ini pada umumnya berinteraksi dengan modul-modul data mining.
6. Antarmuka (*Graphical User Interface*), merupakan modul komunikasi antara pengguna atau user dengan sistem yang memungkinkan pengguna berinteraksi terhadap sistem untuk menentukan proses data mining itu sendiri.



Gambar 2.2 Arsitektur Data Mining

2.4 Data Pada Data Mining

Pada prinsipnya data mining dapat diaplikasikan pada berbagai jenis data yaitu pada database *relasional*, data *warehouse*, database *transaksional*, *advance database system*, *flat-files* dan *world wide web*. Tetapi teknik data mining pada setiap jenis data berbeda.

2.4.1 Database *Relasional*

Suatu database *relasional* terdiri dari kumpulan tabel-tabel, dimana masing-masing tabel mempunyai nama yang unik. Setiap tabel terdiri dari kumpulan atribut (kolom atau *field*) dan biasanya menyimpan banyak *record* atau *tupel*. Setiap *record* pada *database relasional* mewakili suatu objek yang diidentifikasi oleh sebuah kunci yang unik.

Database *relasional* dapat diakses melalui *query* yang ditulis dalam bahasa *query* seperti *SQL*, atau melalui anatarmuka grafis yang tersedia. Database *relasioanal* merupakan database yang sangat popular dan kaya akan informasi yang tersimpan didalamnya oleh karena itu data mining banyak diterapkan pada database *relasional*.

2.4.2 Data *Warehouse*

Suatu data *warehouse* mengkonsolidasikan data dari berbagai sumber data dengan aturan penamaan, ukuran atribut fisik, serta semantik yang konsisten. Data *warehouse* dibangun dari suatu proses pembersihan data, transformasi data dan integrasi data, pengambilan data dan penyegaran data yang dilakukan secara periodik.

Data *warehouse* biasanya dimodelkan dengan suatu struktur database multidimensional yang disebut data *cube*. Dimensi pada data-*cube* dapat bertingkat untuk memudahkan dalam mendapatkan rangkuman informasi dari tingkatan dimensi yang lebih luas atau umum sampai kepada tingkatan informasi yang lebih sempit atau khusus dengan operasi *roll-up* atau sebaliknya dengan operasi *drill-down*. Akan tetapi untuk mendapatkan informasi yang tidak diketahui secara eksplisit diperlukan satu tahap lagi yaitu aplikasi teknik data mining.

2.4.3 Database *Transaksional*

Suatu database *transaksional* terdiri dari sebuah file dimana tiap-tiap *record* mewakili suatu transaksi. Suatu transaksi umumnya mengandung suatu nomer identitas transaksi yang bersifat unik, dan suatu daftar item-item yang membentuk transaksi tersebut. Suatu database *transaksional* dimungkinkan untuk mempunyai tabel-tabel tambahan yang saling terkait yang berisikan informasi-informasi yang saling berhubungan.

2.4.4 *Advance Database System*

Aplikasi database terbaru mampu untuk menangani data spasial (data peta), data-data desain (desain dari sebuah gedung), *hypertext* dan data multimedia (termasuk didalamnya data teks, gambar, *video*, dan *audio*) dan data pada *world wide web*. Aplikasi-aplikasi tersebut membutuhkan struktur data yang efisien dan metode-metode yang skalabilitas untuk menangani struktur objek yang komplek.

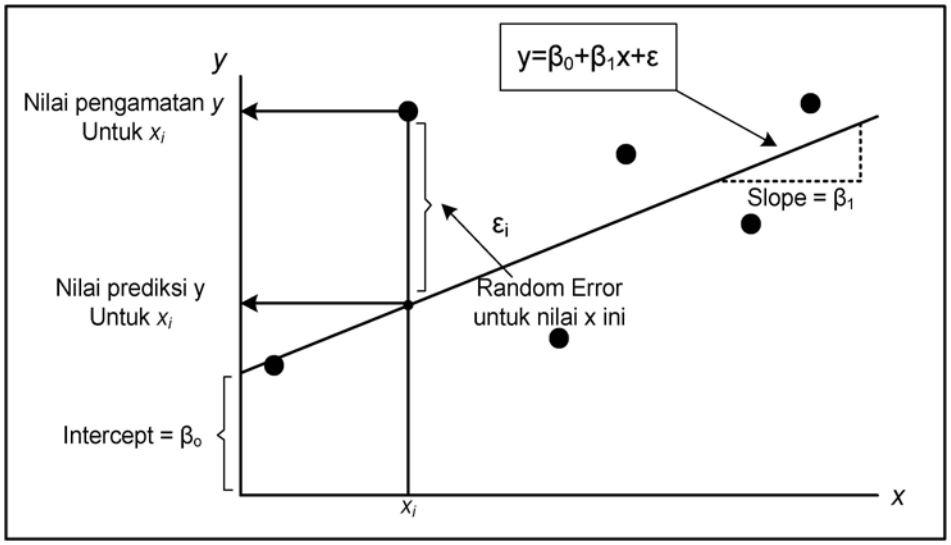
2.5 Teknik Data Mining

Dari definisi data mining yang luas, terdapat banyak jenis teknik analisa yang dapat digolongkan dalam data mining. Dalam penelitian tugas akhir ini teknik analisa yang digunakan yaitu teknik regresi linier. Dibawah ini akan dijelasakan mengenai teknik analisa regresi linier.

2.5.1 Regresi Linier

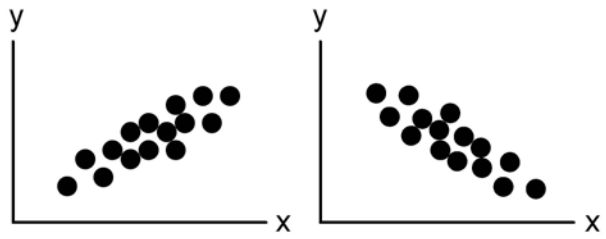
Analisis regresi adalah teknik statistik untuk pemodelan dan investigasi hubungan dua atau lebih variabel. Yang sering dipakai dan paling sederhana adalah regresi linier sederhana. Dalam analisis regresi ada satu atau lebih *variabel independent/prediktor* yang biasa diwakili dengan notasi x dan satu *variabel respon* yang diwakili dengan notasi y. Sesuai namanya, hubungan antara dua variabel yang bersifat linier. Gambar 2.4 dan gambar 2.5 memberi ilustrasi bagaimana hubungan dua variabel ini bersifat linier dan tidak linier. Gambar 2.4 menunjukkan hubungan linier dua variabel. Garis regresi linier akan sangat sesuai untuk mewakili hubungan dua variabel seperti ini. Gambar 2.5 menunjukkan

hubungan tidak linier antara dua variabel. Pendekatan regresi linier kurang sesuai untuk mewakili hubungan dua variabel seperti gambar 2.4 ini. Dalam regresi linier sederhana hanya ada satu *variabel independent/prediktor* dan satu *variabel respon*. Jika *variabel independen*-nya *x* dan *variabel respon* adalah *y* maka model regresi linier sederhana untuk populasi adalah :

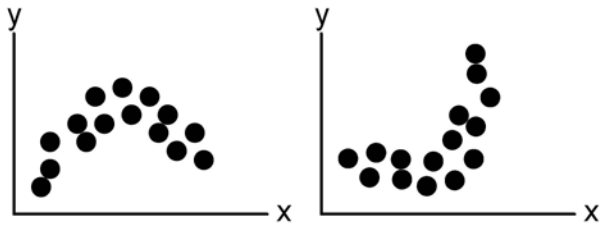


Gambar 2.3 Geometri garis regresi linier

Gambar 2.3 menunjukkan deskripsi geometris dari garis regresi linier dalam dua dimensi.



Gambar 2.4 *Hubungan dua variable yang bersifat linier*



Gambar 2.5 *Hubungan dua variabel yang bersifat tidak linier*

Dalam gambar 2.4 sebelah kiri, slope β_1 bernilai positif dan sebelah kanan β_1 bernilai negatif.

Prediksi nilai dengan pendekatan regresi linier sederhana, didapatkan dari rumus dibawah ini :

$$\bar{y} = b_0 + b_1 x \qquad \text{Rumus (2.1)}$$

Untuk keperluan ini, sebaiknya data hasil pengamatan dicatat dalam bentuk seperti di bawah ini :

Tabel 2.1 Tabel Pengamatan

Variabel tak bebas (y_i)	**Variabel bebas (x_i)**
y_1	x_1
y_2	x_2
.	.
.	.
y_n	x_n

Pada tabel 2.1 terdapat pasangan antara *x* dan *y* dan n, seperti biasa, menyatakan ukuran sampel. Koefisien-koefisien regresi b_0 dan b_1 untuk regresi linier, ternyata dapat dihitung dengan rumus :

$$b_0 = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} \qquad \text{Rumus (2.2)}$$

Jika terlebih dahulu dihitung koefisien b_1, maka koefisien b_0 dapat pula ditentukan oleh rumus:

$$b_0 = \overline{y} - b_1 \overline{x}$$

$$e = error = y - \bar{y} \qquad \text{Rumus (2.3)}$$

Dengan $\bar{x}$ dan $\bar{y}$ masing-masing rata-rata untuk variable-variabel x dan y. Rumus-rumus di atas dipakai untuk menetukan koefisien-koefisien regresi y dan x. Untuk koefisien-koefisien regresi x atas y, rumus yang sama digunakan tetapi harus dipertukarkan tempat untuk simbol-simbol x dan y. Dibawah ini merupakan rumus untuk mencari jumlah kuadrat regresi atau *Sum of Square* (SS) dengan rumus:

$$SS_y = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SS_x = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

Rumus (2.4)

2.5.2 Berbagai Varians Sehubungan Dengan Regresi Linier Sederhana

Untuk analisis selanjutnya tentang regresi linier sederhana beberapa asumsi harus diambil. Pertama, mengingat hasil pengamatan variabel takbebas y belum tentu sama besarnya dengan harga diharapkan, yakni $\bar{y}$ yang didapat dari regresi hasil pengamatan, maka terjadi perbedaan $e = y - \bar{y}$, biasa disebut kekeliruan prediksi atau galat prediksi (*Error*). Dalam populasi, galat prediksi ini dimisalkan berbentuk variabel acak yang mengikuti distribusi normal dengan rata-rata nol dan varians σ_e^2. Tentu saja sudah jelas bahwa kita juga memisalkan tidak terjadi kekeliruan atas pengamatan variabel bebas x.

Asumsi kedua yang diambil adalah bahwa untuk setiap harga x yang diberikan, variabel tak bebas y independen dan berdistribusi normal dengan rata-rata $(\theta_1 + \theta_2 x)$ dan varians $\sigma^2 y.x$. Varians $\sigma^2 y.x$ dimisalkan sama untuk setiap x dan karenanya dapat dinyatakan oleh σ_e^2 yang biasa pula dinamakan varians kekeliruan taksiran sedangkan $\sigma\ y.x$ dikenal dengan kekeliruan baku taksiran.

Berpegang kepada asumsi-asumsi diatas, maka varians σ_e^2 ditaksir oleh rata-rata kuadrat penyimpangan sekitar regresi atau disebut juga rata-rata kuadrat residu, dinyatakan oleh varians *Mean Square Error* (MSE) dengan rumus:

$$MSE = \frac{SSE}{n-2}$$ Rumus (2.5)

Dengan :

$$SSE = \sum (y - \bar{y})^2 = SS_y - \frac{(SS_{xy})^2}{SS_x} = SS_y - b_1 SS_{xy}$$ Rumus (2.6)

Dengan SSE yaitu jumlah kuadrat residu atau *Sum of Square for Error* (SSE), SS_y yaitu jumlah kuadrat regresi y , SS_{xy} yaitu jumlah kuadrat xy dan n yaitu ukuran sampel.

2.5.3 Interval Kepercayaan Sehubungan Dengan Regresi Linier

Kita lihat bahwa regresi linier populasi telah ditaksir oleh regresi linier sampel $\bar{y} = b_0 + b_1 x$ dengan koefisien-koefisien b_0 dan b_1. jadi nampak bahwa b_0 dan b_1 masing-masing merupakan titik taksiran untuk β_o dan β_1. maka berbagai interval taksiran sehubungan dengan regresi linier, termasuk untuk β_0 dan β_1 dapat ditentukan.

A (1-α) 100% interval kepercayaan untuk β_0 yaitu :

$$b_0 \pm t_{(\alpha/2,n-2)} s(b_0)$$ Rumus (2.7)

A (1-α) 100% interval kepercayaan untuk β_1 yaitu :

$$b_1 \pm t_{(\alpha/2,n-2)} s(b_1)$$ Rumus (2.8)

Dengan :

$s(b_0)$ atau yang disebut *standar error* terhadap b_0, yaitu :

$$s(b_0) = \frac{s\sqrt{\sum x^2}}{\sqrt{nSS_x}}$$ Rumus (2.9)

$s(b_1)$ atau yang disebut *standar error* terhadap b_1, yaitu :

$$s(b_1) = \frac{s}{\sqrt{SS_x}}$$ Rumus (2.10)

dan s atau yang disebut *standar error of estimate*, yaitu :

$$s = \sqrt{MSE}$$ Rumus (2.11)

2.5.4 Uji Signifikan Dan Tabel *Analysis of Variance* (ANOVA)

Pada sub-bab ini akan membahas tentang melakukan kriteria uji signifikan dengan menghitung F_{hitung} dan membandingkan hasilnya dengan hasil perhitungan F_{tabel}. Berikut rumus mencari F_{hitung}:

$$F_{(1,n-2)} = \frac{MSR}{MSE} \qquad \text{Rumus (2.12)}$$

Dengan :

- *Mean Square Regression* (MSR) atau rata-rata kuadrat regresi, dengan rumus :

$$MSR = \frac{SSR}{1} \quad \text{dan} \quad SSR = b_1 SS_{xy} \qquad \text{Rumus (2.13)}$$

- *Mean Square Error* (MSE) atau rata-rata kuadrat penyimpangan sekitar regresi. Yang rumusnya telah dijelaskan diatas (Rumus 2-5).

Setelah ditemukan hasil dari F_{hitung} maka selanjutnya kita melaukan uji signifikan dengan membandingkan F_{hitung} dengan F_{tabel} , berikut kaidah pengujian signifikan :

Jika $F_{hitung} \geq F_{tabel}$, maka tolak H_0 (Signifikan)

Jika $F_{hitung} \leq F_{tabel}$, maka tolak H_a (Tidak Signifikan)

Setelah kita menghitung seluruh perhitungan yang ada di proses regresi, maka kita tinggal menyusunnya dalam tabel *analysis of variance* (ANOVA). Berikut skema dari tabel ANOVA.

Tabel 2.2 Tabel ANOVA pada Regresi

Source of Variation	Sum of Squares (SS)	Degress of Freedom (DF)	Mean Square (MS)	F Ratio
Regression	$SSR = b_1 SS_{xy}$	1	$MSR = \frac{SSR}{1}$	$F_{(1,n-2)} = \frac{MSR}{MSE}$
Error	$SSE = SS_y - b_1 SS_{xy}$	n - 2	$MSE = \frac{SSE}{n-2}$	
Total	SST = SSy	n - 1		

2.5.5 Korelasi Pearson

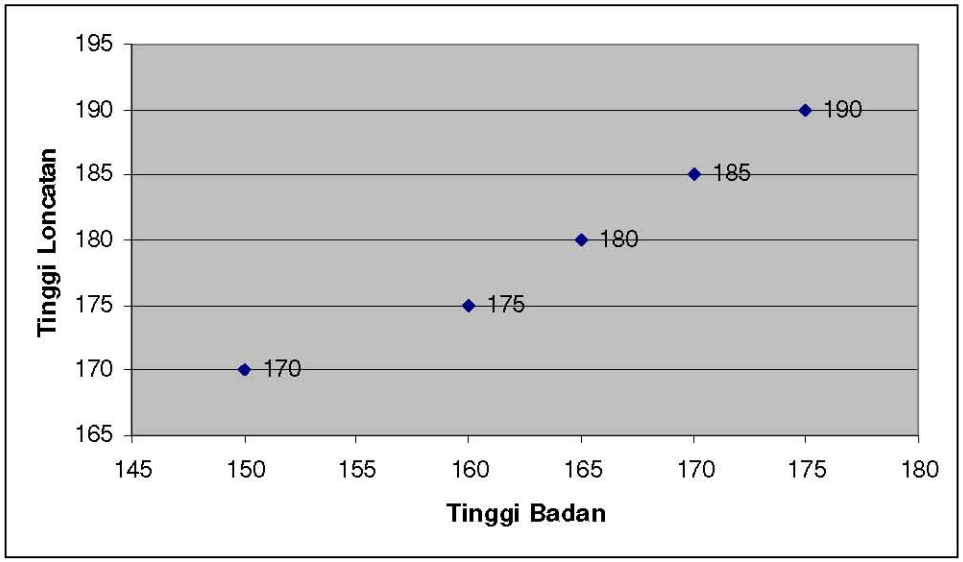
Korelasi merupakan suatu hubungan antara satu variabel dengan variabel lainnya. Hubungan antara variabel tersebut bisa secara korelasional dan bisa juga secara kausal. Jika hubungan tersebut tidak menunjukkan sifat sebab akibat, maka korelasi tersebut dikatakan korelasional, artinya sifat hubungan variabel satu dengan variabel lainnya tidak jelas mana variabel sebab dan mana variabel akibat. Sebaliknya, jika hubungan tersebut menunjukkan sifat sebab akibat, maka korelasinya dikatakan kausal, artinya jika variabel yang satu merupakan sebab, maka variabel lainnya merupakan akibat.

Pembahasan korelasi minimal menyangkut dua kelompok nilai atau variabel. Variabel-variabel tersebut bisa berasal dari subjek penelitian yang sama. Sebelum masuk perhitungan pada korelasi, lebih dulu kita pahami konsep korelasi melalui suatu diagram sederhana yang disusun berdasarkan data sederhana dengan contoh data dibawah ini :

Tabel 2.3 Data Pengukuran Tinggi Badan dan Tinggi Loncatan

Mahasiswa	A	B	C	D	E
Tinggi Badan	150	160	165	170	175
Tinggi Loncatan	170	175	180	185	190

Hasil pengukuran tersebut jika dibuat grafik, hasilnya :



Gambar 2.6 Grafik Pengukuran Tinggi Badan dan Tinggi Loncatan

Apabila antara titik satu dengan titik yang lainnya (yang berdekatan) dihubungkan, maka akan terbentuk suatu garis yang berkemungkinan lurus, melenngkung dan mungkin tidak berketentuan bentuknya (jika n banyak).

Walaupun kita mengalami kesukaran dalam menarik garis yang dapat menghubungkan antar titik dengan jarak terdekat, tetapi kita dapat membuat garis secara intuisi yang mempunyai rata-rata jarak terdekat dengan seluruh titik yang ada. Pembuatan garis tersebut tidak cukup akurat jika dibuat berdasarkan intuisi belaka (lebih-lebih jika titik-titik yang tersebar banyak sekali).

Beberapa penyebaran yang mungkin terjadi sebagai berikut :

1. Memanjang tegak (mendekati sejajar dengan sumbu vertikal).
2. Memanjang rebah (mendekati sejajar dengan sumbu horizontal).
3. Memanjang ke kanan atas.
4. Memanjang ke kanan bawah.
5. Bulat tidak menunjukkan arah pasti.

Korelasi Pearson adalah korelasi yang sering digunakan oleh peneliti, terutama peneliti yang mempunyai data-data interval. Sebelum kita mempergunakan korelasi ini terlebih dahulu kita harus memperhatikan data yang terkumpul, apakah memenuhi persyaratan yang diminta oleh rumus korelasi ini.

Adapun beberapa persyaratan yang harus dipenuhi apabila kita menggunakan rumus ini adalah :

1. Pengambilan sampel dari populasi harus *random* (acak).
2. Data yang dicari korelasinya harus berskala interval atau ratio.
3. Variasi skor kedua variabel yang akan dicari korelasinya harus sama.
4. Distribusi skor variabel yang dicari korelasinya hendaknya merupakan distribusi unimodal.
5. Hubungan antara variabel *x* dan *y* hendaknya linier.

Korelasi Pearson dapat dihitung dengan rumus dibawah ini :

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \qquad \text{Rumus (2.14)}$$

Atau :

$$r = \frac{SSxy}{\sqrt{SSxSSy}} \qquad \text{Rumus (2.15)}$$

Hasil perhitungan korelasi pada dasarnya dapat dikelompokkan menjadi tiga kelompok besar :

1. Korelasi positif kuat, apabila hasil perhitungan korelasi mendekati +1 atau sama dengan +1. ini berarti bahwa setiap kenaikan skor/nilai pada variabel *x* akan diikuti dengan kenaikan skor/nilai variabel *y*. Sebaliknya, jika variabel *x* mengalami penurunan, maka akan diikuti dengan penurunan variabel *y*.
2. Korelasi negatif kuat, apabila hasil perhitungan korelasi mendekati -1 atau sama dengan -1. ini berarti bahwa setiap kenaikan skor/nilai pada variabel *x* akan diikuti dengan penurunan skor/nilai pada variabel *y*. Sebaliknya, apabila skor/nilai dari variabel *x* turun, maka skor/nilai dari variabel *y* akan naik.
3. Tidak ada korelasi, apabila hasil perhitungan korelasi (mendekati 0 atau sama dengan 0). Hal ini berarti bahwa naik turunnya skor/nilai satu variabel tidak mempunyai kaitan dengan naik turunnya skor/nilai variabel yang lainnya. Apabila skor/nilai variabel *x* naik tidak selalu diikuti dengan naik atau turunnya skor/nilai variabel *y*, demikian juga sebaliknya.

Hasil perhitungan korelasi bergerak antara -1 sampai dengan +1. Jadi, kalau ada hasil perhitungan korelasi lebih besar (>) daripada +1 atau kurang dari (<) -1, maka perhitungan tersebut jelas salah. Korelasi pearson hanya diterapkan untuk data yang berskala interval dan ratio.

Untuk menentukan kerataan hubungan dapat digunakan kriteria Guilford, yaitu :

Tabel 2.4 Tabel Kerataan Hubungan

Interval Koefisien	Tafsirannya
0.00-0.20	Hubungan yang sangat kecil dan bisa diabaikan
0.20-0.40	Hubungan yang sangat kecil (tidak erat)
0.40-0.70	Hubungan yang cukup
0.70-0.90	Hubungan yang erat
0.70-1.00	Hubungan yang sangat erat

2.5.6 Pengujian Signifikansi Korelasi

Langkah awal dalam pengujian di sini juga menyusun hipotesis nol dan hipotesis alternatif. Baru kemudian hasil r hitung kita bandingkan dengan hasil r tabel dari tabel r Pearson. Apabila kita menggunakan tabel r Pearson, maka hipotesis nol yang mengatakan tidak ada korelasi (r = 0) ditolak jika hasil perhitungan r > dari pada r tabel, demikian pula sebaliknya apabila r hitung ternyata lebih kecil < dari pada r tabel, maka kita akan menerima H_0 yang menyatakan bahwa dua variabel yang dicari hitungannya nyata-nyata tidak berkorelasi. Untuk lebih jelasnya bisa kita lihat kriteria signifikan sebagai berikut:

- Jika $r_{hitung} \geq r_{tabel}$ maka H_0 ada didaerah penolakan, berarti H_a diterima artinya antara variabel *x* dan *y* ada hubungannya.
- Jika $r_{hitung} \leq r_{tabel}$ maka H_0 ada didaerah penerimaan, berarti H_a ditolak artinya antara variabel *x* dan *y* tidak hubungannya.

2.5.7 Analisis Koefisien Determinasi

Analisis koefisien determinasi adalah menunjukkan seberapa besar pengaruh antar kedua variabel yang diteliti, maka dihitung Koefisien Determinasi (KD) dengan asumsi dasar faktor-faktor lain diluar variabel dianggap tetap atau konstan, koefisien diantara lain $-1 \leq r \geq +1$, tanda (-) berlawanan arah, sedangkan tanda (+) menunjukkan searah. Selanjutnya untuk mengetahui seberapa besar faktor yang berperan antara variabel *x* terhadap variabel *y*, maka hubungan atau pengaruh dihitung koefisien determinasinya dengan rumus :

$$KD = r^2 * 100\% \qquad \text{Rumus (2.16)}$$

2.5.8 Contoh Kasus Regresi Linier

Berikut ini akan dijelaskan contoh kasus dari regresi linier yang diambil dari buku Com*plete Business Statistics Sixth Edition* halaman 437.

Contoh:

Data berikut melukiskan hasil pengamatan pada suatu perusahaan travel mengenai jarak yang ditempuh dalam suatu perjalanan (miles) berpengaruh terhadap beban biaya dalam tiap perjalanan (dollars).

Tabel 2.5 Tabel Pengamatan

Miles	Dollars
1211	1802
1345	2405
1422	2005
1687	2511
1849	2332
2026	2305
2133	3016
2253	3385
2400	3090
2468	3694
2699	3371
2806	3998
3082	3555
3209	4692
3466	4244
3643	5298
3852	4801
4033	5147
4267	5738
4498	6420
4533	6059
4804	6426
5090	6321
5233	7026
5439	6964

Agar rumus yang telah dijelaskan sebelumnya dapat digunakan, kita hitung satuan-satuan yang diperlukan dan sebaiknya data disusun seperti dalam daftar berikut.

Tabel 2.6 Nilai-nilai Untuk Menghitung Koefisien Regresi Linier

No.	Miles X	Dollars Y	X^2	Y^2	XY
1	1211	1802	1466521	3247204	2182222
2	1345	2405	1809025	5784025	3234725
3	1422	2005	2022084	4020025	2851110
4	1687	2511	2845969	6305121	4236057
5	1849	2332	3418801	5438224	4311868
6	2026	2305	4104676	5313025	4669930

7	2133	3016	4549689	9096256	6433128
8	2253	3385	5076009	11458225	7626405
9	2400	3090	5760000	9548100	7416000
10	2468	3694	6091024	13645636	9116792
11	2699	3371	7284601	11363641	9098329
12	2806	3998	7873636	15984004	11218388
13	3082	3555	9498724	12638025	10956510
14	3209	4692	10297681	22014864	15056628
15	3466	4244	12013156	18011536	14709704
16	3643	5298	13271449	28068804	19300614
17	3852	4801	14837904	23049601	18493452
18	4033	5147	16265089	26491609	20757851
19	4267	5738	18207289	32924644	24484046
20	4498	6420	20232004	41216400	28877160
21	4533	6059	20548089	36711481	27465447
22	4804	6426	23078416	41293476	30870504
23	5090	6321	25908100	39955041	32173890
24	5233	7026	27384289	49364676	36767058
25	5439	6964	29582721	48497296	37877196
n	**$\sum X$**	**$\sum Y$**	**$\sum X^2$**	**$\sum Y^2$**	**$\sum XY$**
25	79448	106605	293426946	521440939	390185014

Langkah pertama untuk menjawab uji regresi linier sederhana yaitu dengan membuat H_a dan H_0 dalam bentuk kalimat :

H_a : Terdapat pengaruh yang signifikan antara *miles* terhadap *dollars*.

H_0 : Tidak terdapat pengaruh yang signifikan antara *miles* terhadap *dollars*.

Setelah membuat H_a dan H_0 dalam bentuk kalimat, selanjutnya membuat H_a dan H_0 dalam bentuk statistik :

$H_a : r \neq 0$ $H_0 : r = 0$

Masukkan angka-angka statistik dari tabel 2.4 pada persamaan rumus 2.2, sehingga kita peroleh harga-harga :

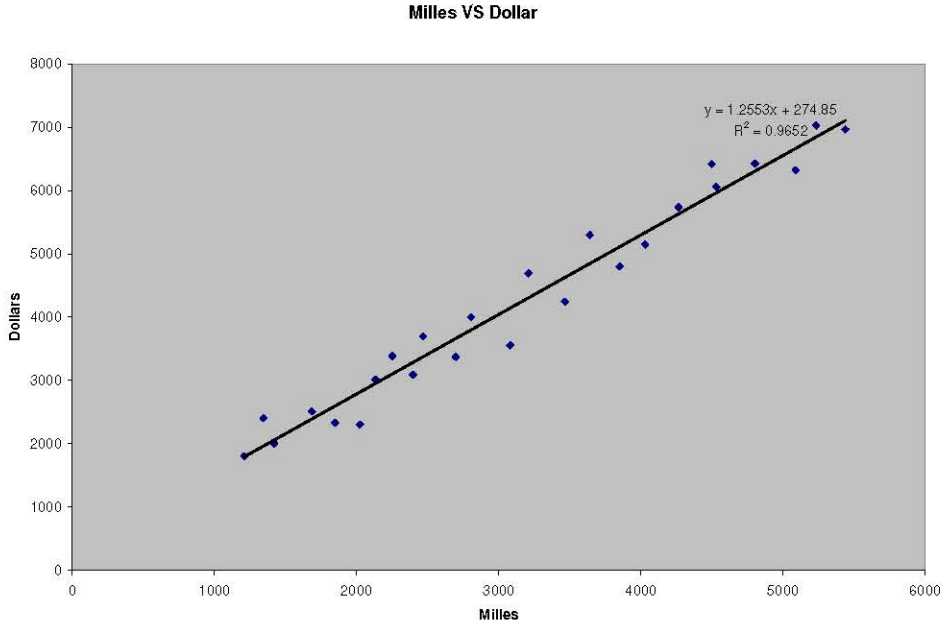
$$b_0 = \frac{(106605)(293426956) - (79448)(390185014)}{25(293426946) - (79448)^2} = 274.8496867$$

$$b_1 = \frac{25(390185014) - (79448)(106605)}{25(293426946) - (79448)^2} = 1.255333776$$

Dengan demikian, persamaan regresi linier y atas x untuk kasus di atas adalah :

$$\bar{y} = 274.85 + 1.26x$$

Variabel takbebas y dalam regresi telah dinyatakan oleh simbol $\bar{y}$ (baca : ye topi) untuk menyatakan bahwa kita berhadapan dengan y yang didapat dari regresi dan untuk membedakannya dengan y dari hasil pengamatan.



Gambar 2.7 Grafik Regresi

Grafik regresi yang di dapat dari persamaan regresi linier diatas dapat dilihat dalam gambar 2.7. Koefisien b_1 dinamakan koefisien arah regresi linier dan menyatakan perubahan rata-rata variabel y untuk setiap perubahan variabel x sebesar satu unit. Perubahan ini merupakan pertambahan apabila b_1 bertanda positif dan penurunan atau pengurangan jika bertanda negatif. Demikianlah misalnya, untuk contoh kita b_1=1.26 bertanda positif, sehingga kita dapat mengatakan bahwa jika x (*Miles*) bertambah, maka rata-rata beban biaya (y) bertambah dengan 274.85 per *Miles*.

Setelah kita mendapatkan persamaan regresi linier dan grafik regresi, maka langkah selanjutnya kita melakukan uji signifikan dengan langkah-langkah berikut :

- Menghitung jumlah kuadrat regresi atau *Sum of Square* (SS), sesuai dengan rumus 2.4 :

$$SS_y = 521440939 - \frac{(106605)^2}{25} = 66855898$$
$$SS_x = 293426946 - \frac{(79448)^2}{25} = 40947557.84$$
$$SS_{xy} = 390185014 - \frac{(79448)(106605)}{25} = 51402852.4$$

- Menghitung jumlah kuadrat residu atau *Sum of Square for Error* (SSE), sesuai dengan rumus 2.6 :

$$SSE = 66855898 - (1.255333776)(51402852.4) = 2328161.2$$

- Menghitung rata-rata kuadrat residu atau *Mean Square Error* (MSE), sesuai dengan rumus 2.5 :

$$MSE = \frac{2328161.2}{25-2} = 101224.4$$

- Menghitung *standar error of estimate* (s), sesuai dengan rumus 2-11 :

$$s = \sqrt{101224.4} = 318.1578225$$

- Menghitung *standar error* terhadap b_0, sesuai dengan rumus 2.9 :

$$s(b_0) = \frac{318.1578225\sqrt{293426946}}{\sqrt{25 * 40947557.84}} = 170.338$$

- Menghitung *standar error* terhadap b_1, sesuai dengan rumus 2.10 :

$$s(b_1) = \frac{318.1578225}{\sqrt{40947557.84}} = 0.04972$$

- Menghitung interval kepercayaan untuk β_0, sesuai dengan rumus 2.7 :

  $t(\alpha/2,n-2) = t(5/2,25-2) = t(0.25,23) = 2.069$ (diperoleh dari tabel nilai kritik sebaran *t*)

$$274.8496866 \pm 2.069(170.338)$$

1-α	(1-α) C.I for β_0
95%	274.8497 + or - 352.3681

- Menghitung interval kepercayaan untuk β_1, sesuai dengan rumus 2.8 :

  $t(\alpha/2,n-2) = t(5/2,25-2) = t(0.25,23) = 2.069$ (diperoleh dari tabel nilai kritik sebaran *t*)

  $$1.255333776 \pm 2.069(0.04972)$$

1-α	(1-α) C.I for β_1
95%	1.2553334 + or - 0.102853

- Menghitung rata-rata kuadrat regresi atau *Mean Square Regression* (MSR), sesuai dengan rumus 2.13 :

  $$SSR = 1.255333776 * 51402852.4 = 64527736.8$$
  $$MSR = \frac{64527736.8}{1} = 64527736.8$$

- Menghitung F_{hitung}, sesuai dengan rumus 2.12 :

  $$F_{hitung} = \frac{64527736.8}{101224.4} = 637.47$$

- Menghitung F_{tabel} :

  Menggunakan $\alpha = 0.01$

  $$F_{tabel} = F(1, n-2)$$
  $$F_{tabel} = F(1,23)$$
  $$F_{tabel} = 7.88$$

  Setelah melakukan perhitungan uji signifikan, maka didapatkan nilai F_{hitung} dan F_{tabel}. Sesuai dengan kaidah uji signifikan maka $F_{hitung} = 637.47$ dan $F_{tabel} = 7.88$ dapat diambil kesimpulan bahwa $F_{hitung} \geq F_{tabel}$ sehingga H_0 ditolak atau signifikan. Sehingga hipotesis awal (H_a) bisa diterima.

- Menghitung korelasi pearson (*r*), sesuai dengan rumus 2.15 :

  $$r = \frac{51402852.4}{\sqrt{40947557.84 * 66855898}} = 0.9824$$

- Menghitung r_{tabel} :

  Menggunakan $\alpha = 0.01$

  $$r_{tabel} = r\ (n-2,0.01)$$
  $$r_{tabel} = r\ (23,0.01)$$
  $$r_{tabel} = 0.505$$

Dengan demikian, maka kita menolak hipotesis nol yang berarti variabel *miles* mempunyai hubungan dengan variabel *dollars*. Hal ini dikarenakan $r_{hitung} > r_{tabel}$, maka jelas korelasi tersebut signifikan.

- Menghitung Koefisien Determinasi (KD), sesuai dengan rumus 2.16 :

  $KD = 0.9824^2 * 100\%$

  $KD = 0.965 * 100\ \% = 96.5\ \%$

  Dengan demikian besaran pengaruh antara variabel *x* (*milles*) terhadap variabel *y* (*dollars*) yaitu sebesar 96.5 %.

Setelah melakukan proses perhitungan analisis regresi, maka beberapa nilai yang telah dapatkan bisa ditampilkan dengan menggunakan tabel ANOVA sesuai dengan penjelasan sebelumnya. Dibawah ini merupakan tabel ANOVA dari beberapa hasil perhitungan yang telah dilakukan dari tiap-tiap proses diatas.

Tabel 2.7 Tabel ANOVA Hasil Perhitungan

Source of Variation	Sum of Squares (SS)	Degress of Freedom (DF)	Mean Square (MS)	F Ratio
Regression	64527736.8	1	64527736.8	637.47
Error	2328161.2	23	101224.4	
Total	66855898	24		

BAB III
ANALISIS DAN PERANCANGAN

Dalam perancangan sistem berbasis komputer, analisis masalah memegang peranan penting dalam membuat rincian aplikasi baru. Analisis masalah merupakan langkah pemahaman persoalan sebelum mengambil tindakan atau keputusan penyelesaian hasil akhir.

3.1 Perumusan Masalah

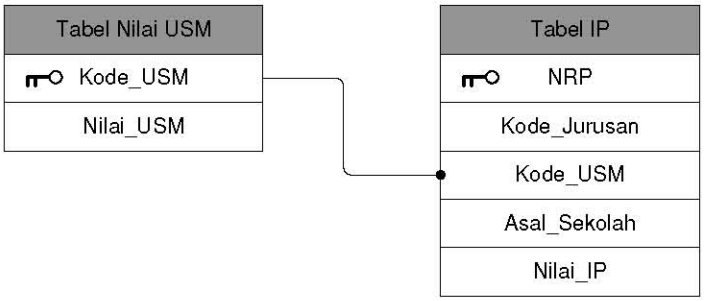
Pada tahap ini dilakukan eksplorasi lebih dalam dan menggali permasalahan yang ada terhadap sistem dan untuk mendefinisikan kebutuhan dari aplikasi yang akan dibangun.

3.1.1 Masalah Yang Dihadapi

Sering kali peneliti ingin melihat kondisi di waktu yang akan datang dengan suatu dasar keadaan sekarang atau ingin melihat kondisi di waktu yang lalu dengan dasar keadaan sekarang. Sifat melakukan prediksi atau taksiran mulai berkembang dalam dunia ekonomi, tetapi sekarang banyak dilakukan di dunia pendidikan. Dewasa ini, melakukan prediksi keadaan siswa untuk waktu yang akan datang merupakan kondisi yang sangat dibutuhkan dalam dunia pendidikan. Salah satunya dalam penelitian ini mencoba untuk memanfaatkan data yang telah ada di Universitas Widyatama, yaitu data Nilai Ujian Saringan Masuk (USM). Data Nilai USM di Universitas Widyatama selama ini belum dimanfaatkan, maka terjadi penumpukan data yang begitu besar dan menghasilkan kuburan data. Oleh karena itu, penelitian ini dilakukan untuk menggali informasi yang bisa di dapat dari data nilai USM mahasiswa baru. Sehingga dilakukan penerapan data mining dalam hal ini adalah mencari pola antara data nilai USM terhadap Indeks Prestasi (IP). Selain itu, penelitian ini pun mencoba untuk bagaimana membangun perangkat lunak yang bisa membantu pencaria pola antara data nilai USM terhadap IP dan membantu mentransformasikan data nilai USM menjadi suatu informasi yang berguna bagi *user*.

3.1.2 Inisialisasi Target Data

Database yang dimiliki secara garis besar yaitu tabel nilai Ujian Saringan Masuk (USM) dan tabel Indeks Prestasi (IP) mahasiswa. Struktur dari tabel-tabel yang akan digunakan dalam penelitian ini adalah sebagai berikut :



Gambar 3.1 Struktur Tabel

3.2 Analisis Kebutuhan Sistem

3.2.1 Analisis Data

Data yang tidak lengkap dan inkonsisten umumnya terjadi hampir pada setiap database, data yang tidak lengkap dapat terjadi karena berbagai macam sebab, dari atribut yang penting yang tidak terdapat dalam suatu database, sampai atribut dengan isi yang salah (*noisy data*). Demikian pula hanya yang terjadi pada database nilai USM banyak atribut yang tidak diperlukan, dan *noisy data* yang ada sehingga proses data *preprocessing* diperlukan sehingga database sesuai dengan ketentuan yang diperlukan sistem.

Data *preprocessing* merupakan hal yang penting dalam proses data mining, hal-hal yang termasuk didalamnya adalah :

3.2.1.1 Pembersihan Data (Data *Cleaning*)

Data *cleaning* diterapkan untuk menambahkan isi atribut yang hilang atau kosong, dan merubah data yang tidak konsisten. Tahapan dalam data *cleaning* yang diterapkan pada database yang akan dijadikan input bagi proses data mining itu sendiri. Ada beberapa teknik pembersihan data yang bisa kita terapkan pada database, yaitu:

1. Mengabaikan tupel, dilakukan bilamana label atau *value* dari suatu tupel hilang atau tidak ada. Metode ini sangatlah tidak efektif apabila terdapat banyak atribut dengan tupel-tupel kosong.
2. Menambahkan isi terhadap atribut yang kosong tersebut secara manual, namun pendekatan ini sangatlah memakan waktu dan tidak efektif bila diterapkan pada data yang sangat besar.
3. Menggunakan konstanta global untuk menambahkan isi atribut yang kosong.
4. Gunakan atribut rata-rata untuk mengisi atribut yang kosong. Sehingga data tersebut akan diisi dengan rata-rata dari data yang ada.

Dalam penelitian ini, teknik yang akan digunakan untuk pembersihan data yaitu menggunakan atribut rata-rata untuk mengisi atribut yang kosong. Sehingga data tersebut akan diisi dengan nilai rata-rata dari masing-masing variabel (teknik ke-4).

3.2.1.2 Transformasi Data (*Data Transformation*)

Dalam proses transformasi data, data ditransformasikan atau di konsolidasikan kedalam bentuk yang sesuai untuk proses data mining. Proses transformasi data yang diberlakukan terhadap penelitian ini yaitu dengan *generalisasi*, dimana data IP *range* nilainya disesuaikan dengan *range* nilai dari data nilai USM. Data nilai USM memiliki nilai maksimal yaitu 100, sedangkan data IP memiliki nilai maksimal yaitu 4. Untuk menyesuaikan *range* nilai antara data nilai IP dengan data nilai USM, maka harus dilakukan pengkalian terhadap data nilai IP dengan nilai 25. Sehingga didapatkan aturan sebagai berikut :

*Data nilai IP = Data nilai IP * 25*

3.2.1.3 Reduksi Data (*Data Reduction*)

Reduksi data dilakukan dengan menghilangkan atribut-atribut yang tidak diperlukan sehingga ukuran dari database menjadi kecil dan hanya menyertakan atribut-atribut yang diperlukan dalam proses data mining. Hal ini dikarenakan proses data mining terhadap data yang lebih kecil akan lebih efisien.

Dari proses data preprocessing tersebut melalui proses pemilihan atribut, maka dihasilkan tabel berikut :

Tabel 3.1 Tabel Hasil *Preprocessing*

Nilai USM	Nilai IP
85	82
79.14	81
78	79
75	46.75
71.90	71.5
…	…

3.2.2 Analisis Kebutuhan Input, Proses, dan Output

Untuk memudahkan implementasi dari sistem yang akan dibangun yaitu suatu *data mining engine* menggunakan teknik regresi adalah dengan mendeskripsikan terlebih dahulu tujuan atau *goal* yang akan dicapai dan menentukan kebutuhan-kebutuhan aplikasi yang akan dibangun.

Tujuan atau *goal* dari sistem yang akan dibangun adalah penerapan suatu teknik *data mining* menggunakan regresi. Kebutuhan aplikasi yang akan dikembangkan meliputi: *input, output,* operasi dan proses dengan menentukan kebutuhan spesifik sistem dalam hal:

1. Input-input yang diperlukan untuk menghasilkan output.

   Masukan atau *input* bagi aplikasi yang akan dibangun adalah data dalam bentuk suatu tabel dari suatu database yang telah melalui berbagai proses normalisasi data sesuai dengan kebutuhan aplikasi.

2. Operasi-operasi yang dilakukan untuk menghasilkan output.

   Operasi yang dihasilkan untuk menghasilkan output adalah suatu proses menggunakan teknik regresi.

3. Output yang harus dihasilkan

   Pola-pola yang ditemukan dari suatu aplikasi data mining dapat ditampilkan dalam berbagai bentuk, seperti grafik regresi atau bentuk visual yang lain. Berkaitan dengan aplikasi yang akan dibangun maka keluaran atau output yang akan dihasilkan adalah suatu hasil analisis regresi.

3.2.3 Analisis Kebutuhan Perangkat Keras

Komponen-komponen yang termasuk dalam perangkat keras untuk kebutuhan sistem adalah sebuah PC atau sebuah *workstation* dengan spesifikasi minimal yang digunakan pada saat implementasi adalah sebagai berikut :

1. Mikroprosesor : *Pentium* II atau lebih
2. Memori : 128 MB atau lebih
3. Media Penyimpanan : *Harddisk* 200 MB atau lebih

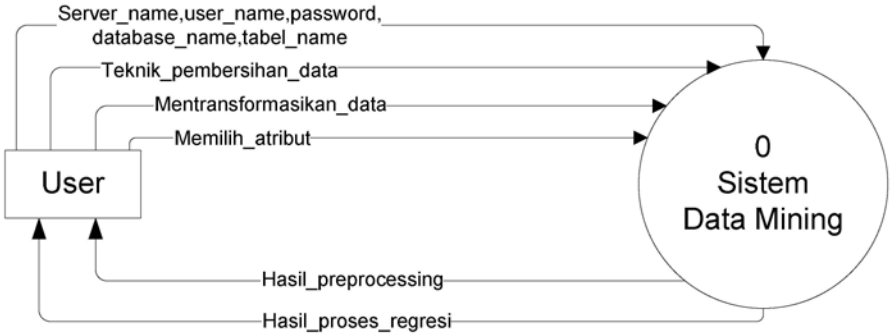
3.2.4 Analisis Kebutuhan Perangkat Lunak

Spesifikasi perangkat lunak yang digunakan untuk implementasi adalah sebagai berikut :

1. Sistem Operasi : *Microsoft Windows Server* 2003 / XP
2. Database System : *Microsoft SQL Server* 2000
3. Program Aplikasi : *Visual Studio .Net* 2005

3.3 Diagram Alur Data

Perancangan proses menggunakan *tools Data Flow Diagram* (DFD). Diagram konteks (*Context Diagram*) adalah DFD level 0 yang menggambarkan hubungan sistem dengan lingkungan *external*.

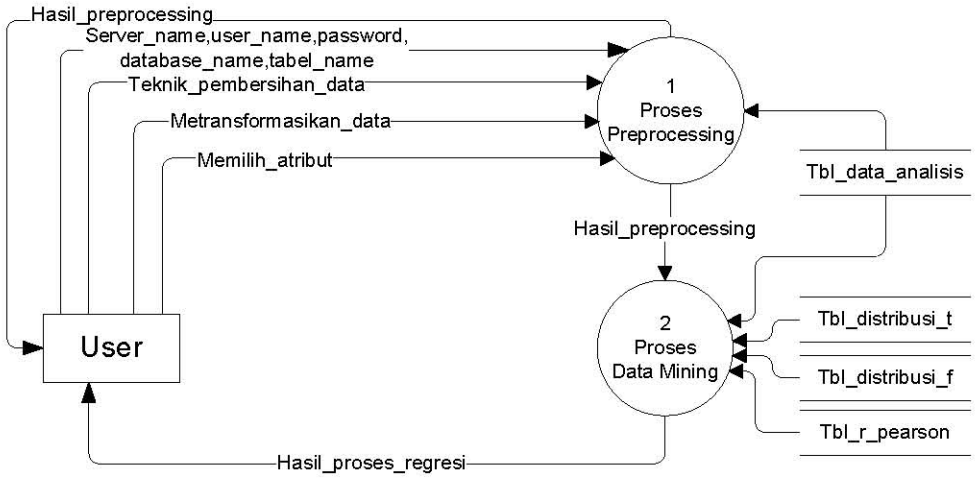


Gambar 3.2 *Data Context Diagram*

Context diagram diatas menggambarkan bahwa aplikasi data mining mempunyai pengguna yaitu user. User adalah orang yang mempunyai kewenangan atas sistem. Dalam sistem data mining ini, user diharuskan untuk memilih tabel data yang akan dianalisis dengan menggunakan teknik regresi. Setelah itu user-pun akan melalui proses preprocessing sebelumnya, untuk

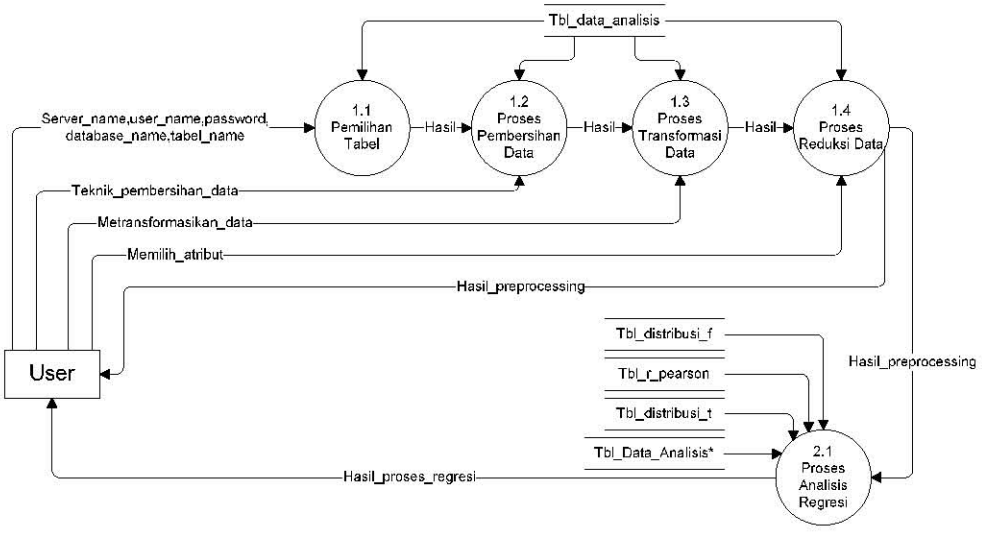
mendapatkan asupan data berupa data hasil preprocessing. Dan data tersebut akan di analisis oleh sistem data mining, dan di dapatkan hasil analisis regresi yang telah dikerjakan oleh sistem data mining tersebut.

3.3.1 Data Flow Diagram (DFD) Level 1



Gambar 3.3 *Data Flow Diagram Level 1*

3.3.2 Data Flow Diagram (DFD) Level 2



Gambar 3.4 *Data Flow Diagram Level 2*

3.4 Perancangan Proses

Pada tahap perancangan proses ini diberikan gambaran proses pada setiap bagian-bagian aktifitas dari sistem data mining yang akan diimplementasikan.

3.4.1 Rancangan Proses Pemilihan Tabel (DFD Level 2 Proses 1.1)

Proses ini melakukan proses koneksi dengan database dan mengambil tabel yang akan dianalisis dengan proses analisis regresi. Proses ini membutuhkan masukan dari *user* berupa *server name*, *user name, password, database name* dan *tabel name*. Setelah masukan diterima oleh proses ini, maka akan dilakukan koneksi dengan database yang dibutuhkan oleh *user*, dan menghubungkan dengan tabel yang akan dianalisis. Keluaran proses ini yaitu berupa tabel yang dipilih oleh *user* untuk dianalisis, dan keluaran proses ini dilanjutkan kepada proses analisis regresi.

3.4.2 Rancangan Proses Pembersihan Data (DFD Level 2 Proses 1.2)

Proses ini melakukan pembersihan data untuk menambahkan isi atribut yang hilang atau kosong, dan merubah data yang tidak konsisten. Proses ini membutuhkan masukan berupa tabel yang telah dipilih oleh user sebelumnya. Setelah mendapatkan tabel yang diinginkan maka dalam proses ini akan membersihkan data-data yang kosong dengan metode yang telah dijelaskan sebelumya.

3.4.3 Rancangan Proses Transformasi Data (DFD Level 2 Proses 1.3)

Proses ini melakukan perubahan data kedalam bentuk yang sesuai untuk proses data mining. Proses ini membutuhkan masukan dari hasil proses pembersihan data sebelumnya. Proses ini akan menghasilkan data dalam bentuk yang sesuai dengan proses data mining yang akan dilakukan.

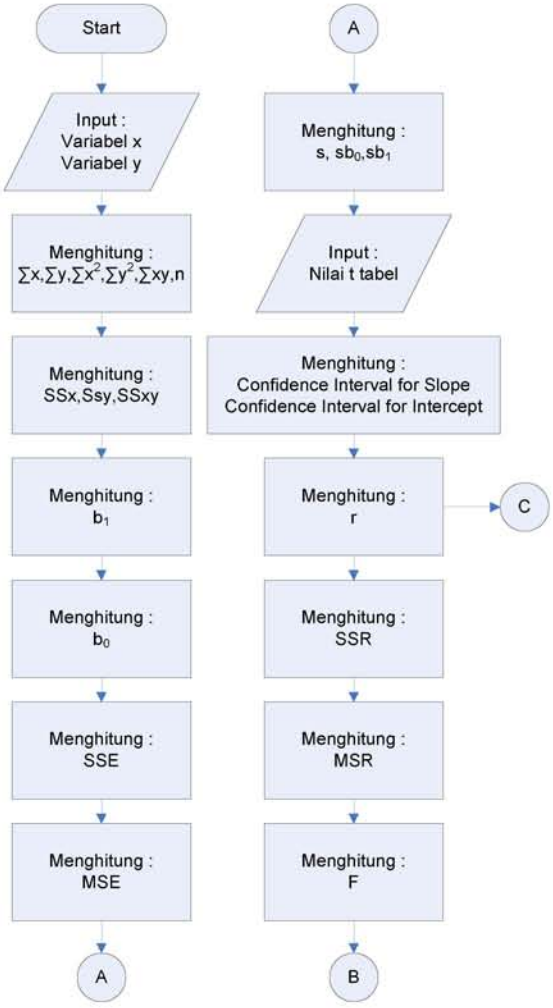
3.4.4 Rancangan Proses Reduksi Data (DFD Level 2 Proses 1.4)

Proses ini melakukan pemilihan atribut-atribut yang dibutuhkan untuk proses data mining. Proses ini membutuhkan masukan dari hasil proses transfromasi data sebelumnya, proses ini akan menghasilkan data dengan atribut-

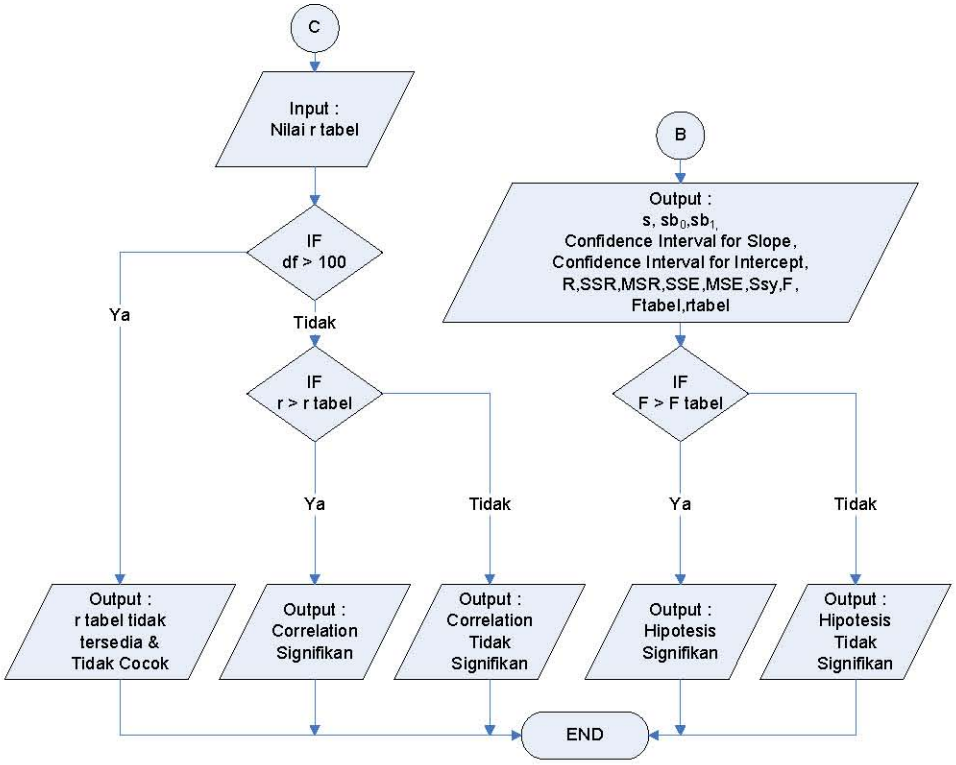
atribut yang dibutuhkan saja sehingga data yang dihasilkan akan lebih kecil dan efisien.

3.4.5 Rancangan Proses Analisis Regresi (DFD Level 2 Proses 2.1)

Proses ini melakukan proses analisis regresi, proses ini melanjutkan dari hasil proses pemilihan tabel yang berupa tabel data yang siap dianalisis. Selain itu proses ini membutuhkan inputan berupa tabel yang mendukung proses analisis regresi, tabel yang digunakan yaitu tabel data analisis, tabel distribusi t, tabel distribusi F dan tabel r pearson. Dalam proses ini terdapat urutan-urutan proses yang dilakukan untuk menghasilkan suatu keluaran berupa hasil analisis regresi. Urutan-urutan proses tersebut digambarkan dalam *flowchart* dibawah ini (gambar 3.5 dan gambar 3.6):



Gambar 3.5 Flowchart Proses Analisis Regresi



Gambar 3.6 Flowchart Proses Analisis Regresi Lanjutan

3.5 Perancangan Basis Data

Tabel basis data yang akan digunakan dalam perancangan, sebagai berikut:

1. Tabel Distribusi t

   Merupakan tabel yang dibutuhkan untuk mendapatkan nilai t tabel, yang mempunyai field df dan alpha, yang mempunyai *primary key* df.

2. Tabel Distribusi F

   Merupakan tabel yang dibutuhkan untuk mendapatkan nilai F tabel, yang mempunyai field df dan numerator, yang mempunyai *primary key* df.

3. Tabel r Pearson

   Merupakan tabel yang dibutuhkan untuk mendapatkan nilai r tabel, yang mempunyai field df dan nilai_r, yang mempunyai *primary key* df.

3.6 Kamus Data

Berikut ini akan dijelaskan data-data yang digunakan dalam perancangan proses (*Data Flow Diagram*).

Tabel 3.2 Tabel Kamus *Data Flow Diagram* (DFD)

No	Data	Kamus Data
1.	Server_name	/* Nama server yang akan digunakan */
2.	User_name	/* User name dari server yang akan digunakan */
3.	Password	/* Password dari server yang akan digunakan */
4.	Database_name	/* Nama dari database yang akan digunakan */
5.	Tabel_name	/* Nama dari tabel yang akan dianalisis */
6.	Tabel_pilihan	/* Tabel yang telah dipilih user untuk dianalisis */
7.	Hasil_proses_regresi	/* Menampilkan hasil proses yang dikerjakan */

3.7 Deskripsi Data

Pada bagian ini akan diberikan keterangan dari tabel-tabel basis data yang akan digunakan dalam sistem data mining.

Tabel 3.3 Tabel Distribusi t

Field	Type	Keterangan
Df	Integer	*Primary key*, nilai dari *Degrees of Freedom*
Alpha	Decimal	Berisikan nilai t berdasarkan perhitungan tabel

Tabel 3.4 Tabel Distribusi F

Field	Type	Keterangan
Df	Integer	*Primary key*, nilai dari *Degrees of Freedom*
Numerator	Decimal	Berisikan nilai F berdasarkan perhitungan tabel

Tabel 3.5 Tabel r Pearson

Field	Type	Keterangan
Df	Integer	*Primary key*, nilai dari Degrees of Freedom
Nilai r	Decimal	Berisikan nilai r berdasarkan perhitungan tabel

3.8 Batasan Perancangan Aplikasi

Batasan perancangan aplikasi yang akan dibangun dalam penerapan teknik data mining menggunakan metode analisis regresi ini adalah sebagai berikut :

1. Input atau target data adalah suatu tabel yang telah melalui proses data *preprocessing*.
2. Melakukan proses analisis regresi dari suatu tabel dalam hal ini adalah table yang akan dianalisis yaitu tabel sample dan tabel populasi.
3. Menampilkan secara grafis data hasi proses analisis regresi dalam bentuk grafik dan data-data berbentuk suatu nilai hasil perhitungan analisis regresi.

3.9 Arsitektur Perangkat Lunak

Dibawah ini merupakan gambar dari arsitektur aplikasi data mining yang akan dibangun :

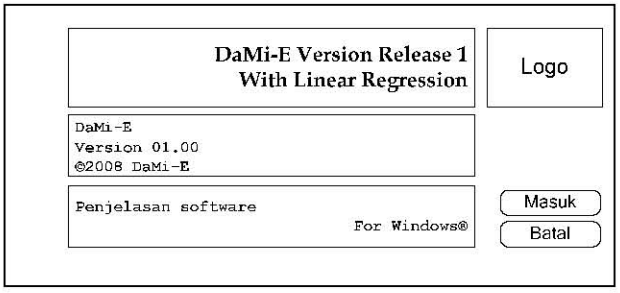
Gambar 3.7 Arsitektur Aplikasi Data Mining

3.10 Perancangan Antarmuka

Perancangan antarmuka menjelaskan rutinitas program yang akan dijalankan oleh sebuah sistem komputerisasi untuk menjelaskan interaksi antara pemakai (*user*) dengan program yang dibuat.

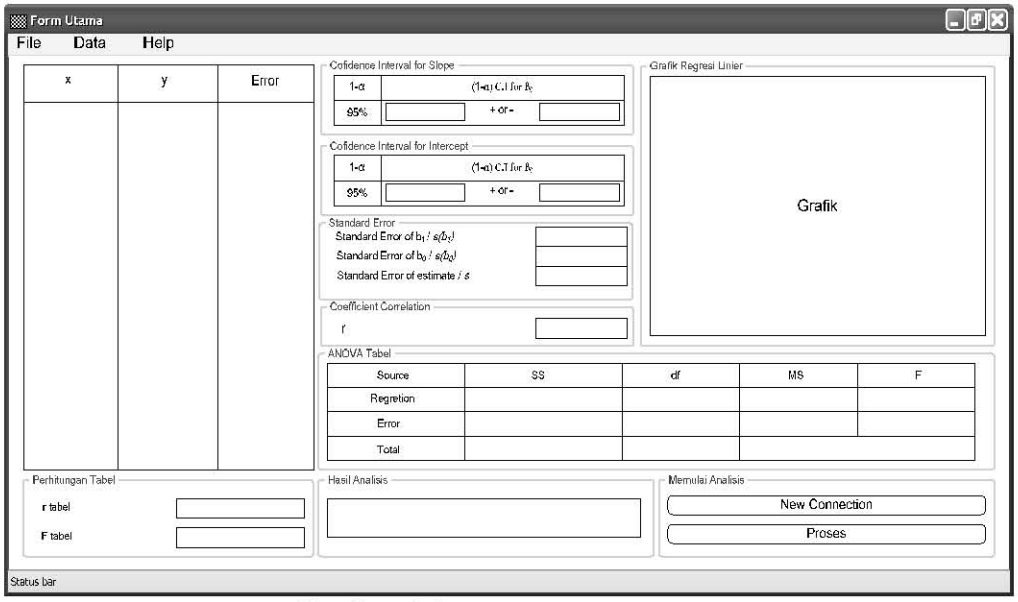
Pada sub bab ini akan digambarkan rancangan antarmuka yang akan digunakan dalam sistem data mining.

a. Desain Form Awal



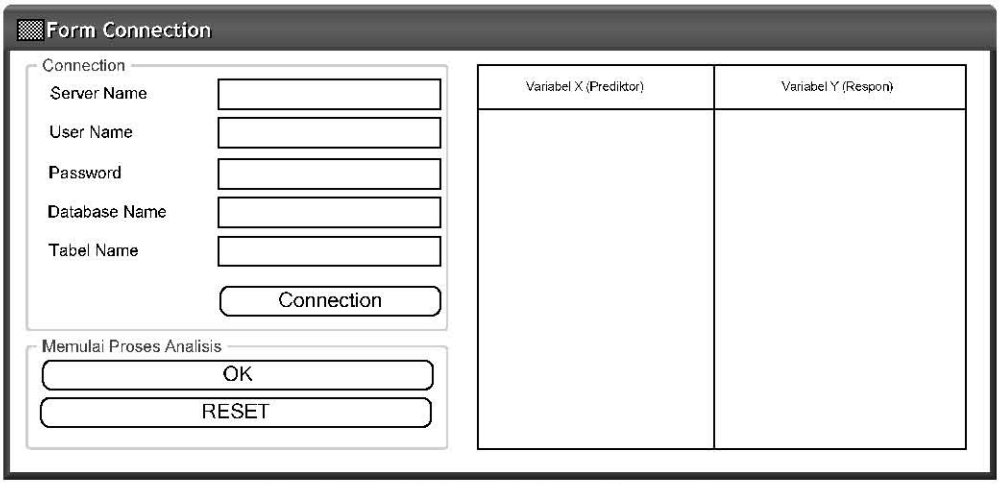
Gambar 3.8 Desain Form Awal

b. Desain Form Utama



Gambar 3.9 Desain Form Utama

c. Desain Form *Connection*



Gambar 3.10 Desain Form *Connection*