BAB 2
LANDASAN TEORI

2.1 Surat

Surat merupakan sarana komunikasi tertulis yang disampaikan kepada pihak lain, adapun Pengertian surat menurut kamus besar bahasa Indonesia “surat merupakan secarik kertas yang berisi maksud dan keterangan dan sebagainya yang dapat digunakan sebagai tanda”. Dra. Sedarmayanti M.Pd menyatakan bahwa definisi surat adalah “ alat komunikasi tertulis yang berasal dari satu pihak dan ditujukan kepada pihak lain untuk menyampaikan berita” [11].

Berdasarkan definisi surat tersebut dapat disimpulkan bahwa surat merupakan sehelai kertas yang berisi paparan mengenai maksud dan tujuannya menulis, yang dapat digunakan sebagai alat komunikasi atas nama pribadi maupun organisasi untuk menyampaikan suatu informasi.

2.1.1 Jenis Surat

Dalam berkomunikasi surat memiliki fungsi masing – masing sesuai dengan macamnya. Adapun surat dapat dibedakan menjadi beberapa macam diantarannya sebagai berikut [12]:

1. Macam surat menurut wujudnya yaitu dapat berwujud kartu pos, warkat pos, surat bersampul, memorandum, telegram dan surat pengantar.
2. Macam surat menurut tujuannya meliputi surat pemberitahuan, perintah, permintaan, peringatan, panggilan, susulan, keputusan, laporan, perjanjian, dan penawaran pemesanan.
3. Macam surat menurut isi dan asalnya dapat dibagi menjadi surat dinas, niaga, pribadi, dan surat yang berisi masalah - masalah sosial.
4. Macam surat menurut jumlah penerima yaitu surat biasa untuk satu orang, surat edaran untuk beberapa orang atau pejabat, sedangkan surat pengumuman untuk masyarakat.
5. Macam surat menurut keamanan isi surat yaitu sangat rahasia, surat rahasia, dan surat biasa.

6. Macam surat menurut urgensi penyelesaiannya yaitu surat sangat rahasia, surat segera dan surat bisa.
7. Macam surat menurut prosedur pengurusannya yaitusurat masuk dan surat keluar.
8. Macam surat menurut jangkauannya yaitu surat intern dan ekstern.

Berdasarkan uraian tersebut macam surat dapat dibedakan menjadi beberapa macam surat. Macam surat menurut wujud, menurut tujuannya, menurut isi, menurut jumlah menurut keamanan dan menurut prosedur pengelolaannya.

2.1.2 Surat Masuk

Surat masuk adalah surat dinas yang diterima oleh sebuah lembaga atau organisasi yang memiliki prosedur dan aturan dalam penyusunan surat. Surat dinas memiliki aturan susunan yang sudah tertulis resmi bahwa aturan surat dinas sebagai berikut:

1. Kepala

   Bagian kepala Surat Undangan Ekstern terdiri atas:

   a. Surat Undangan Ekstern yang ditandatangani oleh pejabat Sekretariat Jenderal DPD RI menggunakan kop surat yang memuat logo Setjen DPD RI disertai “nama lembaga” dengan huruf kapital secara sime- tris;
   b. “nomor”, “sifat”, “lampiran”, dan “hal”, diketik di sebelah kiri di bawah kop surat undangan ekstern;
   c. “tempat” dan “tanggal” pembuatan surat, diketik di sebelah kanan atas sejajar/sebaris dengan nomor;
   d. kata “Kepada Yth.”, ditulis di bawah hal, diikuti den- gan nama jabatan, dan alamat yang dikirimi surat undangan ekstern (jika diperlukan).

2. Batang Tubuh

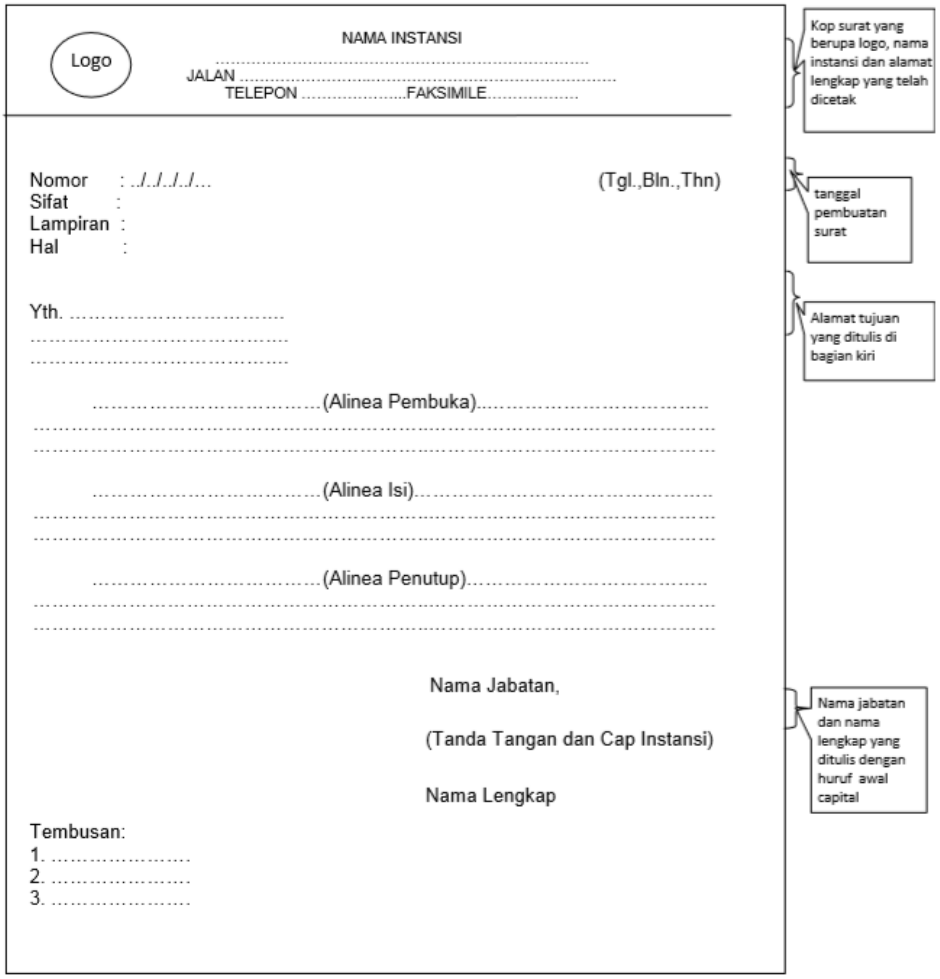
   Bagian batang tubuh Surat Undangan Ekstern terdiri atas:

   a. alinea pembuka;
   b. isi Surat Undangan Ekstern, yang meliputi “hari”, “tanggal”, “waktu”, “tempat”, dan “acara”; dan
   c. alinea penutup.

3. Kaki

Bagian kaki surat undangan ekstern terdiri dari “nama jabatan” yang ditulis dengan huruf awal kapital tanpa mencantumkan gelar, “tanda tangan”, dan “nama peja- bat” yang ditulis dengan huruf awal kapital serta men- cantumkan NIP.

Pada gambar 2.1 adalah contoh surat dinas [13]:



Gambar 2.1 Surat Dinas Resmi

2.2 *Data Mining*

Data Mining adalah langkah analisis terhadap proses penemuan pengetahuan di dalam basis data atau *knowledge discovery in databases* yang disingkat KDD. Pengentahhuan bisa berupa pola data aray relasi antar data yang valid (yang tidak diketahui sebelumnya).

Data mining merupakan gabungan sejumlah disimplin ilmu komputer yang didefinisikan sebagai proses penemuan pola – pola baru dari kumpulan – kumpulan

data sangat besar, meliputi metode – metode yang merupakan irisan dari *artificial intelligence, machine learning, statistics, dan database systems*.

Data mining ditujukan untuk mengekstrak (mengambil intisari) pengetahuan dari sekumpulan data sehingga didapatkan struktur yang dapat dimengerti manusia serta meliputi basisdata dan manajemen data, prapremprosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan = kompleksitas, pascapemrosesan terhadap struktur yang ditemukan, visualisasi dan online updating [14].

2.3 *Text Mining*

Text mining merupakan proses mencari atau mengekstrak informasi yang berguna dari data tekstual. *Text mining* mencoba untuk menemukan pengetahuan dari teks yang tidak terstruktur. Ini juga dikenal sebagai *Text Data Mining* (TDM) dan Penemuan pengetahuan dalam *Textual Databases* (KDT). KDT memainkan peran yang semakin signifikan dalam aplikasi yang muncul, seperti pemahaman teks. Proses penambangan teks sama dengan penambangan data, kecuali, alat penambangan data dirancang untuk menangani data terstruktur sedangkan penambangan teks dapat menangani kumpulan data yang tidak terstruktur atau semi-terstruktur seperti email file HTML dan dokumen teks lengkap dll [15].

Text mining merupakan salah satu proses yang memungkinkan user untuk mengektraksi informasi dari sumber tekstual yang akan berkaitan dengan operasi seperti *retrieval,* klasifikasi (*supervised*, *unsupervised* dan *semi-supervised*) dan *natural language processing*, *data mining,* dan teknik *machine learning* yang bekerja sama untuk mengklasifikasikan secara otomatis dan menemukan pola dari dokumen [16].

Text mining juga dapat diartikan sebagai pemrosesan informasi yang tidak terstruktur secara tekstual, mengektraksi indeks numerik yang bermakna dari teks, dan membuat informasi yang terkandung dalam teks dapat diakses oleh berbagai algoritma *data mining* seperti *machine learning*.

2.4 Ekstraksi Informasi

Ekstraksi Informasi adalah pencarian otomatis pada informasi yang terstruktur seperti entitas, hubungan antar entitas, dan atribut yang menggambarkan

entitas dari sumber yang tidak terstruktur [2]. Pencarian dari suatu ekstraksi informasi merupakan sebuah sistem yang berfungsi untuk mencari data spesifik dalam *Natural Language Text*. Dalam sebuah template formulir atau tabel serta komponen – komponen teks yang ada diambil datanya untuk diekstraksi menjadi data yang memiliki kegunaan [3]. Ekstraksi informasi berperan sebagai sistem yang akan mengenali data yang tidak terstruktur yang memiliki informasi yang belum dikategorikan dan belum memiliki arti yang spesifik. Tidak terstruktur bukan berarti secara data yang membingungkan atau tidak jelas dengan artian bahwa informasti yang terkandung didalamnya tidak dapat langsung diterjemahkan oleh komputer sehingga membutuhkan sistem yang dinamakan ekstraksi informasi.

Ekstraksi informasi dapat digunakan untuk memproses dengan memberikan arti kepada informasi yang tidak terstruktur tersebut, seperti contoh pada data yang terkandung dalam teks, video, gambar dan audio. Pada proses ekstraksi informasi diperlukan suatu proses yang dinamakan *preprocessing* untuk mengidentifikasikan dan mencirikan suatu pola atau tanda dari isi informasi yang terkandung.

2.5 *Converting*

Proses *converting* (konversi) adalah proses pengubahan format data menjadi format data yang diperlukan untuk menuju tahap proses *preprocessing*. Konversi yang dilakukan dalam penelitian ini berupa format file *.pdf menjadi *.txt.

Banyak cara yang dapat dilakukan dengan konversi, seperti menggunakan *library*, menggunakan software ataupun memakai kode API untuk mengakses pada situs yang menyediakan layanan konversi. Tools yang digunakan pada penelitian ini berupa akses API dari pdftables.com untuk mengkonvert data pdf menjadi txt.

2.6 *Preprocessing*

Preprocessing adalah teknik pengambilan data yang melibatkan transformasi data mentah menjadi format yang mudah dipahami. *Preprocessing* memiliki tujuan yaitu untuk mempermudah memahami data sehingga mempermudah pemilihan teknik dan metode *data mining* yang tepat, untuk meningkatkan kualitas data sehingga hasil data mining menjadi lebih baik dan untuk meningkatkan efisiensi dan kemudahan proses data mining [14].

Pada penelitian ini tahapan *preprocessing* meliputi *converting, filtering, tagging,* segmentasi serta pembobotan pada setiap variabel yang akan dijadikan parameter untuk pemrosesan klasifikasi menggunakan algoritma *machine learning*.

2.6.1 *Filtering*

Proses *filtering* dalam istilah *text preprocessing* bisa disebut *Stopwords*. *Stopwords* adalah pembagian bahasa alami. Motif bahwa *stopwords* harus dihapus dari sebuah teks adalah bahwa mereka membuat teks terlihat lebih berat dan kurang penting bagi penelitian. Menghapus *stopword* mengurangi dimensi ruang term. Kata-kata ini diperlakukan sebagai kata-kata henti. *Stopwords* dihapus dari dokumen karena kata-kata tersebut tidak diukur sebagai kata kunci dalam aplikasi penambangan teks [17].

Proses *filtering* (penyaringan) juga merupakan proses dimana segala karakter yang tidak diperlukan akan disaring dan diseleksi agar menjadi teks yang rapi dan bersih agar mudah untuk diidentifikasi.

2.6.2 *Tagging*

Tagging dalam *teks preprocessing* disini bisa disebut sebagai aplikasi kondisional yang digunakan untuk memberikan *tag* kondisional barang atau kata atau frasa tertentu dalam sebuah kalimat tergantung dari kata kunci yang terkandung dalam kalimat yang diidentifikasi oleh *sed command* [18].

Tagging merupakan tahapan untuk memberikan tanda pada kata – kata kunci yang akan menjadi awal kalimat dan sebagai pembatas/pembendung (*stemming*) yang akan di ekstraksi. Proses *tagging* hampir sama dengan *filter* namun pada saat pencarian kata dan pergantian katanya, proses *tagging* disini berperan untuk menandakan suatu kata dengan ekstraksi fitur lalu tag yang sudah dibuat akan diubah menjadi delimiter sebagai pemisah antar kalimat atau kata sehingga bisa suatu kata dapat dilihat kecocokannya terhadap kata kunci yang cocok.

2.6.3 *Tokenization*

Tujuan tahap akhir dari teks *preprocessing* ini adalah untuk proses tokenisasi. Tokenisasi adalah proses memecah aliran konten tekstual menjadi kata-kata, istilah, simbol, atau beberapa elemen bermakna lainnya yang disebut token

[17]. Daftar token berubah menjadi masukan untuk pemrosesan tambahan termasuk penguraian atau penambangan teks.

Jadi tokenisasi merupakan tahapan pembagian string dari data surat yang sudah diberi *tag* sebagai penanda batas yang akan menjadi data inputan yang akan dimasukkan ke dalam tabel untuk pemrosesan.

2.6.4 Pembobotan

Pembobotan atau *term weighting* adalah suatu metode untuk memberikan nilai setiap ciri penting dari setiap ciri yang ada pada dokumen [19]. Pembobotan merupakan faktor penentu sebagai pembeda dan pencirian dalam setiap kalimat yang terkandung. Didalam proses pembobotan akan diperiksa kandungan yang terdapat dalam setiap inputan dari proses *tokenization*.

Pembobotan berperan sebagai proses dimana setiap parameter diberi nilai sesuai dengan fitur yang telah ditentukan sebagai acuan dari pemberian nilai parameter tersebut. *Term weighting* yang dipakai untuk penelitian ini adalah *binary weighting*, yaitu pemberian nilai antara [0,1] jika cocok maka beri nilai 1 begitupun sebaliknya.

Berikut adalah ekstraksi fitur sebagai acuan dari pembobotan dari sumber dan fitur yang diperbaharui yang terdapat pada penelitian ini [3] :

Tabel 2.1 Daftar Ekstraksi Fitur

FITUR LOKAL	
EMAIL	Teks memiliki email
ALLCAPS	Seluruh karakter teks adalah huruf besar (Uppercase)
CONTAINDIGIT	Teks memiliki sedikitnya 1 karakter angka
CONTAINDASH	Teks memiliki sedikitnya 1 strip
CONTAINMORESLASH	Teks memiliki sedikitnya 1 garis miring
INITIALKEYWORD	Teks memiliki sedikitnya 1 kata kunci
WORD	Teks memiliki kata itu sendiri
CONTAINCOLON	Teks memiliki sedikitnya 1 karakter titik dua
STRINGLENGTH	Panjang dari teks lebih 50 karakter
CONTAINCOMMA	Teks memiliki sedikitnya 1 karakter koma
CONTAINDATE	Teks memiliki teks tanggal

2.7 *Artificial Intelligence*

Kecerdasan buatan (*artificial intelligence*) merupakan kawasan penelitian, aplikasi dan instruksi yang terkait dengan pemrograman computer unt7uk melakukan hal yang dalam pandangan manusia adalah cerdas [20].

Artificial intelligence juga disebut sebagai kemampuan komputer digital atau robot yang dikendalikan komputer untuk memecahkan masalah yang biasanya terkait dengan kemampuan pemrosesan sesuai intelektual manusia.

2.8 *Machine Learning*

Machine learning (ML) adalah cabang kecerdasan buatan yang secara sistematis menerapkan algoritma untuk mensintesis hubungan yang mendasar antara data dan informasi [21].

Machine learning berperan sebagai sistem yang mempelajari data untuk diolah yang diprogram sedemikian rupa sesuai dengan kebutuhan untuk menghasilkan data yang diinginkan dari pemrosesan *machine learning* tersebut.

2.9 *Naïve Bayes*

Naïve Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang sesuai dengan teorema Bayes dengan menggunakan asumsi independensi atau ketidaktergantungan yang kuat [14].

Naïve Bayes adalah metode probabilitas dan statistik dari pengklasifikasian yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi atau memperkirakan peluang di masa yang akan *dating* berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema *Bayes*.

Naïve Bayes memiliki kinerja yang bagus dibanding dengan model classifier lainnya. Hal ini dibuktikan oleh Xhemali, Hinde dan Stone dalam jurnalnya "*Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages*" mengatakan bahwa "*Naïve Bayes Classifier* memiliki tingkat akurasi yang lebih baik dibanding model classifier lainnya" [3].

Tabel 2.2 Nilai Perbandingan

Classifier	*Accuracy*	*Precision*	*Recall*	*F-Measure*
NB Classifier	95.20%	99.37%	95.23%	97.26%
DT Classifier	94.85%	98.31%	95.90%	97.09%

The Naïve Bayesian classifier, atau *Simple Bayesian Classifier*, dapat dijelaskan sebagai berikut [14]:

1. Misalkan D adalah himpunan data latih (*training set*) yang berisi sejumlah tuple beserta label kelasnya. Setiap tuple adalah berdimensi n yang dinyatakan sebagai X = (x_1,x_2,….,$x_{n)}$ yang didapat dari n atribut A_1,A_2,…,A_n.
2. Misalkan ada kelas m, C1, C2, ..., Cm. Untuk sebuah tuple, masukan *X* classifier akan memprediksi bahwa tuple X termasuk kelas C_i jika dan hanya jika :

$$P(C_i \mid X) > P(C_j \mid X) \text{ untuk} \leq j \leq m\ , j \neq i \quad (2.1)$$

Dengan kata lain, *Naïve Bayes classifier* bekerja dengan cara memaksimalkan $P(C_i \mid X)$. Kelas C_i yang membuat $P(C_i \mid X)$ bernilai maksimal disebut *maximum posteriori hypothesis.* Dengan teorema Bayes, $P(C_i \mid X)$ diestimasi menggunakan formula

$$P(C_i \mid X) = \frac{P(X \mid Ci)\ P(Ci)}{P(X)} \quad (2.2)$$

Keterangan:

P(Ci|X) = Probabilitas hipotesis Ci jika diberikan fakta atau record X (Posterior probability)

P(X|Ci) = mencari nilai parameter yang memberi kemungkinan yang paling besar (likelihood)

P(Ci) = Prior probability dari X (Prior probability)

P(X) = Jumlah probability tuple yg muncul

1. Mengingat P (X) adalah sama untuk semua kelas, hanya P (X | Ci) P (Ci) yang perlu dimaksimalkan. Jika probabilitas prior untuk setiap kelas tidak diketahui, maka probabilitas setiap kelas biasanya diasumsikan sama, P (C1) = P (C2) = · · · = P (Cm), dengan demikian, *Naïve Bayes Classifier* hanya memaksimalkan P (X | Ci). JNamun, jika probabilitas *prior* untuk setiap kelas bisa berbeda-beda, maka *Naïve Bayes classifier* harus memaksimalkan P(X | Ci) P(Ci).
2. Jika memiliki himpunan data yang memiliki sangat banyak atribut, maka kompleksitas perhitungan P(X|Ci) dapat direduksi dengan asumsi naif tentang independensi bersyarat kelas, yaitu : nilai – nilai atribut adalah saling independen. Artinya antar atribut adalah saling bebas, tidak ada ketergantungan sama sekali. Dengan demikian, Naïve Bayes classifier memaksimalkan :

$$P(C_i \mid X) = \prod_{K=1}^{n} P(x_k \mid C_i)$$
$$= P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \ldots \times P(x_n \mid C_i) \quad (2.3)$$

Berdasarkan tuple – tuple himpunan data latih maka dapat diestimasikan P (x1 | Ci), P (x2 | Ci),. . . , P (xn | Ci) berdasarkan tipe atributnya. Untuk atribut yang bernilai kategorial $P(x_k \mid C_i)$, didefinisikan sebagai jumlah tuple di kelas C_i dalam himpunan data *D* yang memiliki nilai xk pada atribut A_k dibagi dengan jumlah semua tuple di kelas C_i dalam *D* yang disimbolkan sebagai $|C_i , {}_D|$. Sementara itu, untuk atribut yang bernilai kontinu, yang umumnya diasumsikan memiliki distribusi *Gaussian*, $P(x_k \mid C_i)$ didefinisikan sebagai

$$P(x_k \mid C_i) = \frac{1}{\sigma_{c_i}\sqrt{2\pi}} e^{\frac{\left(x-\mu_{c_i}\right)}{2\sigma c_i^2}} \quad (2.4)$$

Dimana μCi dan σCi adalah rata – rata dan deviasi standar dari nilai – nilai pada atribut $_{Ak}$ untuk kelas C_i.

3. Untuk memprediksi label kelas dari tuple *X* anda harus menghitung probabilitas $P(X|C_i)P(C_i)$ untuk setiap kelas C_i. Selanjutnya probabilitas tersebut dimaksimalkan, yaitu mencari kelas C_i yang menghasilkan

probabilititas $P(X|C_i)P(C_i)$ maksimum sebagai kelas keputusan. Secara matematis, tuple *X* diberi label kelas C_i jika dan hanya jika

$$P(X|C_i)P(C_i) > P(X|C_j)\,P(C_j) \text{ untuk } 1 \leq j \leq m, j \neq i \quad (2.5)$$

Dengan kata lain, label kelas diprediksi adalah Ci yang mana P (X | Ci) P (Ci) adalah maksimal. Pengklasifikasi Bayesian memiliki tingkat kesalahan minimal dibandingkan dengan klasifikasi lainnya. Namun, dalam prakteknya hal ini tidak selalu terjadi, karena ketidakakuratan asumsi yang dibuat untuk penggunaannya, seperti kondisi kelas independen, dan kurangnya data probabilitas yang tersedia.

Pengklasifikasi *Bayesian* juga berguna dalam memberikan pembenaran teoritis untuk pengklasifikasi lain yang tidak secara eksplisit menggunakan teorema *Bayes*.

2.10 Pemrograman Terstruktur

Pemrograman Terstruktur adalah proses pengimplementasian urutan langkah penyelesaian suatu masalah dalam bentuk program dan aktifitas pemrograman yang dilakukan secara berurutan dari setiap langkah perintah yang dikerjakan secara sistematis, logis, dan tersusun berdasarkan algoritma yang sederhana dan dapat dengan mudah dipahami [22].

Melalui pendekatan terstruktur, permasalahan yang komplek di organisasi dapat dipecahkan dan hasil dari sistem akam mudah untuk dipelihara, fleksibel, lebih memuaskan pemakainya, mempunyai dokumentasi yang baik, tepat waktu, sesuai dengan anggaran biaya pengembangan, dapat meningkatkan produktivitas dan kualitasnya akan lebih baik (bebas kesalahan).

2.11 *Entity Relationship Diagram*

Entity Relationship Diagram (ERD) merupakan teknik yang digunakan untuk memodelkan kebutuhan data dari suatu organisasi, biasanya oleh Sistem Analis dalam tahap analisis persyaratan proyek pengembangan sistem. Sementara seolah-olah teknik diagram atau alat peraga memberikan dasar untuk desain *database* relasional yang mendasari sistem informasi yang dikembangkan [23].

Diagram Entity-Relationship melengkapi penggambaran grafik dari struktur logika . Dengan kata lain Diagram E-R menggambarkan arti dari aspek data seperti bagaimana *entity*, atribut-atribut dan *relationship* disajikan.

2.12 *Data Flow Diagram*

Data Flow Diagram (DFD) adalah teknik yang menggambarkan proses bisnis dan data yang lewat di antara mereka [24]. DFD berfungsi sebagai alat pembuatan model yang memungkinkan profesional sistem untuk menggambarkan sistem sebagai suatu jaringan proses fungsional yang dihubungkan satu sama lain dengan alur data, baik secara manual maupun komputerisasi.

DFD ini merupakan alat perancangan sistem yang berorientasi pada alur data dengan konsep dekomposisi dapat digunakan untuk penggambaran analisa maupun rancangan sistem yang mudah dikomunikasikan oleh profesional sistem kepada pemakai maupun pembuat program.

2.13 PHP

PHP merupakan secara umum dikenal sebagai bahasa pemrograman script yang membuat dokumen HTML secara *on the fly* yang dieksekusi di server web, dokumen HTML yang dihasilkan dari suatu aplikasi bukan dokumen HTML yang dibuat dengan menggunakan editor teks atau editor HTML dikenal juga sebagai bahasa pemrograman *server side*.

Dengan menggunakan PHP makan *maintenance* suatu situs web menjadi lebih mudah. Proses *update* data dapat dilakukan dengan menggunakan aplikasi yang dibuat dengan menggunakan script PHP.

PHP/FI merupakan nama awal dari PHP. PHP – *Personal Home Page*, FI adalah *Form Interface.* Dibuat pertama kali oleh Rasmus Lerdoff PHP awalnya merupakan program CGI yang dikhususkan untuk menerima iinput melaului form yang ditampilkan dalam browser web. *Software* ini disebarkan dan dilisensikan sebagai perangkat lunak *Open Source.*

Kini, PHP adalah kependekan dari PHP : *HyperText Preprocessor* (rekursif, mengikut gaya penamaan di *nix), merupakan bbahasa utama script *server side* yang disisipkan pada HTML yang dijalankan di server, dan juga bisa digunakan untuk membuat aplikasi desktop [25].

PHP merupakan bahasa pemrograman yang cukup mudah untuk dipelajari dan diakses karena langsung diaplikasikan kedalam kode HTML yang merupakan untuk perancangan website yang dapat diaplikasikan langsung pada browser masing-masing.

2.14 XAMPP

XAMPP adalah *software* gabungan dalam pelayanan *software* Apache, MySQL sebagai *database system*. XAMPP digunakan untuk untuk pemrograman PHP dan pengaksesan database MySQL dalam perancangan web programming. XAMPP secara default menyatukan antara HTML dan PHP, dengan direktori penyimpanann script yang terletak di c:\xampp\htdocs ini disebut sebagai direktori root, isi dari direktori ini yang akan ditampilkan jjika kita mengakses lewat browser web dari komputer yang sama.

XAMPP digunakan karena fiturnya yang sudah cukup lengkap dan dengan GUI yang simple untuk diakses sehingga tidak menyulitkan untuk dipakai.

2.15 Apache

Apache adalah server *web* yang bisa digunakan di Windows. Apache sendiri adalah server *web* yang ada pada setiap sistem operasi yang popular, seperti keluarga Unix, Linux, MacOS, dan Windows. Pemasangan PHP untuk server *web* Apache termasuk paling mudah, karena sudahbanyak paket pemasangan yang memudahkan proses pemasangan, proses pemasangan tidak harus satu per satu komponennya [25].

Apache juga dilengkapi dengan fitur-fitur canggih seperti pesan kesalahan yang dapat dikonfigur, autentikasi berbasis basis data dan lain-lain. Apache pun memili antarmuka pengguna berbasis grafik (GUI) yang cukup mudah diperlajari. Apache merupakan *software* yamg yang telah dikembangkan oleh komunitas terbuka yang terdiri dari pengembang-pengembang dibawah naungan Apache Software Foundation.

2.16 MySQL

SQL adalah bahasa standar yang digunakan untuk melakukan akses dan memanipulasi *database*, dalam hal ini *database* relasional. Perintah SQL yang akan digunakan untuk mengakses harus disusun menjadi string SQL terlebih dahulu [25].

Mysql merupakan sistem manajemen basis data relasional atau disebut RDBMS (relational database management system) yang mendukung adanya *relationship* atau hubungan antar tabel. Mysql dapat diakses secara gratis dibawah lisensi GPL (General Public License). DBMS atau (Database Management System) adalah pengelola perangkat lunak untuk digunakan sebagai database, dengan DBMS user dapat mengakses dan mengatur database secara efsien, user pun secara penuh dapat memanipulasi data yang ada pada DBMS.

MySQL adalah sistem penyimpan data (*database*) yang intinya adalah bahwa *database* merupakan tempat yang dapat kita gunakan untuk "menyimpan" data pada web. Dalam sistem informasi komputer *database* adalah sebuah dapat diartikan kumpulan data yang tersusun secara sistematis sehingga dapat memudahkan pengguna untuk mengakses dan mengatur sehingga hasilnya bisa mendapatkan sebuah informasi yang efektif dan efisien.

2.17 Pengujian *Black Box*

Pengujian *black box* merupakan pendekatan komplementer dari teknik *white box*, karena pengujian *black box* diharapkan mampu mengungkap kelas kesalahan yang lebih luas dibandingkan teknik *white box*. Pengujian *black box* berfokus pada pengujian persyaratan fungsional perangkat lunak, untuk mendapatkan serangkaian kondisi input yang sesuai dengan persyaratan fungsional suatu program [26].

Pengujian *black box* adalah pengujian aspek fundamental sistem tanpa memperhatikan struktur logika internal perangkat lunak. Metode ini digunakan untuk mengetahui apakah perangkat lunak berfungsi dengan benar. *Black box* juga dapat dikatakan sebagai metode perancangan data uji yang didasarkan pada spesifikasi perangkat lunak. Data uji dibangkitkan, dieksekusi pada perangkat lunak dan kemudian keluaran dari perangkat lunak dicek apakah telah sesuai dengan yang diharapkan.

Pengujian *black box* berusaha menemukan kesalahan dalam kategori :

1. fungsi-fungsi yang tidak benar atau hilang,
2. kesalahan interface,
3. kesalahan dalam struktur data atau akses database eksternal,
4. kesalahan kinerja, serta
5. inisialisasi dan kesalahan terminasi.

2.18 *Confusion Matrix*

Pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting. Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. *Confusion matrix* merupakan salah satu metode untuk pengukuran kinerja suatu metode klasifikasi. *Confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya [13].

Berdasarkan jumlah keluaran kelasnya, sistem klasifikasi memiliki empat jenis yaitu klasifikasi *binary*, *multi-class*, *multi-label* dan *hierarchical* [10]. Pada klasifikasi *binary*, data masukan dikelompokkan ke dalam salah satu dari dua kelas. Jenis klasifikasi ini merupakan bentuk klasifikasi yang paling sederhana dan banyak digunakan. Contoh penggunaannya antara lain dalam sistem yang melakukan deteksi orang atau bukan, sistem deteksi kendaraan atau bukan, dan sistem deteksi pergerakan atau bukan.

Terdapat empat istilah dalam confusion matrix sebagai representasi hasil proses klasifikasi yaitu (*True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN)). TN atau *True Negative* adalah jumlah tuple negative yang dilabeli dengan benar oleh *classifier*, sedangkan FP atau *False Positive* adalah jumlah tuple *negative* yang salah dilabeli oleh *classifier*. TP atau *True Positive* adalah jumlah tuple positif yang dilabeli dengan benar oleh *classifier*, sedangkan FN atau *False Negative* merupakan jumlah tuple positif yang salah dilabeli oleh *classifier*.

Pada jenis klasifikasi binary yang hanya memiliki 2 keluaran kelas, confusion matrix dapat disajikan seperti pada Tabel 2.2.

Tabel 2.3 *Confusion Matrix*

Kelas	Terklasifikasi Postif	Terklasifikasi Negatif
Positif	TP (True Positive)	FN (False Negative)
Negatif	FP (False Positive)	TN (True Negative)

Berdasarkan nilai *True Negative* (TN), *False Positive* (FP), *False Negative* (FN), dan *True Positive* (TP) dapat diperoleh nilai akurasi, presisi dan *recall*. Nilai akurasi mencerminkan akurasi pada sistem dalam mengklasifikasikan data secara benar dengan membandingan data yang terklasifiakasi dengan benar dengan semua data yang ada

Berikut adalah persamaan untuk mendapatkan nilai akurasi:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (2.6)$$

Nilai presisi adalah gambaran jumlah data kategori positif yang di klasifikasikan dengan benar dibagi dengan total data yang diklasifikasi positif.

Berikut adalah persamaan untuk mendapatkan nilai Presisi:

$$Presisi = \frac{TP}{FP+TP} * 100\% \quad (2.7)$$

Recall menunjukan persentasi dari data kategori positif yang diklasifikasi benar oleh sistem.

Berikut adalah persamaan untuk mendapatkan nilai recall:

$$Recall = \frac{TP}{FN+TP} * 100\% \quad (2.8)$$

dimana:

TP atau *True Positive* : jumlah data positif yang terklasifikasi dengan benar oleh sistem.

TN atau *True Negative* : jumlah data negatif yang terklasifikasi dengan benar oleh sistem.

FN atau *False Negative* : jumlah data negatif namun terklasifikasi salah oleh sistem.

FP atau *False Positive* : jumlah data positif namun terklasifikasi salah oleh sistem.

Pada klasifikasi dengan jumlah keluaran yang *multi-class* atau yang memiliki lebih dari dua kelas adalah sebagai berikut [6]:

$$Akurasi = \frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{l} * 100\% \quad (2.9)$$

$$Presisi = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (FP_i + TP_i)} * 100\% \quad (2.10)$$

$$Recall = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_i + FN_i)} * 100\% \quad (2.11)$$

dimana:

TPi adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem untuk kelas ke-i.

TNi adalah *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem untuk kelas ke-i.

FNi adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem untuk kelas ke-i.

FPi adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem untuk kelas ke-i

l adalah jumlah kelas.