BAB 2
LANDASAN TEORI

2.1 Data

Data merupakan fakta atau potongan informasi dan dua hal tersebut merupakan hal yang berbeda. Data pun sering kali memerlukan suatu konteks dalam membuat pengetahuan [5]. Data juga diartikan sebagai data mentah *(raw data)* yang merupakan kumpulan simbol, angka, dan teks tanpa arti. Oleh karena itu, data harus melalui tahap pemrosesan agar dapat memiliki makna [6]. Data dibagi menjadi dua buah berdasarkan tipe, yaitu data kualitatif dan data kuantitatif. Penjelasannya adalah sebagai berikut:

A. Data Kualitatif

Data kualitatif atau yang sering disebut sebagai data kategorial adalah data yang dapat ditempatkan ke dalam beberapa kategori berbeda. Terkadang, data kualitatif dapat disusun dalam suatu urutan yang memiliki makna. Namun pada data kualitatif, tidak terdapat operasi aritmetika yang dapat diterapkan [7]. Terdapat beberapa contoh untuk data kualitatif seperti golongan darah yang terdiri dari empat buah yaitu AB, A, B, dan O, dan evaluasi penugasan dengan empat buah indikator yaitu gagal, lulus, baik, dan sangat baik.

B. Data Kuantitatif

Data kuantitatif adalah data yang bersifat numerik. Pemberian peringkat secara berurutan serta penggunaan operasi aritmetika, merupakan salah satu ciri dari data kuantitatif. Pengklasifikasian data kuantitatif, dibagi menjadi dua buah kelompok, antara lain [7]:

1. Data Diskrit

Data diskrit adalah data dengan asumsi nilai yang tertera dapat dihitung. Namun, semua nilai pada data diskrit terletak pada garis bilangan, tidak dapat diasumsi. Contoh data diskrit adalah jumlah anak dalam sebuah keluarga.

2. Data Kontinu

Data kontinu dapat diasumsikan semua nilai dalam rentangnya terletak dalam suatu garis bilangan. Perolehan nilai didapatkan dengan cara mengukur. Contoh data kontinu adalah suhu.

C. Tingkat skala pengukuran pada data

Walaupun terdapat dua buah klasifikasi pada data seperti data kualitatif dan kuantitatif seperti yang dijelaskan di atas, terdapat tingkat skala pengukuran pada data. Tingkatan tersebut dibagi menjadi empat buah, yaitu nominal, ordinal, interval, dan rasio. Pemaparannya adalah sebagai berikut [7].

Tabel 2.1 Tingkat Skala Pengukuran pada Data

Level	Peringkat	Operasi Aritmetika	Nol Dalam Aritmetika
Nominal	Tidak	Tidak	Tidak
Ordinal	Ya	Tidak	Tidak
Interval	Ya	Ta	Tidak
Rasio	Ya	Ya	Ya

1. Nominal

Data dengan kategori yang tak tumpang tindih serta lengkap dengan tidak ada pemeringkatan, termasuk dalam tingkat nominal. Level nominal, tergolong ke dalam data kualitatif. Contohnya adalah golongan darah.

2. Ordinal

Pengklasifikasian data dalam kategori yang tak dapat dihitung dengan aritmetika, namun dapat diurutkan dalama kategori, termasuk data ordinal. Level ordinal, tergolong ke dalam data kualitatif. Contohnya adalah evaluasi penugasan.

3. Interval

Data yang dapat dihitung dengan aritmetika dan mengklasifikasikan tingkat pengukuran serta perbedaan antara satuan ukuran, termasuk dalam data interval. Salah satu ciri dari data interval adalah nilai nol yang memiliki arti tidak ada, dalam suatu interval. Level interval, tergolong ke dalam data kuantitatif. Contohnya adalah suhu dalam Fahrenheit.

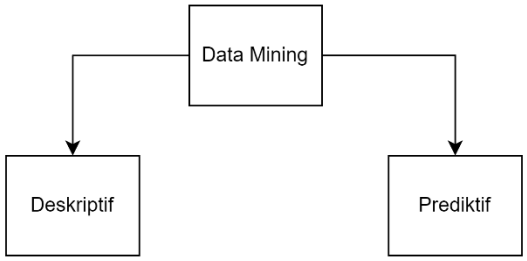
4. Rasio

Semua karakteristik yang terdapat pada level interval, termasuk pula dalam level rasio. Hal yang membedakan antara rasio dan interval adalah nilai nilai nol yang berarti. Hal tersebut mengakibatkan rasio berada di unit ukuran yang berbeda. Level rasio, tergolong ke dalam data kuantitatif. Contohnya adalah jumlah anak dalam satu keluarga.

2.2 *Data Mining*

Jika didefinisikan, *data mining* adalah suatu proses menemukan pola yang menarik. Pola tersebut sebelumnya tidak diketahui dan berpotensi berguna dari sejumlah besar data. Pola tersebut dapat ditemukan dari berbagai jenis seperti perubahan anomali, sub graf, asosiasi, dan struktur signifikan [8]. Keterkaitan antara *data mining* dan *knowledge discovery* adalah *data mining* merupakan langkah integral dalam melakukan proses *knowledge discovery*.

Beberapa model yang digunakan dalam teknik *data mining* adalah klasifikasi *(classification)*, klasterisasi *(clustering)*, *sequence and link analysis*, regresi *(regression)*, dan permodelan ketergantungan *(dependency modeling)* [9]. Berdasarkan tugasnya, *data mining* dibagi menjadi dua, yaitu deskriptif dan prediktif. Pemaparannya tertera pada gambar di bawah ini.



Gambar 2.1 Pembagian Task pada *Data Mining*

A. Deskriptif

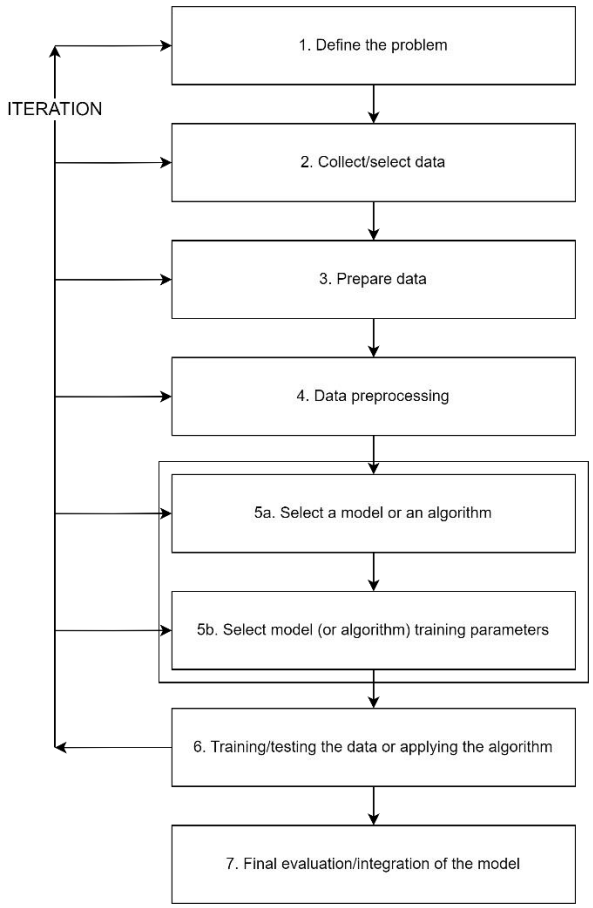
Data mining deskriptif digunakan untuk menemukan data yang menggambarkan pola. *Data mining* deskriptif pun akan menghasilkan informasi baru yang signifikan dari kumpulan data yang telah ada [10].

B. Prediktif

Data mining prediktif digunakan untuk memprediksi nilai yang tidak diketahui dari suatu kumpulan data yang lain. Contohnya adalah ketika dokter yang mencoba mendiagnosis penyakit berdasarkan hasil tes medis pasien [10].

2.3 Metode *Data Mining Life Cycle*

Data Mining Life Cycle merupakan sebuah metode yang digunakan untuk mendapatkan sebuah *knowledge discovery*. Berdasarkan proses prediktif yang terdapat dalam *data mining*, disebutkan bahwa terdapat proses *data mining*, khususnya *Data Mining Life Cycle* (siklus hidup *data mining*) [11][12], [13]. Mengenai siklus hidup *data mining*, dapat dilihat di gambar di bawah ini.



Gambar 2.2 Proses Permodelan Data dan *Data Mining Life Cycle*

Terdapat tujuh buah proses yang tertera pada **Gambar 2.2**. Penjelasan mengenai tahapan dalam siklus hidup *data mining* adalah sebagai berikut.

1. *Define the problem* (Mendefinisikan masalah)

Define the problem atau mendefinisikan permasalahan memiliki arti di mana masalah yang terjadi, didefinisikan secara akurat dengan menentukan kelayakan dari penggunaan *data mining*. Pada tahap pertama, terbagi menjadi tiga buah tahap. Pertama yaitu mengidentifikasi tujuan agar upaya melakukan *data mining* dapat menghasilkan hasil yang terstruktur. Kedua yaitu mendefinisikan tujuan atau sasaran dari *data mining.* Terakhir adalah analisa masalah dengan tujuan untuk melihat apakah permasalahan yang terjadi dapat ditangani dengan menggunakan *data mining*. Akan dilakukan identifikasi pada beberapa hal seperti ketersediaan serta teknologi yang digunakan dalam *data mining*.

2. *Collect and select data* (Mengumpulkan dan memilih data)

Pada bagian ini, ditentukan data yang akan dikumpulkan, jumlah data yang akan dikumpulkan, serta bagaimana cara mengumpulkannya. Selain itu, dilakukan pula seleksi atribut yang diperlukan untuk *data mining.*

3. *Prepare data* (Mempersiapkan data)

Proses mempersiapkan data atau yang dapat disebut juga sebagai tahap *data preprocessing* (Pra-pemrosesan data), dilakukan dengan tujuan untuk meningkatkan kualitas data. Tahapan ini merupakan tahapan yang paling penting karena sekitar 50%-80% proses *data mining* dilakukan di tahapan ini. Terdapat beberapa faktor yang mempengaruhi kualitas data pada tahap pemrosesan data, seperti keakuratan, konsistensi, keutuhan, aktualitas, serta penafsiran [14]. Oleh karena itu akan dilakukankan proses verifikasi kualitas data, pembersihan data *(data cleaning)* yang terdiri dari pembersihan data kosong, data duplikat, dan pengurangan baris atau kolom. Proses pengabungan dan transformasi data pun dilakukan pada tahapan ini. Pada tahap ini pula, kelak akan dilakukan tahapan yang termasuk ke dalam *predictive data mining*, seperti set pelatihan yang digunakan untuk membangun model awal, set pengujian yang digunakan untuk menyesuaikan model awal agar lebih umum, serta evaluasi set untuk melihat keefektifan model.

4. *Select an appropriate mining method* (Memilih metode *data mining* yang tepat)

Algoritma yang akan digunakan, harus ditentukan untuk memasuki tahapan *data mining*. Pada bagian ini pun, dilakukan seleksi model atau parameter untuk dilakukan *training*, seperti jumlah *node* pada setiap *level* ketika algoritma tersebut digunakan.

5. *Training/testing* (Melatih dan menguji)

Menerapkan algoritma pada data untuk dilatih.

6. *Final integration and evaluation of the generated model* (Integrasi akhir dan evaluasi dari model yang dihasilkan)

Setelah dilakukan tahap *training* dan *testing*, dilakukan evaluasi apakah model tersebut bekerja dengan baik atau tidak. Jika belum sesuai target, maka akan kembali ke tahap pertama, yaitu *define the problem* (mendefinisikan masalah).

2.4 Bahan Makanan

Bahan makanan atau yang dapat disebut juga sebagai bahan pangan, merupakan bahan yang dihasilkan dari perkebunan, perternakan, pertanian, serta teknologi makanan. Selain itu, bahan makanan merupakan bahan makanan yang belum tercampur dengan pengawet, segar, serta mengandung semua unsur gizi [15]. Pada umumnya, dibagi menjadi dua komponen besar, yaitu komponen makro yang merupakan komponen utama, serta komponen mikro. Komponen makro terdiri dari air, karbohidrat, lemak dan turunannya, serta protein. Sementara itu, komponen mikro terdiri dari mineral, vitamin, serta pigmen. Setiap komponen tersebut memiliki variasi jenis dan jumlahnya, sehingga akan membentuk struktur, rasa, tekstur, aroma, warna, serta kandungan gizi yang berlainan [16].

Bahan makanan juga dapat dibagi menjadi dua golongan yaitu *perishable* dan *groceries* [17]. Untuk bahan makanan dengan golongan *perishable*, bahan makan tersebut cenderung mudah rusak, sehingga memerlukan suatu perlakuan khusus untuk menjaga kualitasnya. Berbeda dengan golongan *groceries*. Walaupun bahan makanan tersebut tidak mudah rusak, namun proses penyimpanannya perlu diperrhatikan pula agar tidak cepat berjamur.

2.5 *Text Mining*

Text mining merupakan sebuah teknik yang digunakan untuk menangani *klasifikasi, clustering, information extraction,* dan *information retrival* [18] . Dalam *text mining,* terdapat beberapa tahapan, diantaranya [19].

1. *Parsing*

Tahap pertama adalah *parsing* yang merupakan pemecahan dokumen dalam suatu atribut, menjadi komponen-komponen terpisah.

2. *Lexical Analysis*

Dalam tahapan *lexical analysis,* terdapat *case folding* dan penghilangan tag HTML, angka, dan karakter spesial.

1) *Case Folding*

Biasanya terdapat huruf kapital dan singkatan dalam huruf kapital. Tujuan dilakukan *case folding* adalah agar pada kata yang sama tidak dihitung sebagai kata yang berbeda oleh komputer.

2) Pembersihan tag HTML, angka, dan karakter spesial

Pada saat dilakukan *scraping* dari website internet, terdapat beberapa komponen selain huruf yang terbawa. Komponen-komponen tersebut seperti tag HTML, angka, dan karakter spesial. Tujuan dilakukan pembersihan adalah agar komponen selain huruf yang terbawa pada saat *scraping*, tidak diikutsertakan pada saat perhitungan oleh komputer.

3. *Tokenizing*

Tokenizing atau yang biasa disebut dengan sebutan tokenisasi merupakan tahapan pada *text mining* yang melakukan pemisahan kalimat menjadi kata, karakter, serta tanda baca, menjadi sebuah token.

4. *Phrase Detection*

Phrase Detection merupakan tahapan untuk mendeteksi dua kata atau lebih yang tergolong dalam satu frase. Contohnya adalah kata *'search engines'* yang merupakan suatu kesatuan, dikarenakan memiliki arti mesin pencarian. Agar kata tersebut tetap terhitung sebaga satu frase, maka dilakukan penambahan *dash* (-) jika terdeteksi dua kata atau lebih.

5. *Term Frequency Inverse Document Frequency* (TF-IDF)

Salah satu metode pembobotan kata yang populer dengan hasil efisien, mudah dan akurat adalah TF-IDF [20]. TF-IDF merupakan salah satu cara yang dapat digunakan dalam memberikan bobot terhadap setiap kata yang terdapat pada teks [21]. Metode ini merupakan metode yang digunakan untuk menghitung bobot setiap kata *(term)* yang paling umum digunakan pada *information retrieval*. Metode ini menggabungkan dua konsep perhitungan bobot. Konsep tersebut adalah frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu serta *inverse* frekuensi dokumen yang mengandung kata tersebut.

Frekuensi kemunculan kata yang terdapat pada dokumen yang diberikan menunjukkan seberapa penting kata yang dimaksud dalam dokumen tersebut.

Sementara frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Berdasarkan pemaparan tersebut, maka bobot hubungan antara sebuah kata dan sebuah dokumen dikatakan tinggi apabila frekuensi kata tersebut tinggi dalam dokumen. Serta frekuensi keseluruhan dokumen yang mengandung kata yang dimaksud rendah pada kumpulan dokumen.

Berikut adalah rumus yang digunakan untuk menghitung *Term Frequency Inverse Document Frequency* (TF-IDF) [22].

1) Binary Term-Weighting

$$w_{t,d} = \begin{cases} 1, jika\ d\ mengandung\ t \\ 0, jika\ d\ tidak\ mengandung\ t \end{cases} \quad (2.1)$$

$$w = bobot$$

$$t = term$$

$$d = dokumen$$

2) Raw-Term Frequency (TF)

$$w_{t,d} = tf_{t,d} \quad (2.2)$$

$$w = bobot$$

$$t = term$$

$$d = dokumen$$

$$tf_{t,d} = jumlah\ kemunculan\ (frekuensi)\ term\ t\ pada\ dokumen\ d$$

3) Log-Frequency Weighting

$$w_{t,d} = \begin{cases} 1 + log_{10} tf_{t,d}, jika\ tf_{t,d} > 0 \\ 0, jika\ tf_{t,d} = 0 \end{cases} \quad (2.3)$$

$$w = bobot$$

$$t = term$$

$$d = dokumen$$

$$tf_{t,d} = jumlah\ kemunculan\ (frekuensi)\ term\ t\ pada\ dokumen\ d$$

4) Inverse Document-Frequency

$$IDF_t = log_{10}\left(\frac{N}{DF_t}\right) + 1 \quad (2.4)$$

$$df_t = jumlah\ dokumen\ yang\ mengandung\ term\ t$$

$$df_t\ merupakan\ ukuran\ kebalikan\ dari\ keinformatifan\ term\ t$$

$df_t \leq N$ = *Nilai* df_t *lebih kecil atau sama dengan jumlah dokumen N*

5) TF-IDF Weighting

$$TF - IDF = tf_{t,d} * idf_t \quad (2.5)$$

$$TF - IDF = tf_{t,d} * log_{10}\left(\frac{N}{DF_t}\right) \quad (2.6)$$

$$\text{Jika } IDF_t = 0 \text{ maka } IDF_t = log(N) \text{ dan TF-IDF} = TF_i * log(N) \quad (2.7)$$

2.6 Metode Klasifikasi

Klasifikasi adalah suatu proses untuk mendapatkan sejumlah model atau fungsi yang dapat menggambarkan, mengenali, serta membedakan kelas dari suatu data, Tujuan dari klasifikasi adalah untuk dengan menggunakan model yang diperoleh, dapat dilakukan prediksi kelas objek yang label kelasnya belum diketahui. Label dari suatu kelas harus berbentuk atribut diskret. Hal tersebut merupakan kunci yang membedakan antara regresi dengan klasifikasi.

Proses klasifikasi menurut Gorunescu terbagi menjadi empat buah komponen dasar, yaitu [23]:

A. *Class* / Kelas atau label kelas

Merupakan variabel *dependen* dari model yang merupakan variabel kategori. Variabel kategori tersebut menjelaskan sebuah ‘label’ pada objek setelah dilakukan proses klasifikasi. Contohnya adalah skala loyalitas pelanggan dengan label loyal dan tidak loyal.

B. *Predictor* / Prediktor atau *Atribute* / Atribut

Merupakan variabel *independent* dari suatu model. Variabel tersebut diwakili oleh karakteristik (atribut) dari data yang akan diklasifikasikan serta berdasarkan klasifikasi yang telah dibuat. Contohnya adalah tekanan darah dan kecepatan angin.

C. *Training Data set* / *Dataset* Latihan

Merupakan suatu kumpulan data yang berisik *record* (baris) untuk dua komponen sebelumnya yaitu kelas dan atribut. Dapat juga berupa variable kategoris atau kontinu. *Training set* digunakan untuk ‘pelatihan’ atau pembangunan model

dengan tujuan untuk menyesuaikan kelasnya berdasarkan prediktor yang tersedia. Contohnya adalah kelompok pasien yang diidentifikasi pada kasus serangan jantung.

D. *Testing Dataset* / *Data set* pengujian

Testing dataset berisikan data baru yang akan dilakukan klasifikasi dengan metode klasifikasi. *Dataset* ini akan digunakan pula untuk mengukur tingkat akurasi dari klasifikasi. Tujuannya adalah agar performa dari metode tersebut dapat dilakukan evaluasi.

2.7 *k-Nearest Neighbor* (k-NN)

Metode *k-Nearest Neighbor* atau yang biasa dikenal dengan sebutan k-NN adalah metode yang bekerja dengan cara mencari sejumlah *k* pada suatu objek data atau pola dari semua pola latih yang ada. Nilai *k* tersebut merupakan nilai yang paling dekat dengan pola masukan. Parameter *k* merupakan suatu parameter yang tak dapat ditentukan dengan rumus matematika. Pemilihan kelas akan terjadi berdasarkan jumlah pola terbanyak di antara *k* pola tersebut. Dalam menentukan *k* pola terdekat, harus dilakukan berdasarkan ukuran *similarity* atau *dissimilarity* dan jarak. Hasil akhir dari proses k-NN akan memberikan akurasi yang paling tinggi dalam mengeneralisir data-data yang akan datang.

Metode k-NN merupakan metode yang menarik. Hal tersebut dikarenakan keputusan kelas dapat ditelusuri dengan mudah, sehingga dapat digunakan untuk memperbaharui model klasifikasi. Terakhir, metode ini bekerja secara lokal dengan memperhitungkan sejumlah *k* data, sehingga sesuai untuk himpunan data yang terkelompok secara lokal.

Walaupun metode ini sangat menarik karena beberapa kelebihan yang dimilikinya, terdapat pula kelemahan pada metode ini. Apabila nilai *k* diatur dengan parameter yang kecil, maka tingkat generalisasinya akan tinggi. Jika sebaliknya, maka kemungkinan data menjadi *overfit* di mana, data baru akan gagal diklasifikasikan [24]. Oleh karena itu parameter *k* harus diatur dengan tepat pada saat proses *training*.

Metode.k-NN biasanya dihitung dengan metode *distance measure*. Terdapat enam buah metode, yaitu *Minkowski*, *Manhattan*, *Euclidean*, *Tanimoto*, *Jaccard*, dan *Mahalanobis* jika berupa data kontinu, dan *Hamming* jika data kategori [23].

A. *Minkowski Distance (MD)*

Merupakan bentuk umum dari *Manhattan* dan *Euclidean Distance*. Apabila diketahui titik koordinat kartesian antara himpunan dua titik $x = (x_1, x_2, \dots, x_n)$ dan $y = (y_1, y_2, \dots, y_n)$, maka:

$$D(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \quad (2.8)$$

B. *Manhattan Distance*

Setara dengan *Minkowski* dengan P = 1. Titik p dan q dalam n dimensional *real* vektor *space* merupakan total jumlah proyeksi dari segmen setiap titik ke *axis* koordinatnya. Apabila titik $x = (x_1, x_2, \dots, x_n)$ dan $y = (y_1, y_2, \dots, y_n)$ merupakan dua titik dalam *Euclidean u-space* maka:

$$D(x,y) = ||x - y||i = \sum_{i=1}^{n} |xi - yi| \quad (2.9)$$

Jika dimensi yang diambil berasal dari ruang dimensi dua, maka jarak *Manhattan* p_1, p_2 dan q_1, q_2 diperoleh dengan cara:

$$|x_1 - y_1| + |x_1 - y_1|$$

C. *Euclidean Distance*

Merupakan panjang segmen $\underset{pq}{\rightarrow}$ antara titik-titik yang terdapat dalam suatu himpunan p dan himpuan q. Apabila diketahui titik koordinat kartesian antara himpunan dua titik $x = (x_1, x_2, \dots, x_n)$ dan $y = (y_1, y_2, \dots, y_n)$, maka jarak antara p ke q dalam *Euclidean u-space* adalah seperti berikut:

$$D(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (2.10)$$

$$= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2.8 Usability Testing

Usability Testing merupakan sebuah bagian dari kelompok keilmuan HCI atau *Human Computer Interactions* [25]. *Usability Testing* memfokuskan pada desain antarmuka, serta interaksi antara manusia dengan komputer. Pada *usability testing*, terdapat beberapa indikator yang digunakan untuk mengukur kepuasan pengguna dalam menggunakan aplikasi. Kepuasan tersebut memiliki arti apakan tujuan yang diinginkan pengguna tercapai atau tidak.

Selain itu digunakan untuk mengetahui seberapa efektif dan efisien sebuah aplikasi dapat membantu pengguna mencapai tujuannya. Pada *usability testing*, terdapat 5 buah tahapan yang menjadi penentu usabilitas, antara lain [26]

1. *Learnability:* seberapa mudah pengguna menyelesaikan tugas-tugas dasar pada aplikasi.
2. *Efficiency:* seberapa cepat pengguna dapat melakukan tugas yang diberikan.
3. *Memorability:* apakah pengguna dapat mengingat kembali tampilan serta alur pada aplikasi tersebut.
4. *Error:* seberapa banyak serta berat kesalahan-kesalahan yang dilakukan pengguna. Selain itu berkaitan dengan seberapa mudah pengguna mengatasi permasalahan tersebut pada aplikasi.
5. *Satisfaction:* kepuasan pengguna setelah menggunakan aplikasi tersebut.

Usability testing akan menggunakan kuesioner yang disebarkan pada responden. Perhitungannya menggunakan *System Usability Scale (SUS)* dengan syarat hasil skor pada pertanyaan bernomor ganjil dikurangi satu, 5 dikurangi hasil skor pada pertanyaan genap. Hasil skor untuk setiap pertanyaan dijumlahkan lalu dikalikan 2,5 per respondennya. Hasil skor SUS diperoleh dengan cara membagi total skor responden dengan jumlah responden [27].

Tabel 2.2 Skala Interpretasi Hasil Perhitungan dengan SUS

Grade	SUS	Percentile Range	Adjective	Acceptable	NPS
A+	84,1 – 100	96 - 100	*Best Imaginable*	*Acceptable*	*Promoter*
A	80,8 – 84,0	90 – 95	*Excellent*		
A-	78,9 – 80,7	85 – 89	*Good*		
B+	77,2 – 78,8	80 – 84			*Passive*
B	74,1 – 77,1	70 – 79			
B-	72,6 – 74,0	65 – 69			
C+	71,1 – 72,5	60 – 64			
C	65,0 – 71,0	41 – 59	*OK*	*Marginal*	
C-	62,7 – 64,9	35 – 40			
D	51,7 – 62,6	15 – 34			*Detractor*