



Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

BAB II

LANDASAN TEORI

Penelitian ini bertujuan untuk membuat visualisasi data dan juga prediksi data dengan menggunakan *Tools Microsoft PowerBI* untuk memvisualisasikan data penggunaan IBM SPSS untuk memprediksi data yang menggunakan algoritma *Linear Regression*. Pada bab ini menjelaskan tentang visualisasi dan juga algoritma *Linear Regression* menurut berbagai sumber, buku, pengertian, jurnal, dan juga penelitian terdahulu.

2.1. Unit Pengolahan Ikan

Ikan merupakan salah satu bahan pangan yang cukup mudah didapatkan di Indonesia mengingat bahwa potensi laut yang dimiliki oleh Indonesia begitu luas ditambah dengan Sumber air tawar yang cukup banyak untuk pengembangan perikanan darat, oleh karena itu ikan merupakan bahan pangan yang cukup penting bagi ketersediaan pangan di Indonesia. Hal ini dapat dilihat dari data konsumsi ikan per kapita di Indonesia dari tahun ke tahun yang terus meningkat, sejalan dengan terjadinya perubahan kecenderungan konsumsi dunia yang beralih dari protein hewan ke protein ikan (Shoimah, Purnaweni, & Yulianto, 2013).

Produk perikanan di Indonesia dikonsumsi dalam bentuk olahan dan ikan segar. Teknik pengolahan ikan yang banyak dilakukan di Indonesia adalah penggaraman/pengeringan, pemindangan, pengolahan lain, pengasapan/pemanggangan, fermentasi, penanganan segar,

pengekstraksian/pereduksian, pembekuan, pengolahan jelly ikan/surimi dan pengalengan (Shoimah, Purnaweni, & Yulianto, 2013).

Sektor Pengolahan ikan juga menjadi faktor pendukung tingginya nilai Produk Domestik Bruto, berikut detail tabel Produk Domestik Bruto pada sektor perikanan. (Fransiska, 2009)

2.2.Pengertian Prediksi

Prediksi adalah suatu proses memperkirakan secara sistematis tentang sesuatu yang paling mungkin terjadi di masa depan berdasarkan informasi masa lalu dan yang dimiliki sekarang, agar kesalahannya (selisih antara sesuatu yang terjadi dengan hasil perkiraan) dapat diperkecil. Prediksi tidak harus memberikan jawaban secara pasti kejadian yang akan terjadi, melainkan berusaha untuk mencari jawaban sedekat mungkin yang akan terjadi (Herdianto, 2013).

Berdasarkan teknik yang digunakan untuk memprediksi maka prediksi dapat dibagi menjadi dua bagian yaitu prediksi kualitatif dan prediksi kuantitatif (Baafi, 2015):

1. Prediksi Kualitatif

Prediksi kualitatif didasarkan atas data kualitatif pada masa lalu. Metode kualitatif digunakan jika data masa lalu dari variabel yang akan diprediksi tidak ada, tidak cukup atau kurang dipercaya. Hasil prediksi yang dibuat sangat tergantung pada individu yang menyusunnya. Hal ini penting karena hasil prediksi tersebut ditentukan berdasarkan pemikiran yang bersifat judgement atau opini, pengetahuan dan pengalaman dari penyusunnya. Oleh

karena itu metode kualitatif ini disebut juga *judgemental*, *subjective*, *intuitive*.

2. Prediksi Kuantitatif

Prediksi kuantitatif didasarkan atas data kuantitatif pada masa lalu. Hasil prediksi yang dibuat sangat tergantung pada metode yang dipergunakan dalam prediksi tersebut. Dengan metode yang berbeda akan diperoleh hasil prediksi yang berbeda. Hal yang perlu diperhatikan dari penggunaan metode tersebut adalah baik tidaknya metode yang digunakan dan sangat ditentukan dari penyimpangan antara hasil prediksi dengan kenyataan yang terjadi. Metode yang baik adalah metode yang memberikan nilai-nilai perbedaan atau penyimpangan yang mungkin. Prediksi kuantitatif hanya dapat digunakan apabila terdapat tiga kondisi sebagai berikut :

- a. Adanya informasi tentang keadaan yang lain.
- b. Informasi tersebut dapat dikuantifikasikan dalam bentuk data.
- c. Dapat diasumsikan bahwa pola yang lalu akan berkelanjutan pada masa yang akan datang

2.3. Metodologi Prediksi

Untuk melakukan prediksi data diperlukan model-model algoritma prediksi. Model-model algoritma tersebut adalah:

2.3.1. *Linear Regresion*

Linear Regresion digunakan untuk estimasi, prediksi, forecast pada nilai-nilai suatu *variable* berdasarkan nilai-nilai dari *variable* lainnya, serta

hubungan antar 2 *variable*. Variabel Y adalah nilai-nilai yang akan di prediksi, *variable* Y disebut juga *criterion*. *Variable* X adalah nilai-nilai sebagai dasar untuk memprediksi, X disebut predictor (Djie, 2013).

Model persamaan *Linear Regression* sederhana adalah seperti berikut:

$$Y = a + bX$$

Rumus 2.1. Rumus Linear Regression

Dimana:

Y = Variabel Response atau Variabel Akibat (Dependent)

X = Variabel Predictor atau Variabel Faktor Penyebab (Independent)

a= konstanta

b = koefisien regresi (kemiringan); besaran Response yang ditimbulkan oleh Predictor.

Nilai-nilai a dan b dapat dihitung dengan menggunakan rumus dibawah ini:

$$a = \frac{(\sum y)(\sum X^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

Rumus 2.2. Linear Regression a

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Rumus 2.3 Linear Regression b

Berikut ini adalah Langkah-langkah dalam melakukan Analisis Regresi Linear Sederhana (Djie, 2013) :

1. Tentukan Tujuan dari melakukan Analisis Regresi Linear Sederhana
2. Identifikasikan Variabel Faktor Penyebab (Predictor) dan Variabel Akibat (Response)
3. Lakukan Pengumpulan Data
4. Hitung X^2 , Y^2 , XY dan total dari masing-masingnya
5. Hitung a dan b berdasarkan rumus diatas.
6. Buat Model Persamaan Regresi Linear Sederhana.
7. Lakukan Prediksi atau Peramalan terhadap Variabel Faktor Penyebab atau Variabel Akibat.

2.3.2. Decision Tree

Decision Tree adalah satu metode klasifikasi yang paling populer karena mudah diinterpretasikan oleh manusia. *Decision tree* digunakan untuk pengenalan pola dan termasuk dalam pengenalan pola secara statistic (Hastuti, 2012).

Decision Tree dibentuk dari 3 tipe simpul (Hastuti, 2012):

a. Simpul Leaf

Membuat suatu keputusan akhir atau kelas target untuk suatu pohon keputusan

b. Simpul root

Merupakan titik awal dari suatu decision tree

c. Simpul perantara

Setiap simpul perantara berhubungan dengan suatu pertanyaan atau pengujian.

Decision Tree memiliki algoritma seperti berikut (Hastuti, 2012):

i. Pohon dibangun dalam suatu metode rekursif *topdown*

divide and-conquer

1. Seluruh contoh pelatihan dimulai dari simpul root, lalu dilakukan pengujian

2. Mencabang ke jalur yang benar berdasarkan hasil pengujian

3. Apakah simpul leaf ditemukan? Jika ya, masukan contoh ini ke kelas target, jika tidak akan kembali ke langkah 1.

ii. Atribut-atribut berada dalam suatu kategori (jika bernilai kontinu, nilai-nilai tersebut didistributkan terlebih dahulu)

iii. Contoh-contoh dipartisi secara rekursif berdasarkan atribut terpilih

- iv. Atribut-atribut uji dipilih berdasarkan heuristic atau pengukuran statistic (misal, *information gain*).

Decision tree menggunakan nilai *entropy* sebagai nilai pembandingan (Hastuti, 2012), rumus *entropy* adalah:

$$\text{Average Entropy} = \sum_b \left(\frac{n_b}{n_i} \right) \times \left[\sum_c - \left(\frac{n_x}{n_b} \right) \log_2 \left(\frac{n_x}{n_b} \right) \right]$$

$$P_b = \frac{\text{Probabilty an instance on a branch } b \text{ is positive number of positive instance on branch}}{\text{total number of instance on branch}} = \frac{n_{bx}}{n_b}$$

Rumus 2.4. Rumus Decision Tree

2.3.3. Naïve Bayes

Naïve Bayes Classifier (NBC) merupakan sebuah pengklasifikasi probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi ketidaktergantungan (independent) yang tinggi. Keuntungan penggunaan NBC adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian, karena yang diasumsikan sebagai *variable independent*, maka hanya varians dari suatu *variable* dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians (Bustami, 2014). Model untuk Naïve Bayes adalah:

$$p(\mathbf{x}|\mathbf{i}) = \frac{p(\mathbf{A}_i) \cdot p(\mathbf{B}|\mathbf{A}_i)}{\sum_{j=1}^c p(\mathbf{j}) \cdot p(\mathbf{x}|\mathbf{j})}$$

Rumus 2.1. Rumus Naïve Bayes

2.4. Pengertian Visualisasi

Visualisasi data merupakan interpretasi di antara banyak bidang. Visualisasi data menggabungkan pembuatan dan kajian dari representasi visual dari data, informasi yang telah diabstraksikan dengan skematis, termasuk atribut atau variabel dari unit informasi (Cleveland, 1993).

Tujuan utama dari visualisasi data adalah untuk mengkomunikasikan informasi secara jelas dan efisien kepada *user* lewat grafik informasi, seperti tabel dan grafik. Visualisasi yang efektif membantu pengguna dalam menganalisa dan penjabaran tentang data dan bukti. Visualisasi membuat data yang kompleks bisa diakses, dipahami dan berguna. *User* bisa melakukan pekerjaan analisis tertentu, seperti melakukan perbandingan atau memahami kausalitas, dan prinsip perancangan dari grafik, seperti memperlihatkan perbandingan atau kualitas. (Cleveland, 1993).

Tabel pada umumnya digunakan saat pengguna akan melihat ukuran tertentu dari sebuah variabel, sementara grafik dari berbagai tipe digunakan untuk melihat pola atau keterkaitan dalam data untuk satu atau lebih variabel (Cleveland, 1993).



Gambar 2.1 Visualisasi Microsoft PowerBI
Sumber (Microsoft.powerbi.com)

2.5. Metode Visualisasi

Dalam membuat visualisasi terdapat beberapa teknik yang dapat digunakan seperti *Pureshare*, *Noetix*, *User Centered Design*, *Pixel*, *Visual Data minig* dan masih banyak metode-metode yang dapat digunakan untuk visualisasi, berikut adalah penjelasan-penjelasan untuk metode-metode dan teknik-teknik visualisasi tersebut

2.5.1. Metode *Pureshare*

Pureshare adalah sebuah metode yang dikembangkan untuk memfasilitasi sebuah *project* yang berhubungan dengan pengukuran dan pengelolaan kinerja suatu organisasi. *Pureshare* menggunakan pendekatan *top-down design* dan *bottom-up implementation* (Hariyanti, 2008).

Pureshare memiliki tahapan-tahapan dalam metodologi, yaitu:

a. Perencanaan dan desain

Melakukan identifikasi kebutuhan data dan informasi, melakukan analisis metainformasi dan konten informasi, dan perbaikan sumber data.

b. Perancangan *prototype*

Melakukan perancangan *prototype* yang didalamnya terdapat perancangan konten informasi dashboard dan juga *layout* desain dashboard.

c. Review Prototype

Melakukan peninjauan kembali terhadap *prototype* dan melakukan perancangan mekanisme navigasi.

d. *Release*

Melakukan implementasi *prototype*, integrasi sumber data, penerapan kontrol *security* dan melakukan validasi dan pengujian sistem.

e. *Maintance*

Pemeliharaan dan perbaikan *dashboard* agar tetap sesuai dengan kebutuhan bisnis dan organisasi.

2.5.2. Metode *Noetix*

Metodologi yang dikembangkan oleh *Noetix* memperhatikan tahapan tahapan dalam SDLC (*Software Development Life Cycle*) dan

memperlihatkan kemampuan *developer* (Hariyanti, 2008). *Noetix* memiliki tahapan-tahapan, yaitu:

a. Identifikasi kebutuhan

Melakukan abstraksi skenario (*high level scenario*) dashboard untuk mengidentifikasi kebutuhan data dan informasi

b. Perencanaan

Melakukan perencanaan hierarki *user interface* dan perencanaan komunikasi dan navigasi dashboard.

c. Perancangan *prototype*

Perancangan desain dashboard dan perancangan konten informasi dashboard.

d. *Implementation & Validation*

Melakukan validasi terhadap informasi dan juga *KPI (Key Performance Indicator)* agar tetap relevan

e. Pengembangan

Melakukan penerapan ke lingkungan operasional, integrasi sistem dengan lingkungan operasionalnya, sosialisasi dan pelatihan penggunaan dashboard.

2.5.3. Metodologi *User Centered Design*

User Centered Design menempatkan *user* sebagai pusat dari proses pengembangan sistem dan tujuan dari pengalaman *user* dalam hal

memperhatikan kepentingan organisasi dalam pembuatan *Dashboard* (Winarko & Mulyana, 2009).

Prinsip yang harus di perhatikan dalam *User Centered Design* adalah:

A. Fokus pada pengguna

Perancangan harus terkoneksi langsung dengan pengguna akhir melalui wawancaram survey dan *workshop* pada saat perancangan. Tujuannya adalah untuk memahami kognisi, karakter dan sikap pengguna. Serta karakteristik *anthropometric*. Aktifitas utamanya mencakup pengambilan data, analisis dan integrasikedalam informasi perancangan dari pengguna tentang karakteristik tugas, lingkungan teknis didalam organisasi.

B. Perancangan Terintegrasi

Perancangan harus mencakup antar muka pengguna, sistem bantuan, dukungan teknis serta prosedur untuk instalasi dan pengaturan konfigurasi

C. Dari Awal Berlanjut Pada Pengujian Pengguna

Satu-satunya pendekatan yang berhasil dalam perancangan sistem yang berpusat pada pengguna adalah secara empiris dibutuhkan *observasi* tentang kelakuan pengguna, evaluasi *feedback* yang cermat, wawasan pemecah terhadap

masalah yang ada, dan motivasi yang kuat untuk mengubah rancangan

D. Perancangan Interaktif

Sistem yang sedang dikembangkan harus didefinisikan, dirancangan, dan diuji berulang kali. Berdasarkan hasil *test* kelakuakn dari fungsi, antarmuka, sistem bantuan, dokumentasi pengguna, dan pendekatan pelatihan

Tahapan-tahapan dalam metodologi *User Centered Design* adalah (Winarko & Mulyana, 2009):

a. *Plan the human centered proses*

Memahami dan menentukan konteks pengguna. Tahap ini dilakukan diskusi terhadap orang-orang yang akan mengerjakan proyek, untuk mendapatkan komitmen bahwa proses pembangunan proyek adalah berpusat kepada pengguna atau *user*. Proyek akan memiliki waktu dan tugas untuk melibatkan prngguna dalam awal dan akhir proses atau dimana mereka dibutuhkan.

b. Mendefinisikan konteks

Tahapan ini harus bias memahami kebutuhan pengguna dan mampu berkomunikasi dengna pengguna untuk menentukan fitur-fitur kunci yang akan di tampilkan pada *dashboard*.

- c. Mendefinisikan *user requirement* dan *organizational requirement*

Mengidentifikasi kebutuhan pengguna dan kebutuhan organisasi dan juga tahapan ini dilakukan dengan pendekatan *bottom-up implementation*, yaitu dengan melakukan identifikasi sumber data, cara pengaksesan dan membuat ukuran suatu kualitas data.

- d. *Product design solution*

Membangun desain sebagai solusi dari produk yang sedang dianalisis atau membangun sebuah *prototype*

- e. Mengevaluasi desain dengan *user requirement*

Dari *prototype* yang telah dibuat dan melalui tahap *review* bersama dengan *user* untuk mengumpulkan umpan balik dan dari umpan balik maka akan di kembangkan dan disesuaikan dengan kebutuhan *user*.

2.5.4. Metode Visualisasi *Data minig*

Teknik visualisasi *data minig* merupakan teknik visualisasi yang diaplikasikan pada tahap *pre-processing* pada *data minig*, sedangkan teknik visualisasi terdapat pada tahap *post-prosessing* (Soukup & Davidson, 2002).

Terdapat 8 langkah dalam pembuatan *visual data minig*, yaitu (Soukup & Davidson, 2002):

A. *Justify and Planning the Data Visualization and Data*

Pada tahap ini menjelaskan tentang bagaimana melakukan dan membuat perencanaan dalam membuat data visualisasi, mulai waktu dan sumber daya yang dibutuhkan. Tahap ini merupakan penggerak utama dari proses pembuatan visualisasi data dan sangat berperan dalam kesuksesan sebuah data visualisasi.

B. *Identifying the Top Business Questions*

Tahap ini akan menentukan pertanyaan-pertanyaan bisnis yang digunakan untuk menjawab pertanyaan bisnis.

Langkah-langkah tersebut adalah:

- a. Mengidentifikasi pertanyaan bisnis yang akan di analisa
- b. Menentukan pertanyaan bisnis yang harus di investigasi dan dapat dipetakan ke dalam definisi permasalahan yang dapat di atasi dengan model data visualisasi
- c. Jenis *output* yang diharapkan
- d. Menentukan dan mengukur tujuan dan kriteria kesuksesan yang ingin di capai

C. *Choosing the Data*

Mendapatkan data-data yang akan digunakan bukanlah hal yang mudah, hal pertama yang harus

dilakukan dalam membangun sebuah data visualisasi adalah mengetahui sumber data yang akan digunakan, bagaimana cara mendapatkan data tersebut, tipe data apa saja yang akan digunakan. Data-data yang telah di dapat kemudian akan dipilih.

D. *Transforming Data Set*

Tahapan ini merupakan tahapan lanjutan dari tahap *Choosing the Data* dimana pengolahan data akan dilakukan untuk mendapatkan data baru untuk pengembangan *data minig*. Hal ini dilakukan untuk meningkatkan hasil dari visualisasi tersebut.

E. *Verifying the Data Set*

Identifikasi dan pengujian terhadap *data set* yang telah diolah dengan teknik *ETL* untuk memastikan tidak adanya *error* dan tidak menimbulkan bias.

F. *Choosing the Visualization or Data minig Tools*

Pemilihan *Tools* yang tepat untuk digunakan dan sesuai dengan kebutuhan untuk menginvestigasi *data set* baru yang berasal dari *exploratory data mart* kemudian berfokus pada teknik visualisasi.

G. *Analyzing the Visualization or Data mining Model*

Tahap ini merupakan tahap analisa data visualisasi dan secara visual mengevaluasikan hasil dari visualisasi yang ada untuk mencapai tujuan utama.

H. *Verify and Presenting Analysis*

Tahap ini merupakan tahap terakhir pada *Visual Data Mining*, dimana terdapat tiga bagian, yaitu:

- a. Memverifikasi bahwa model visualisasi sudah memuaskan, telah mencapai tujuan, memenuhi tujuannya dan secara signifikan menjawab pertanyaan bisnis
- b. Mempersiapkan presentasi dari visualisasi yang telah dibuat atas temuan yang terdapat di dalamnya untuk diberikan kepada pengambil keputusan.
- c. Mengembangkan atau membuat visualisasi

Berikut beberapa teknik dalam *visual data mining*: (Soukup & Davidson, 2002):

A. *A Proof of Concept VDM*

A proof of concept visual data mining memiliki cakupan yang terbatas, dimana ruang lingkup keseluruhan proyek proof of concept adalah untuk menentukan visualisasi dan *data mining*

akan bermanfaat bagi suatu proyek untuk membuktikan kepada pembuat keputusan tentang nilai visualisasi dan *data minig* dengan visualisasi dan konsep data. konsep *data minig* biasanya memiliki satu atau dua pertanyaan bisnis yang diselidiki. Kumpulan data yang telah dianalisis secara terbatas pada sampel data yang ada. Rata-rata durasi *proof of concept project* adalah beberapa minggu.

B. A Pilot VDM

A *pilot visual data minig* memiliki cakupan ruang lingkup yang terbatas. Ruang lingkup keseluruhan *pilot visual data minig* adalah untuk menyelidiki, menganalisis, dan menjawab satu atau lebih pertanyaan bisnis untuk menentukan apakah ROI dari penelitian tersebut dapat memastikan sebuah produksi proyek. Data yang di analisis merupakan data terbatas yang mewakili Sumber data asli. Rata-rata durasi data yang digunakan pada *pilot visual data minig* adalah beberapa bulan.

C. Production VDM

A *production visual data minig* memiliki jangkauan project yang hampir sama dengan *a pilot visual data minig* yang memiliki hasil berupa visualisasi dan juga model *data minig* yang di implementasikan ke cakupan produksi. Cakupan proyek adalah untuk meninvegasi penuh, menganalisa, dan menjawab pertanyaan-pertanyaan bisnis dan mengimplementasikan suatu tindakan dari rencana dan mengukur hasil akhir dari produksi. *Production data*

minig memiliki rata-rata durasi data yang digunakan adalah data bertahun tahun yang lalu.

2.6. Data Staging

Data Staging menyediakan sebuah tempat dan area dengan sekumpulan fungsi untuk membersihkan, merubah, mengkombinasikan, penggandaan dan menyiapkan Sumber data untuk penyimpanan dan digunakan pada *data warehouse* (Ali, 2011).

Terdapat tiga proses pada *Data Staging*, yaitu: *Extraction, Transformation, Loading*. Tujuan dari *Data Staging* adalah mengumpulkan, menseleksi, mengolah dan menggabungkan data yang relevan dari berbagai Sumber untuk disimpan dalam *data warehouse* (Ali, 2011).

2.6.1.Extraction

Langkah pertama dari skenario *ETL(Extract, Transform, Loading)* adalah *extract*. *ETL* memiliki tugas dalam mengekstrak data dari sumber sistem. Setiap sumber data memiliki karakteristik yang berbeda yang perlu di atur dalam sebuah proses, proses tersebut perlu di mengintegrasikan berbagai sumber data yang memiliki berbeda-beda tipe (Berliantara, Wicaksono, & Pinandito, 2017).

Pada saat melakukan proses ekstraksi dari data hal-hal yang harus diperhatikan adalah (Berliantara, Wicaksono, & Pinandito, 2017):

- a. Menggunakan ODBC\JDBC (*Open Database Connectivity\ Java Database Connectivity*) drivers untuk mengkoneksikan *database*

- b. Mengerti struktur dari sumber data
- c. Sumber data memiliki karakteristik yang berbeda-beda seperti *mainframes*

Proses *Extraction* terdiri dari dua fase yaitu *initial extraction* dan *changed data extraction*. *Extraction* dapat memungkinkan kita untuk mendapatkan data dari berbagai Sumber operasional dan kemudian dimuat dalam *data warehouse*. Proses ini hanya dilakukan sekali setelah membangun *data warehouse* dengan data dalam jumlah besar dari *source system* (Berliantara, Wicaksono, & Pinandito, 2017).

2.6.2. Transformation

Tahap ke dua dalam pembuatan *ETL* adalah *Transformation*, dimana pada langkah ini data akan di *cleaning* dan akan menyesuaikan data untuk mendapatkan data yang *correct, complete, consistent*, dan *unambiguous* (Berliantara, Wicaksono, & Pinandito, 2017).

Data Cleansing adalah tindakan untuk mendeteksi dan memperbaiki suatu *record* yang memiliki *noise* (data tidak lengkap, tidak benar, tidak relevan) dan kemudian harus diganti, dimodifikasi atau dihapus dengan tujuan mendapatkan data yang berkualitas tinggi. Menurut Mark Mosley dalam bukunya "*Dictionary of Data Management*", pengertian kualitas data adalah *level data* yang menyatakan data tersebut akurat (*accurate*), lengkap (*complete*), *timely (update)*, konsisten

(*consistent*) sesuai dengan semua kebutuhan peraturan bisnis dan relevan.
(Winarko, 2009)

Data berkualitas tinggi memiliki seperangkat kriteria (Aryani1, Honggowati, Djuminah, & Widyaningdyah, 2014), diantaranya:

1. *ACCURACY*: data yang tersimpan nilainya benar
2. *DOMAIN INTEGRITY*: nilai atributnya sesuai batasan yang diperkenankan (nilai atribut laki dan perempuan)
3. *DATA TYPE*: Nilai data disimpan dalam tipe data yang sesuai (data nama disimpan dengan tipe text)
4. *CONSISTENCY*: nilai sebuah *field* data akan sama semua dalam berbagai berkas (*field* produk A dgn kode 123, akan selalu sama kodenya di setiap berkas lain)
5. *REDUDANCY*: tidak boleh ada data yang sama disimpan di tempat yang berbeda dalam satu sistem
6. *COMPLETENESS*: Tidak ada nilai attribut salah yang diberikan dalam sistem
7. *DATA ANOMALY*: sebuah *field* hanya digunakan sesuai kegunaannya. (*Field* nama digunakan untuk nama, dan tidak digunakan untuk memasukan nomor fax atau alamat)
8. *CLARITY*: Kejelasan arti kegunaan dan cara penulisan sebuah data (penamaan khusus)

9. *TIMELY*: merepresentasikan waktu dari data yang dimasukkan (jika data digunakan perhari maka data pada warehouse harus juga dibuat per hari)
10. *USEFULNESS*: setiap data harus benar digunakan oleh *user*
11. *ADHERENCE TO DATA INTEGRITY RULES*: taat pada aturan keterhubungan data

Ada beberapa faktor yang membuat kualitas data sulit untuk diwujudkan yang di sebut *Data Quality Challenges* (Aryani1, Honggowati, Djuminah, & Widyaningdyah, 2014), faktor-faktor tersebut adalah:

a. *Data Source* :

- i. Penuaan Data
- ii. Kesalahan Inputan
- iii. Penipuan
- iv. Sistem Konversi yang salah

b. Validasi Data :

- i. Terjadi duplikasi data
- ii. Beberapa atribut data tercampur dalam *field* yang sama
- iii. Ejaan yang berbeda untuk nama yang sama
- iv. Tidak adanya atribut kunci (*Primary Key*)

Data yang bersih ataupun berkualitas baik dapat memberikan beberapa manfaat, yaitu (Aryani1, Honggowati, Djuminah, & Widyaningdyah, 2014):

- a. Meningkatkan kepercayaan diri dalam pengambilan keputusan

- b. Memperbaiki pelayanan kepada *customer*
- c. Meningkatkan kesempatan memperbaiki kinerja,
- d. Mengurangi resiko dari keputusan yang berbahaya,
- e. Mengurangi biaya, terutama untuk pemasaran,
- f. Mengembangkan strategi untuk pembuatan keputusan,
- g. Meningkatkan produktivitas dengan memangkas beberapa proses
- h. Menghindari efek komplikasi dari data yang terkontaminasi

Data Cleansing digunakan dalam *staging area*. *Data Cleansing* berguna untuk melakukan standarisasi, dan verifikasi data terhadap data yang tidak dikenali.

Beberapa fitur yang terdapat dalam *Data Cleansing* adalah (Mosley, 2008):

- a. *Data Parsing*: Memecah nilai dalam bentuk yang kecil untuk digunakan pada tahap selanjutnya.
- b. *Data Standarization*: Mengubah elemen – elemen data kedalam bentuk standar.
- c. *Data Correction and verification*: Melakukan pengecekan data terhadap data yang tidak dikenali, terutama pada lokasi atau alamat.
- d. *Data Transformation*: Menjamin konsistensi pemetaan data dari sistem Sumber dengan data warehouse.
- e. *Householding*: Menggabungkan individual data yang memiliki kesamaan alamat

2.6.3. Data Loading

Loading data ke target struktur *multidimensional* adalah langkah terakhir pada *ETL*. Pada tahap ini data yang telah di *extract* dan di *Transform* akan tertulis di struktur dimensional yang dapat di akses oleh *end users* dan juga dari *application systems*. *Loading* mencakup langkah *dimension tables* dan *loading fact tables*. Data yang dikirim akan masuk ke data warehouse atau ke aplikasi *reporting* lainnya dalam bentuk yang berbeda, misalkan data yang semula berformat .doc akan menjadi csv atau format data lainnya seperti sql (Berliantara, Wicaksono, & Pinandito, 2017).

2.7. Chart

Berikut pengertian masing-masing *chart*:

a. Grafik batang (Bar diagram)

Merupakan grafik yang berbentuk batang pada bidang sumbu tegak dimana setiap batang mewakili nilai dari suatu variabel atau komponen variabel. Tinggi dari batang bersifat proporsional yang menunjukkan nilai dari variabel yang diwakili. Grafik batang digunakan untuk membandingkan beberapa variabel dalam waktu dan tempat yang sama atau satu *variable* dalam waktu dan tempat yang berbeda. Ada dua jenis grafik batang yaitu (Wahyuni, 2015):

- i. Histogram dimana batang batang disusun berimpitan satu dengan yang lainnya tanpa ruang antara yang merupakan data kuantitatif yang *continue* dalam bentuk distribusi frekuensi.
 - ii. Poligon dimana titik tengah dari puncak setiap batang dihubungkan oleh suatu garis Grafik batang bisa digunakan baik pada analisis *univariate* maupun *bivariate*.
- b. Grafik lingkaran atau diagram *pie/ circle cart*

Penyajian data dalam bentuk lingkaran dimana frekuensi variabel yang dilukiskan sesuai dengan luas (jumlah derajat) dalam lingkaran tersebut. Umumnya grafik ini digunakan untuk menarik perhatian, terutama untuk presentasi oral. Untuk memaksimalkan penggunaan diagram lingkaran, jumlah komponen dari suatu lingkaran sebaiknya tidak lebih dari 10 buah (meskipun ada juga yang berpendapat bahwa enam adalah jumlah yang maksimal). Grafik lingkaran biasanya digunakan untuk penyajian data *univariate* (Wahyuni, 2015).

c. Grafik garis

Informasi digambarkan dalam bentuk garis pada bidang satu sumbu tegak. Garis bisa dalam bentuk tunggal, ganda maupun majemuk. Grafik garis sangat tepat digunakan untuk menggambarkan pertumbuhan/ perkembangan (Wahyuni, 2015).

d. Grafik pencar (scatter plot)

Grafik pencar dihasilkan dari titik koordinat (x,y) dan merupakan grafik korelasi atau kecenderungan karena digunakan untuk mengetahui hubungan antara dua variabel yang berpasangan. Titik titik pada scatter menunjukkan pengaruh suatu variabel terhadap variabel yang lain. Grafik pencar biasanya digunakan jika data yang besar. Hubungan antara variabel bisa positif, negatif atau tidak berhubungan. Grafik pencar digunakan menyajikan data dari analisis bivariate (Wahyuni, 2015).

e. Grafik peta/map diagram

Grafik ini digunakan untuk memperlihatkan perbedaan situasi atau kondisi area dalam suatu wilayah yang digambarkan dengan menggunakan warna. Penggunaan peta biasanya digunakan untuk data deskriptif (Wahyuni, 2015).

2.8. Tools

Berikut adalah alat bantu untuk membuat visualisasi dan prediksi:

2.8.1. Tools Visualisasi



Gambar 2.2. Magic Quadrant for BI Sumber (Microsoft.powerbi.com)

Terdapat beberapa *Tools* untuk memvisualisasikan data akan tetapi berdasarkan *Magic Quadrant For Business Intelligence and Analytic Platforms* terdapat 3 *Tools* yang menjadi leaders di *Business Intelligence and Analytic Platforms*, yaitu:

1. *Microsoft PowerBI*



Gambar 2. 3. Microsoft PowerBI Sumber (Microsoft.powerbi.com)

Microsoft PowerBI Desktop dibangun untuk menganalisis. *Tools* Ini menggabungkan visualisasi interaktif termutakhir, dengan industri terkemuka permintaan data bawaan. Membuat dan menerbitkan laporan ke *PowerBI*. *PowerBI Desktop* membantu Anda memberdayakan orang lain dengan wawasan kritis yang tepat waktu, kapan pun, di mana pun.

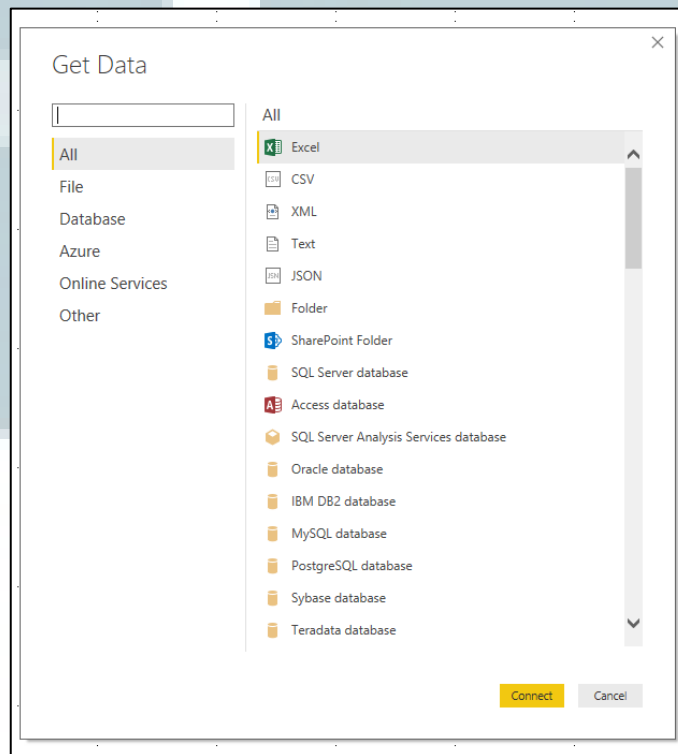
Microsoft PowerBI Desktop adalah sebuah aplikasi *Desktop* pendamping untuk *PowerBI*. Fasilitas yang di berikan oleh *PowerBI* antara lain (Microsoft, 2017):

a. Mendapatkan data:

PowerBI Desktop memudahkan menemukan data.

Anda dapat mengimpor data dari berbagai Sumber data.

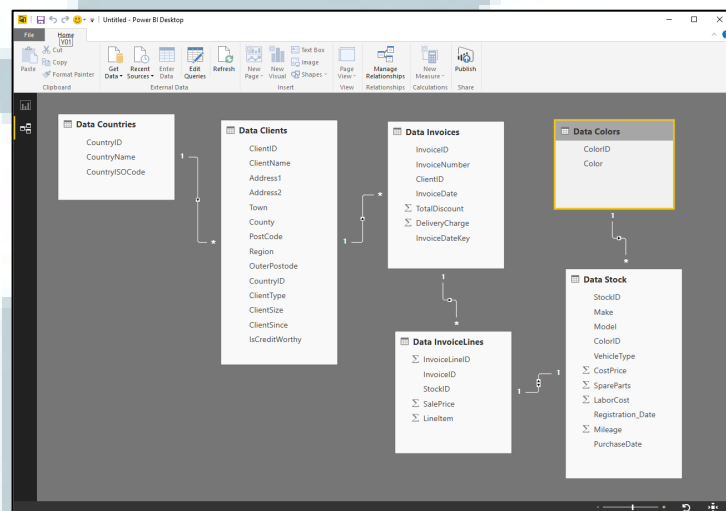
Setelah Anda terhubung ke Sumber data, Anda dapat membentuk data untuk mencocokkan kebutuhan analisis dan pelaporan Anda.



Gambar 2.4. Data Input PowerBI Sumber (Microsoft.powerbi.com)

b. Membuat hubungan dan memperkaya model data dengan format pengukuran baru dan format data:

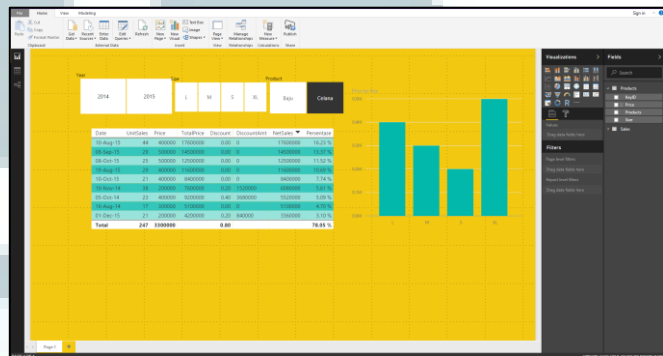
Bila kita mengimpor dua tabel atau lebih, seringkali Anda harus untuk membuat hubungan antar tabel tersebut. *PowerBI Desktop* menyertakan dialog *ETL* (Extract, Transform, Load) dan tampilan *ETL*, yang dapat menggunakan Autodetect agar *PowerBI Desktop* dapat menemukan dan menciptakan hubungan apa pun antar tabel, atau kita dapat membuat hubungannya sendiri. Kita juga dapat dengan mudah membuat pengukuran baru dan perhitungan kita sendiri atau menyesuaikan format data dan kategori untuk memperkaya data dan wawasan tambahan.



Gambar 2. 5. Relationship PowerBI

c. Membuat laporan

PowerBI Desktop juga menyediakan tampilan laporan. Kita bisa memilih grafik yang kita inginkan, menambahkan filter, kita dapat memilih puluhan visualisasi, laporan yang dapat dikustom, gradien, dan beberapa pilihan lain.



Gambar 2. 6. Visualisasi PowerBI

d. Menyimpan laporan Anda

Dengan *PowerBI Desktop*, kita dapat menyimpan pekerjaan kita sebagai berkas *PowerBI Desktop*. *PowerBI Desktop* file memiliki ekstensi *.pbix*.

e. Mengunggah atau menerbitkan laporan Anda

PowerBI dapat mengunggah laporan yang dibuat dan disimpan di *Desktop* ke situs *PowerBI*. *PowerBI* juga dapat menerbitkannya langsung ke *PowerBI.com* dari *PowerBI Desktop*.

B. Tableau



**Gambar 2. 7. Logo Tableau
Sumber (tableau.com)**

Tableau adalah alat *Business Intelligence* untuk menganalisis data. *User* dapat membuat dan mendistribusikan *dashboard* interaktif dan dapat menggambarkan tren, variasi dan kepadatan data dalam bentuk grafik dan diagram. *Tableau* dapat terhubung ke file, relasional dan Sumber data yang besar untuk memperoleh dan mengolah data. Perangkat lunak ini memungkinkan pencampuran data dan kolaborasi *real time*, yang membuatnya menjadi unik. Hal ini digunakan oleh bisnis, peneliti akademis dan banyak pemerintah untuk melakukan analisis data visual. Hal ini juga diposisikan sebagai pemimpin *Business Intelligence* dan *Analytics platform* di *Gartner Magic Quadrant*. *Tableau* dapat menyambungkan berbagai Sumber data baik *data cloud* maupun *data prem*. Data yang dapat digunakan oleh *tableau* adalah (Tableau, 2017):

- a. *Database SQL*
- b. *Spreadsheet*
- c. *Google Ananytics*

d. *Salesforce*

User juga dapat melakukan *pivot*, *split*, dan mengelola metadata untuk mengoptimalkan data.

Tableau dapat membuat analisis data yang luar biasa, bukan hanya sekedar *dashboard* yang cantik, *tableau* dapat membuat perhitungan yang cepat dari data yang ada.



**Gambar 2. 8. Visualisasi *Tableau*
Sumber (*Tableau.com*)**

C. Qlik



**Gambar 2.9. Logo Qlik
Sumber (Qlik.com)**

Qlik Sense adalah aplikasi business intelligence yang paling powerful dan sangat mudah digunakan untuk standar visualisasinya. *User interface* yang digunakan mudah untuk dimengerti untuk orang awam sekalipun. *Qlik Sense* juga powerful dari segi integritas data dan pengolahan data yang sangat dinamis. Banyak fitur-fitur yang dapat disajikan di *Qlik Sense* antara lain adalah grafik-grafik berbentuk bar, pie, map maupun kombinasi *chart* yang dinamis dan saling terkait satu sama lain yang sangat praktis untuk anda yang ingin menganalisa data secara cepat dan akurat tanpa harus mengolah data-data. Selain fitur grafik ada juga fitur-fitur lain seperti filter, table maupun pivotable dan banyak fitur2 lain yang sangat membantu dalam mengolah ataupun menganalisa data.

Dari *user* sendiri relatif mudah yang menggunakan drop and drag saja dan sedikit pengaturan propertise nya dan Sumber data bisa melalui data dengan format *Excel*, atau bisa juga dikoneksikan dengan database lain misalnya sql server, mysql

maupun akses dengan pengkoneksian yang cukup mudah dengan menggunakan odbc.

Keunggulan dari *Qlik* (Qlik.com, 2017):

a. Mudah dan intuitif.

Membuat dan berbagi menampilkan data dinamis dengan *drag-and-drop visualisasi dan layout*. Tidak menggunakan *coding*.

b. Visual dan fleksibel.

Menerjemahkan data besar yang kompleks dari berbagai Sumber ke dalam grafis interaktif yang mudah dipahami dan siap digunakan.

c. Dinamis dan interaktif.

Dapat membantu membuat keputusan yang tepat waktu tanpa dibatasi oleh laporan statis. *Eksplorasi data yang real-time*.

d. Dapat digunakan banyak perangkat

Qlik menggunakan HTML 5, teknologi *browser* dan desain sepenuhnya responsif memberikan *Qlik* pengalaman penuh pada perangkat apapun .

e. Data Integration

Menghubungkan beberapa Sumber data.

f. Aman dan efisien .

Qlik menyediakan data yang lengkap dan tata pengguna melalui data terpusat dan terkontrol.

g. Dapat menggunakan database apapun.

Menggabungkan semua data *user* ke *Data Qlik* atau menggunakan investasi yang ada (*Sql Server, Mysql, Excel, Oracle dan lain-lain*).

h. Lebih cepat.

Menghitung dan agregat dengan cepat . Super responsif memori dalam pengolahan data dan mengoptimalkan daya.



Gambar 2. 10. Visualisasi *Qlik*

2.8.2. Tools Prediksi



Gambar 2.11. Magic Quadrant Mining Tools
Sumber (ibm.com)

Menurut Gather Magic Quadrant for Data Science Platforms, terdapat beberapa *Tools* untuk *data mining*. Terdapat 3 *platforms* yang memimpin platform-platform tersebut, yaitu:

1. IBM SPSS Statistics



Gambar 2. 12. Logo SPSS
Sumber (ibm.com)

SPSS Statistics® adalah *software* statistik utama yang digunakan untuk memecahkan berbagai masalah bisnis dan penelitian. SPSS menyediakan berbagai teknik termasuk analisis ad-hoc, pengujian dan pelaporan hipotesis sehingga memudahkan pengelolaan data dan melakukan analisis, dan membagikan hasil perhitungan. Fitur baru termasuk statistik Bayesian dan integrasi terhadap perangkat lunak pihak ketiga yang disempurnakan. Statistik SPSS menawarkan edisi dasar dengan *add-on* opsional untuk memperluas kemampuan analisis prediktif. SPSS juga menawarkan beberapa fitur, yaitu (IBM SPSS Statistics, 2018):

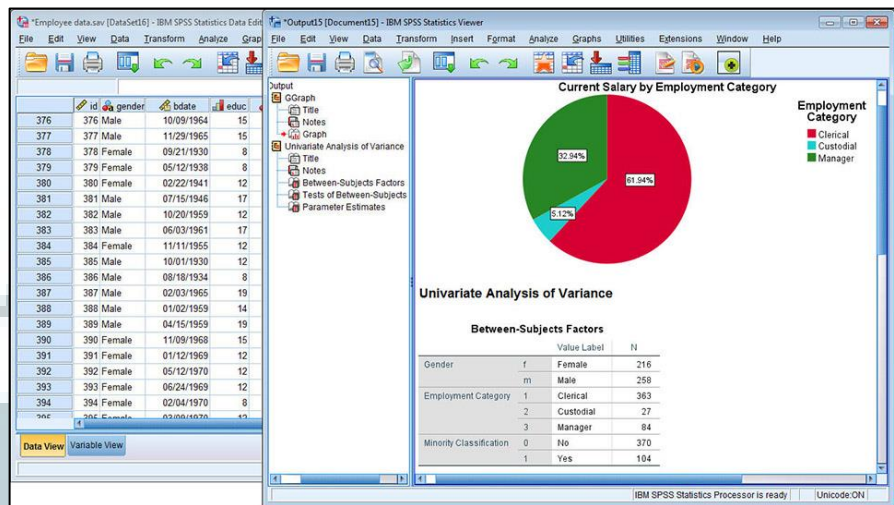
SPSS bekerja pada satu *user interface* terpadu untuk menjalankan statistik deskriptif, regresi, statistik lanjut dan banyak lagi. Membuat publication-ready *charts*, tabel, dan juga *decision trees* dalam satu *tool*

a. *Easy statistical analysis*

SPSS menggunakan *user interface drag and drop* sederhana untuk mengakses berbagai fitur dan mengerjakan beberapa *task* dari berbagai Sumber data.

b. *Integration with Open Source*

SPSS juga terintegrasi dengan beberapa *Open Source* seperti R, Python dan 130+ *hub* lainnya.



Gambar 2. 13. SPSS Output

2. SAS Enterprise Miner



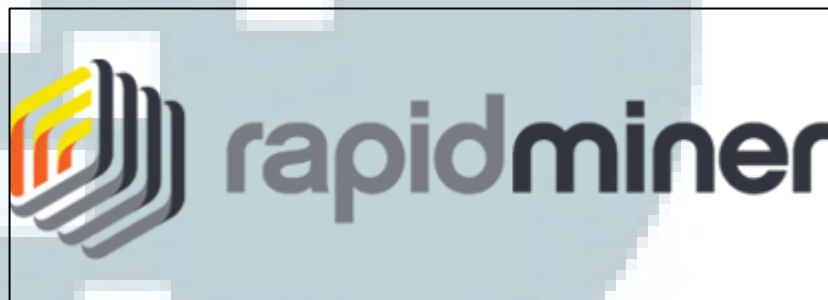
Gambar 2. 14. Logo SAS
Sumber (sas.com)

Merupakan *Tools* data yang *descriptive* dan *predictive* yang memberikan wawasan yang membantu dalam pengambilan keputusan yang lebih baik. Proses *data mining* akan menjadi lebih mudah untuk pengembangan model.

Keunggulan *SAS Enterprise Miner* adalah (SAS Enterprise Miner, 2008):

- a. Mempersingkat waktu ketika mendevlop model

- b. Proses *self-documenting*
 - c. Banyak model predictive
 - d. Dapat membuat model sendiri
 - e. GUI yang mudah digunakan
 - f. Hasil analisa mudah untuk dimengerti untuk pengambilan keputusan yang lebih baik
 - g. Meningkatkan akurasi prediksi
3. RapidMiner



**Gambar 2.15. Logo Rapidminer
Sumber (Rapidminer.com)**

RapidMiner adalah salah satu software untuk pengolahan *data mining*. Pekerjaan yang dilakukan oleh RapidMiner text mining adalah berkisar dengan analisis teks, mengekstrak pola-pola dari data set yang besar dan mengkombinasikannya dengan metode statistika, kecerdasan buatan, dan *database*. Tujuan dari analisis teks ini adalah untuk mendapatkan informasi bermutu tertinggi dari teks yang diolah.

RapidMiner menyediakan prosedur *data mining* dan *machine learning*, di dalamnya termasuk: *ETL* (*extraction, Transformation, loading*), data preprocessing, visualisasi, modelling dan evaluasi. Proses *data mining* tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI. Penyajiannya dituliskan dalam bahasa pemrograman Java (nendrabertus, 2018).

