DAFTAR TABEL

BAB 1 PENDAHULUAN

1.1 Latar belakang

Pada zaman sekarang bisnis berkembang menjadi lebih pesat dibandingkan dengan masa lalu, dimana tujuan perusahaan mengedepankan kebutuhan pelanggan untuk menjadi lebih kompetetif dan dinamis dibandingkan dengan perusahaan atau organisasi lainnya. Dengan demikian perkembangan Business Intelligence yang ada dimanfaatkan untuk strategi bisnis yang tepat. Business intelligence merupakan cara untuk mengumpulkan, menyimpan, mengorganisasikan, membentuk ulang, meringkas data serta menyediakan informasi baik berupa data aktifitas bisnis internal perusahaan termasuk aktifitas bisnis pesang yang mudah diakses serta dianalisis untuk berbagai kegiatan manajemen (David,2000) [1].

Business intelligence juga dapat dispesifikasi melalui empat dimensi yaitu : Strategis, Taktis, Operasional dan *Real-time* (Asghar, Fong, & Hussain, 2009) [2].Pada Real-time BI (rt-BI) akurasi generalisasi dan waktu respon merupakan kriteria penting . Menganalisis data sesegera mungkin ketika data masuk kedalam organisasi merupakan bagian dari real-time BI.

Akurasi generalisasi dan waktu respon adalah dua kriteria penting untuk mengevaluasi adanya suatu pengklasifikasian bila diterapkan dalam real-time BI (rt-BI). Pengklasifikasian diperlukan tidak hanya untuk menggambarkan pelatihan data tetapi juga untuk memperediksi data yang tidak terlihat (Jiaqi Wang, 2005) [3].

Oleh karena itu kebutuhan perkembangan BI yang meningkat,maka diperlukan adanya analisis data sesegera mungkin ketika data masuk kedalam suatu organisasi. Proses memasukan data sesegera mungkin yang terjadi secara berulang akan menimbulkan penumpukan terhadap data sehingga mempengaruhi pencarian informasi terhadap data tersebut.

Pada karya tulis ini,salah satu metode yang digunakan untuk melakukan proses data mining dengan menggunakan stream mining . Metode ini memiliki tujuan untuk mengatasi masalah dari jumlah data yang banyak dan di proses secara terus menerus dan membutuhkan waktu yang lama . Dengan menggunakan metode ini diharapkan dapat mengeluarkan hasil analisis BI dengan tingkat keterlambatan yang rendah. Untuk metode pengelompokkan data ,penulis menggunakan *Ensemble Method*. Metode ini digunakan untuk mempercepat pengelompokan data dan memiliki tingkat akurasi analisis yang tinggi. Menggunakan *Ensemble Method* dikarenakan data yang dipergunakan ukurannya sangat besar sehingga satu algoritma tidak cukup untuk mengatasi masalah ini sehingga metode *Ensemble Method* yang dipilih untuk menyelesaikan masalah ini. *Ensemble Method* yang dipilih untuk membahas masalah ini adalah K-means untuk clustering dan Algoritma Naïve Bayes untuk klasifikasi. Akurasi yang didapatkan pada penlitian tugas akhir ini yaitu 93.33%.

1.2 Perumusan Masalah

Berdasarkan uraian latar belakang diatas,maka perumusan masalah yang dicapai pada penelitian ini adalah sebagai berikut :

1. Bagaimana cara mengimplementasikan ensemble method pada sistem Business Intelligence ?
2. Bagaimana metode *k-means clustering* dan algoritma *Naïve Bayes* dapat mempercepat waktu respons pengelompokan data ?
3. Bagaimana metode *k-means clustering* dan algoritma *Naïve Bayes* dapat meningkatkan tingkat akurasi pemrosesan analisis ?

1.3 Tujuan

Berdasarkan uraian latar belakang diatas, maka tujuan yang dicapai pada penelitian ini adalah sebagai berikut :

1. Mengimplementasikan algoritma K-Means clustering dan algoritma Naïve Bayes untuk proses klasifikasi pada sistem Real-time Business Intelligence
2. Menganalisa kineja kedua algoritma yang digunakan dalam menghasilkan sebuah model clustering dan klasifikasi terbaik yang akan di implementasikan kedalam sebuah sistem
3. Menganalisa performansi sistem dalam prediksi pengelompokan pelanggan berdasarkan data sales dan pelanggan.

1.4 Batasan Masalah

Adapun batasan masalah yang dibahas pada penelitian ini adalah sebagai berikut :

1. Data yang digunakan adalah data customer dan data sales dari sebuah perusahaan.
2. Data yang masuk kedalam sistem diasumsikan sudah melewati tahap *ekstrasi data*.
3. Data yang dianalisis memiliki atribut kategorik.
4. Metode yang digunakan adalah Clustering Kmeans dan Klasifikasi Naive Bayes.
5. Data yang akan dikelompokan tidak dapat dilakukan melalui proses online melainkan data diproses secara offline.
6. Sistem yang dibuat tidak berorientasi objek.

1.5 Metodologi

Metodologi untuk menyelesaikan masalah pada karya tulis ini adalah sebagai berikut :

1. Identifikasi masalah
   Masalah yang ada didapat dari beberapa karya ilmiah yang ada dan masalah yang didapat dalam kehidupan sehari-hari.Kemudian masalah yang ada dirumuskan menjadi permasalahan yang utama.

2. Studi literature

   Melakukan pembelajaran mengenai hal-hal yang berkaitan dengan masalah yang dibahas pada karya tulis ini (*Real-time Business Intelligence*, *K-means Clustering*, dan *Data Mining*) melalui berbagai paper,jurnal,buku dan artikel di internet .

3. Pemodelan sistem

   Setelah studi literature sudah dilakukan yang berikutnya melakukan pemodelan sistem dari karya tulis yang dibuat. Pemodelan sistem dilakukan untuk mempermudah proses pengerjaan dan mencegah adanya perubahan-perubahan pada saat tahap implementasi sampai dengan tahap akhir.

4. Implementasi sistem

   Setelah pemodelan sistem dilakukan proses selanjutnya adalah proses abstraksi.proses pengubahan komponen komponen yang ada menjadi sebuah bahasa pemrograman.

5. Pengujian sistem

   Proses ini adalah proses pengujian yang dilakukan untuk mengetahui keakuratan dari sistem yang telah dibuat. Pengujian dilakukan lebih dari satu kali agar tingkat keakuratan dapat terlihat sebagai bahan untuk analisis.

6. Pembuatan Laporan

   Proses yang terakhir adalah pembuatan laporan,pada proses ini analisis dilakukan untuk mendapatkan suatu kesimpulan dari sistem yang ada apakah sesuai dengan kriteria yang ditentukan.selain itu juga pada laporan adanya suatu evaluasi dari sistem ini berupa suatu kelebihan dan kelemahan sistem sehingga nantinya dapat dikembangkan. Pada laporan semua aspek yang ada ditulis sesuai dengan ketentuan yang ada.

BAB 2 TINJAUAN PUSTAKA

2.1 Business Intelligence

BI merupakan sebuah proses analisis yang mengubah data internal dan data eksternal menjadi informasi tentang kapabilitas, posisi pasar, aktifitas dan tujuan yang harus dikerjar oleh sebuah perusahaan untuk tetap bersifar kompetitif. *Business Intrlligence* berdiri untuk konsep sistem informasi seperti *Online Analytical Processing* (*OLAP*), membuat *query* dan pelaporan, atau *data mining* yang menyediakan metode-metode berbeda untuk sebuah analisis data bisnis berdasarkan tujuan yang fleksibel dan berdasarkan sebuah *data pool* yang terpusat(Schiefer & Seufert, 2005) [4].

BI dapat memfasilitasi hubungan antara organisasi dengan bentuk baru, membawa informasi *real-time* ke tempat penyimpanan yang terpusat dan membantu analisis yang dapat di-eskploitasi pada setiap level *horizontal* dan *vertical* baik di dalam maupun di luar perusahaan [5].

Karena BI memiliki banyak konsep umum, *Business Intelligence* tidak mempunyai istilah yang terdefinisi dengan baik.Pebedaaan cara pandang ini memperjelas bahwa BI memiliki banyak segi. Untuk menangkap segi-segi tersebut, *Business Intelligence* didefinisikan sebagai semua tentang bagaimana caranya untuk menangkap, mengakses, memahami, menganalisis dan mengubah salah satu aset paling berharga dari sebuah perusahaan melalui data mentah kemudian menjadi informasi yang dapat ditindak lanjuti dengan tujuan untuk meningkankan performansi bisnis(Azvine, Cui, Nauck, & Majeed, 2006) [6].

2.2 Real-time Business Intelligence

Real-time Business Intelligence berhubungan dengan banyak teknologi dan alat revolusi dari Business Intelligence baik secara strategis maupun taktis. Real-time juga dapat diartikan sebagai sebuah kemampuan untuk memperoleh pengukuran performansi kunci yang berhubungan dengan situasi pada masa sekarang dan tidak hanya beberapa situasi [6].

2.3 Ensemble Methods

Berbagai macam metode bersaing tersedia untuk menginduksi model dari data,dan bergantung pada relative data. Akurasi perbandingan algoritma bergantung pada rincian data yang ada [7].

Model Ensemble meningkatkan akurasi dan ketahanan atas metode model tunggal, contoh aplikasi yang menggunakan ensemble methods adalah sebagai berikut :

- Komputasi Terdistribusi
- Aplikasi Privasi
- Data berskala besar dengan model yang dapat digunakan kembali
- Berbagai sumber data

2.4 Clustering pada Data Mining

Pada data mining, *Clustering* merupakan pembagian menjadi grup yang serupa atau sama . Pada proses *Clustering* beberapa perincian diabaikan dalam suatu pertukaran

untuk penyederhanaan data . *Clustering* terkait dengan berbagai peranan yang penting dalam jangkauan suatu aplikasi yang lebih luas . Aplikasi *clustering* biasanya berurusan dengan dataset dan data yang memiliki banyak atribut . [7]

2.5 K-means

K-means Clustering adalah suatu metode yang umum digunakan untuk partisi data set secara otomatis ke dalam kelompok K. Proses dilakukan dengan cara memilih k pusat cluster awal dan kemudian melakukan iterative sebagai berikut:

- Setiap contoh Di ditugaskan untuk pusat terdekat pada cluster.
- Setiap pusat cluster Cj diperbarui menjadi sebuah mean.

Algoritma *K-means clustering* akan menyatu ketika tidak ada lagi perubahan penugasan kasus ke dalam masing masing cluster [8]

Berikut merupakan formula yang digunakan setelah menentukan nilai K dan titik centroid

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(xi - yi)2} \quad \text{.....(2.1)}$$

2.6 Algoritma Naïve Bayes

Model naïve bayes adalah model probabilitas yang disederhanakan dalam Bayesian. Naïve bayes classfier beroperasi pada asumsi yang kuat. Ini disebabkan bahwa probabilitas dari satu atribut tidak mempengaruhi probabilitas yang lain, sehingga serangkaian atribut yang ada di dalam algoritma naïve bayes membuat asumsi yang bersifat independen. Hasil klasifikasi Naïve Bayes seringkali benar atau tepat . Kesalahan yang biasa terjadi pada algoritma naïve bayes disebabkan oleh tiga faktor,yaitu : Data Latih,bias dan nilai varians. Data latih hanya dapat diminimalkan dengan memilih data training yang baik. Bias merupakan kesalahan pada pengelompokan data latih yang jumlahnya besar, Sedangkan varians merukana kesalahan dalam jumlah data yang sedikit atau kecil. Dalam (Hamzah, A., 2012) menyebutkan bahwa algoritma naïve bayes merupakan algoritma yang memiliki kinereja tinggi untuk proses pengklasifikasian.

Naïve bayes merupakan sebuah classifier probabilistik berdasarkan Bayes Rule Of Conditional Probability. Naïve Bayes menggunakan probabilitas untuk mengklasifikasi kelas baru [12]. Cara kerja algoritma Naïve Bayes dengan mencari peluang terbesar dari kemungkinan klasifikasi dengan melihat frekuensi tiap klasifikasi pada data training.

Terdapat dua peluang dalam algoritma Naïve Bayes yaitu Posterior dan Prior [7]. P(H/X) merupakan probabilitas Posterior dari H yang dikondisikan dalam X, Sedangkan P(H) merupakan probabilitas Prior dari H. Probabilitas Posterior merupakan probabilitas yang didasarkan pada informasi-informasi yang ada, sedangkan probabilitas Prior merupakan probabilitas yang independen. Secara sama,P(X|H) merupakan probabilitas dari Xyang dikondisikan dalam H. Sedangkan P(X) adalah probabilitas Prior dari X [7].

Bentuk umum dari teori bayes adalah sebagai berikut:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad \text{.....(2.2)}$$

Keterangan dari bentuk umum teori bayes diatas adalah :

X : Data tanpa label atau belum diketahui kelas

H : Pelabelan data X yang merupakan suatu class spesifik

P(H/X) : Posterior

P(H) : Prior

P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) : Probabilitas kondisi X

Naïve Bayes merupakan bentuk sederhana dari teori bayes yang ada , Bentuk umum dari Naïve Bayes adalah sebagai berikut

$$P(H|X) = (P(X|H)P(X) \quad \text{.....(2.3)}$$

2.7 Klasifikasi pada Data Mining

Klasifikasi adalah proses dalam data mining yang digunakan untuk memprediksi nilai-nilai yang belum diketahui dengan menggunakan kelas yang sudah diketahui nilainya. Proses pada klasifikasi dibagi menjadi dua fase,yaitu learning dan test. Pada fase learning sebagian data yang telah diketahui kelas datanya dijadikan Classifier sebagai bentuk model perkiraan. Kemudian untuk fase test model yang sudah terbentuk diuji dengan menggunakan sebagian data lain untuk mengetahui nilai akursi dari model tersebut [10].

Klasifikasi adalah proses sebagai dasar dalam pembangunan sebuah model. Sebelum melakukan proses pembelajaran data, data yang akan digunakan telah memilki kelas label pada setiap barisnya sehingga bisa melakukan suatu klasifikasi secara otomatis [11]. Pengukuran kinerja *classifier, ad*alah sebagai berikut.

- Sensitivitas dan Spesifisitas

Sensitivitas merupakan bagian dari nilai true positive yang telah diklasifikasikan oleh classifier, sedangkan spesifitas adalah bagian dari nilai true negative yang diklasifikan dengan benar oleh classfier.

True positive merupakan nilai positif tuple yang diklasifikasikan secara benar atau tepat, sedangkan True negative merupakan nilai negative tuple yang diklasifikasikan secara benar [11].

Formula sensitivitas dan Spesifitas dapat dihitung dengan cara sebagai berikut:

$$Sensitivitas = \frac{TRUE\ POSITIVE}{POSITIVE} \quad \text{.....(2.4)}$$

$$Spesifisitas = \frac{TRUE\ NEGATIVE}{NEGATIVE} \quad \text{.....(2.5)}$$

- Akurasi

Akurasi merupakan nilai presentase dari jumlah keseluruhan data yang diklasifikasikan secara benar oleh classifer. [11]
Perhitungan akurasi dapat dilakukan dengan cara sebagai berikut :

$$akurasi = Sensitivitas\frac{Positive}{Positive+Negative} + Spesifisitas\frac{Negative}{Positive+Negati} \quad \ldots(2.6)$$

- Precision dan Recall

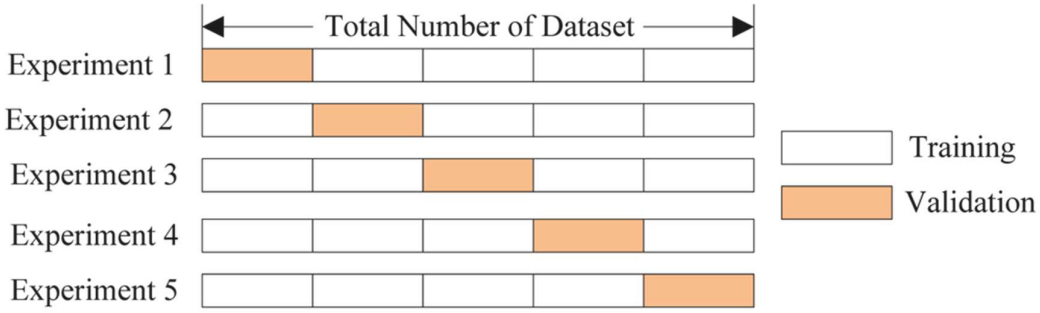
Precision merupakan persentase dalam pengukuran kinerja sistem untuk mendapatkan data yang sesuai, sedangkan recall adalah presentase dalam pengukuran kinerja sistem untuk memperoleh data yang relevan [10].
Untuk mengetahui nilai precision dan recall adalah sebagai berikut :

$$Precision = \frac{True\ Positive}{True\ Positive+False\ Positive} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.7)$$

$$Recall = \frac{True\ Positive}{True\ Positive+F\quad Negative} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.7)$$

2.8 K-Fold Cross Validation

Cross Validation merupakan sebuah teknik pengujian akurasi dari classifier. Dalam k-fold cross validation, data dipartisi kedalam beberapa bagian, dimana setiap bagian tersebut memiliki ukuran yang sama.
Misalkan dalam iterasi yang dilakukan pertama kali, bagian k=1 hingga bagian k=9 merupakan data training dan bagian k=10 merupakan data testing. Dilanjutkan hingga masing masing k menjadi data testing.



Gambar 2-1 Skenario K-Fold validation

BAB 3 PERANCANGAN SISTEM

Pada bab ini akan dibahas mengenai gambaran umum sistem, data, clustering dengan menggunakan algoritma K-Means, klasifikasi dengan menggunakan algoritma Naives Bayes dan kebutuhan sistem.

3.1 Gambaran Umum Sistem

Pada sistem yang akan dibangun terdapat basis aturan yang berisikan aturan yang akan dianalisis kepeda pengguna dan basis ensemble method yang berisi model-model yang telah ditetapkan di dalam sistem .Sistem ini memiliki dua proses utama dalam menghasilkan analisis *Business Intelligence*, yaitu proses stream mining dan pengelompokan data .

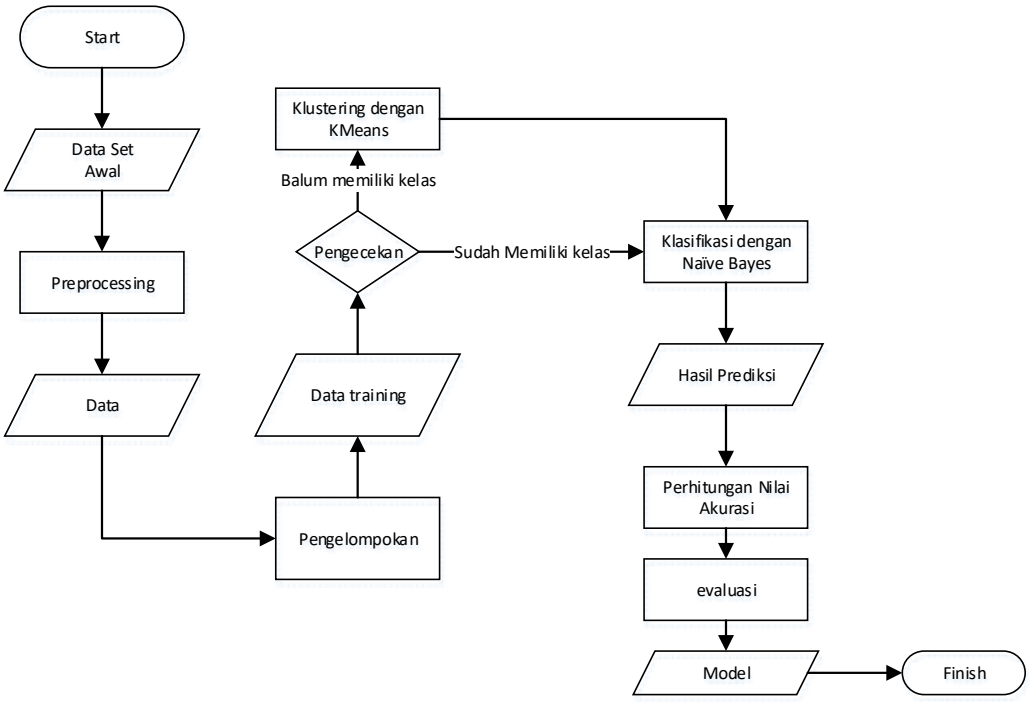
Proses ini berawal dari masuknya data yang sudah melewati tahap pre-processing atau data yang sudah siap masuk ke dalam sistem. Sistem akan melakukan pengecekan terhadap data tersebut dengan method yang sudah tersedia didalam basis data apakah data tersebut memiliki kelas atau tidak.

Jika data yang masuk telah memiliki kelas maka selanjutkan akan ketahapan pencocokan data yang terdapat pada basis aturan. Jika ternyata data yang masuk sesuai sudah cocok dengan basis aturan maka selanjutnya hasil analisis BI dapat dikeluarkan dan ditampilkan kepada pengguna. Jika ternyata data tidak cocok dengan seua aturan yang ada pada basis aturan, maka data akan dialihkan pada proses clustering.

Proses clustering ini akan dilakukan secara terus menerus, ARBBapabila data belum sesuai dengan basis aturan.

Pada akhir proses analisis secara real-time ,selain hasil dari analisis BI akan ditampilkan kepada pengguna,informasi pengelompokan data akan di integrasikan kedalam basis aturan. Selama proses analisis secara real-time berlangsung,semua informasi akan diperbarui didalam basis data. Dengan alur seperti ini maka sistem BI bergantung pada data yang masuk.

Secara umum, alur proses yang terjadi pada sistem digambarkan dalam gambar 3.1 berikut :



Gambar 3-1 Rancangan Sistem

Berdasarkan gambar 3.1, Dilakukan beberapa langkah untuk membangun sistem yaitu :

1. Dataset yang digunakan untuk melakukan proses pengelompokan dalam tugas akhir ini adalah dataset Sales Report dan dataset Customer dari PT.Intergrated Logixtrem.
2. Dilakukan proses preprocessing pada dataset yang ada sehingga data set tersebut sesuai dengan model yang akan digunakan.
3. Dilakukan proses pengecekan data yang masuk kedalam sistem sudah memiliki kelas atau belum,jika data masuk dan belum memiliki kelas atau masih disebut dengan data mentah maka data tersebut dilakukan proses klustering dengan menggunakan Algoritma K-Means.
4. Sedangkan data yang masuk telah memiliki kelas pada proses pengecekan ketika akan dilakukan suatu pengelompokan maka data tersebut akan diproses dengan menggunakan proses klasifikasi dengan menggunakan Algortima Naïve Bayes
5. Proses pengelompokan akan menghasilkan suatu model data yang diuji dengan data uji dengan menghitung suatu nilai akurasi.
6. Jika belum menghasilkan data yang tepat proses pengelompokan dengan menggunakan klustering dan klasifikasi akan dilakukan terus menerus.
7. Setelah mendapatkan hasil terbaik akan dicocokan dengan hasil prediksi.
8. Proses akan menghasilkan sebuah model yang digunakan sebagai model BI dan proses pembangunan sistem selesai.

3.2 Dataset

Dataset yang digunakan dibagi menjadi dua yaitu data Sales Report dan data Customer, berikut merupakan deskripsi dari masing-masing dataset tersebut:

1. Data Sales Report tahun 2013,2014,2015,2016 yang terdiri dari atribut berikut:

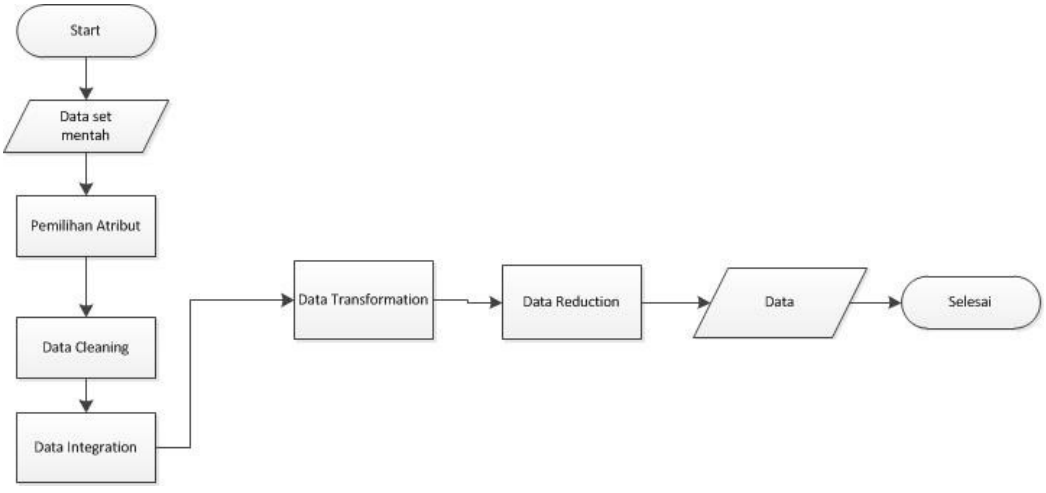
- Tahun : Menunjukan tahun terhadap dataset yang ada
- Jenis : Menunjukan barang pada proses transaksi dalam data tersebut
- SI (Shipping Instruction)
- Tanggal : Menunjukan waktu transaksi
- Quanitity : Menunjukan jumlah transaksi yang dilakukan
- Amount : Menunjukan total harga transaksi yang dilakukan
- Nama cust.: Menunjukan nama customer yang melakukan transaksi

2. Data Costumer yang didapatkan dari PT.Intergrated Logixtream, yang terdiri dari atribut sebagai berikut :

- Customer no
- Nama cust.
- Phone
- Contact name
- Balance prime
- Balance tax
- Suspended
- Address

3.3 Preprocessing

Pada tahap preprocessing ini data Sales Report dan data Costumer mentah diolah agar siap dilakukan proses mining. Proses preprocessing dalam tugas akhir ini digambarkan pada gambar 3.2 sebagai berikut :



Gambar 3-2 Tahapan Pre-Processing