BAB II
KAJIAN PUSTAKA DAN DASAR TEORI

Bab ini berisi tentang kajian pustaka dan pembahasan tentang teori dasar yang berkaitan dengan implementasi Modified K-Nearest Neighbour dalam menentukan tingkat resiko penyakit lemak dalam darah *(Profil Lipid).* Kajian pustaka membahas tentang penelitian yang telah ada dan yang akan diusulkan. Dasar teori membahas teori yang diperlukan untuk menyusun penelitian yang diusulkan, meliputi konsep dasar dari penyakit lemak darah, data mining, Algoritma K-Nearest Neighbour, dan Algoritma Modified K-Nearest Neighbour.

2.1. Kajian Pustaka

Kajian pustaka pada penelitian ini akan membahas tentang penelitian sebelumnya yang dapat mendukung pengerjaan penelitian ini. Pada studi kasus pengklasifikasian tingkat resiko penyakit lemak darah masih belum diterapkan dalam suatu metode. Akan tetapi, berbagai riset dengan menggunakan Algoritma *Modified K-Nearest Neighbor* (MKNN) telah diimplementasikan pada beberapa macam kasus data. Hal tersebutlah yang akan dibahas pada kajian pustaka ini, diantaranya :

1. Penelitian [ADE-13] membahas tentang pengimplementasian metode MKNN dalam menentukan potensi Tsunami akibat gempa bumi yang terjadi, dengan menggunakan 3 parameter gempa bumi yakni : kedalaman (km), magnitude (SR) dan letak *epicenter* gempa bumi. Pengujian sistem tersebut dilakukan dengan nilai k yang berbeda, yaitu k = 1 hingga k = 20. Setiap nilai k dilakukan uji coba sebanyak 5 kali dan didapatkan akurasi rata-rata. Rata-rata nilai akurasi maksimum yang dihasilkan sistem sebesar 73,74% pada saat jumlah dataset 100 dan rata-rata akurasi minimum sistem sebesar 69,63% pada saat jumlah dataset 300.
2. Penelitian lain oleh [ZAI-13] membahas tentang implementasi metode *Modified K-Nearest Neighbour* untuk klasifikasi penyakit tanaman kedelaidengan menggunakan data morfologi tanaman kedelai sebagai variabel. Pada penelitian

tersebut dibuat sebuah aplikasi pengklasifikasi berbasis desktop dengan tujuan untuk memudahkan dalam mengkasifikasikan penyakit tanaman kedelai. Terdapat 34 variabel pada data morfologi tanaman kedelai untuk setiap jenis penyakit. Pengujian dilakukan dengan mengubah nilai k. Tingkat akurasi tertinggi dari sistem klasifikasi penyakit tanaman kedelai dengan menggunakan algoritma *Modified K-Nearest Neighbour* menggunakan 300 data latih adalah sebesar 92.74%, dengan nilai k=3 .

3. Pada penelitian yang dilakukan ini bejudul "Implementasi Algoritma *Modified K-Nearest Neighbor* (M-KNN) Untuk Menentukan Tingkat Resiko Penyakit Lemak Darah". Penelitian ini bertujuan untuk mengetahui tingkatan resiko penyakit lemak darah berdasarkan hasil uji tes darah.

2.2 Lemak Darah *(Profil Lipid)*

Lemak dalam darah adalah komponen lemak yang terdapat pada pembuluh darah yang berfungsi sebagai sumber energi, membentuk dinding sel-sel dalam tubuh, dan sebagai bahan dasar pembentukan hormon -hormon steroid. Lemak dalam darah memang dibutuhkan oleh tubuh, akan tetapi dapat membentuk endapan pada dinding pembuluh darah. Oleh karena itu, HDL bertugas mengambil kolesterol jahat serta fosolipida dari darah dan menyerahkan pada lipoprotein lain, untuk diangkut kembali ke hati [SPI-12].

Di hati, reseptor LDL mengatur lemak dalam darah. Jika LDL meningkat, sel-sel perusak menumpuk di dinding pembuluh darah dan membentuk plak, yang memperkecil diameter pembuluh darah. Plak yang bercampur dengan protein akan ditutupi oleh sel-sel otot dan kalsium dan dalam jangka waktu bertahun-tahun bisa terjadi *atherosclerosis* (pengerasan dan penyempitan pembuluh darah). Akibatnya, suplai oksigen dan nutrisi ke seluruh tubuh terhambat. Jika dibiarkan, dapat mengakibatkan gangguan jantung, stroke, dan gangguan lain [SPI-12].

Lemak dalam darah manusia terbagi menjadi 2 jenis yakni kolesterol LDL (kolesterol jahat) dan HDL (kolesterol baik). LDL apabila terlalu tinggi dan tidak seimbang dengan kolesterol baik HDL dapat menyebabkan penempelan di dinding

pembuluh darah. Lemak dalam darah yang berlebihan bisa menempel di dinding pembuluh darah sehingga pembuluh darah menyempit dan aliran darah tidak lancar. Hal inilah yang menjadi salah satu faktor resiko penyakit jantung [SPI-12].

2.2.1. Faktor Resiko Penyakit Lemak Darah

Pendeteksian penyakit lemak darah terdapat beberapa faktor yang mempengaruhi tingkatan resiko penyebab penyakit lemak darah, seperti : LDL, HDL, *trigliserida* dan kadar kolesterol total. Dalam hal ini akan dibahas secara detail mengenai faktor-faktor resiko penyakit lemak darah yang menjadi batasan permasalahan pada penelitian ini. Faktor-faktor resiko penyakit lemak darah pada penelitian ini adalah :

1. *Low Density Lipoprotein* (LDL)

   LDL (*Low Density Lipoprotein*) merupakan jenis lipoprotein yang mengangkut kolesterol dan trigliserida dari hati ke jaringan perifer. Kadar LDL dinyatakan dalam satuan mg/dl. Klasifikasi kadar LDL ditunjukkan pada tabel 2.1 [TJO-01] :

Tabel 2.1 Klasifikasi Kadar LDL

Klasifikasi	Kadar Kolesterol
Normal	< 100
Mendekati Normal	100 - 129
Borderline High	139 - 159
Tinggi	160 – 189
Sangat Tinggi	≥ 190

2. *High Density Lipoprotein* (HDL)

   HDL (*High Density Lipoprotein*) merupakan jenis kolesterol yang besifat baik atau menguntungkan. Hal ini dikarenakan HDL mengangkut kolesterol dari pembuluh darah kembali ke hati untuk dibuang, sehingga dapat mencegah penebalan dinding pembuluh darah atau mencegah terjadinya proses *aterosklerosis.* Klasifikasi kadar HDL di tunjukan pada tabel 2.3 [TJO].

Tabel 2.2 Klasifikasi Kadar HDL

Klasifikasi	Kadar Kolesterol
Rendah	<40
Tinggi	≥ 40

3. *Trigliserida*

Trigliserida merupakan komponen yang normal dari darah, baik yang datang dari diet maupun yang dihasilkan oleh tubuh. Sebagian besar lemak yang dimakan berbentuk *trigliserida*. Lemak yang berasal dari buah-buahan seperti kelapa, durian, dan alpukat tidak mengandung kolesterol, melainkan mengandung kadar *trigliserida* yang tinggi. Klasifikasi kadar *trigliserida* ditunjukkan pada tabel 2.3 [TJO-01].

Tabel 2.3 Klasifikasi Kadar *Trigliserida*

Klasifikasi	Kadar Kolesterol
Normal	< 150
Borderline High	150 – 199
Tinggi	200 – 499
Sangat Tinggi	≥ 500

4. Kolesterol Total

Kolesterol total merupakn susunan dari banyak zat, yang di dalamnuya terdiri dari trigliserida, LDL kolesterol dan HDL kolesterol. Kolesterol yang ada di dalam zat makanan akan meningkatkan kadar kolesterol dalam darah. Peningkatan kadar kolesterol total berbanding lurus dengan peningkatan resiko penyakit lemak darah. Semakin tinggi kolesterol total, resiko penyakit lemak darah semakin tinggi juga. Klasifikasi kadar kolesterol total ditunjukan pada tabel 2.2 [TJO-01].

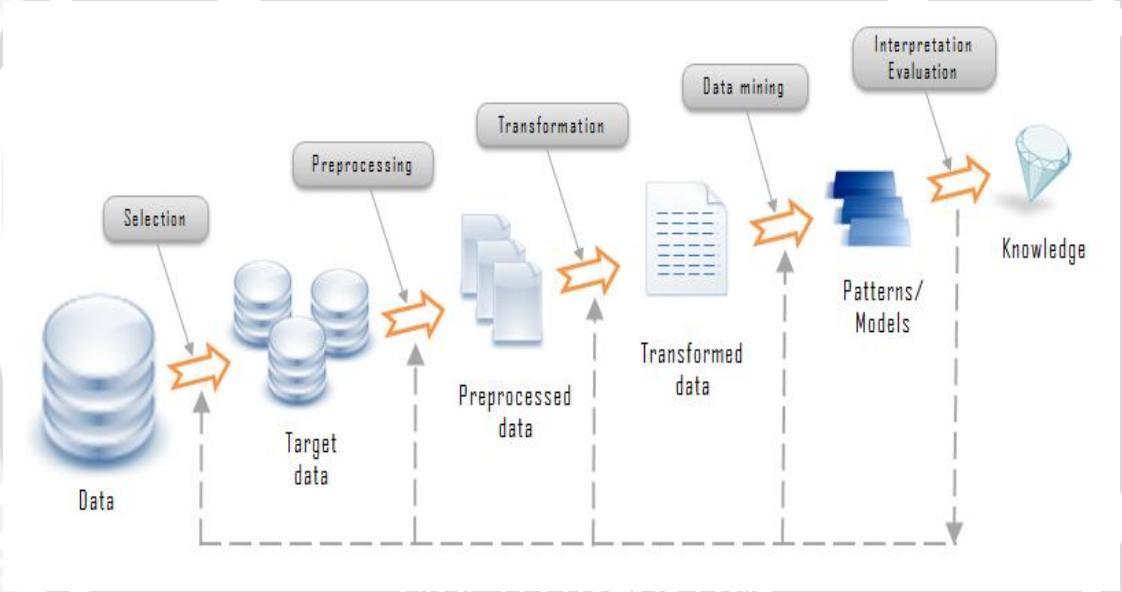
Tabel 2.4 Klasifikasi Kadar Kolesterol Total

Klasifikasi	Kadar Kolesterol
Normal	< 200
Borderline (Waspada)	200 – 239
Tinggi	≥ 240

2.3. *Data Mining*

2.3.1 Definisi *Data Mining*

Data Mining adalah proses pencarian pola dan relasi yang tersembunyi dalam sejumlah data yang besar. *Data mining* dapat juga dikatakan sebagai suatu proses untuk menggali informasi (knowledge) dari sejumlah data yang besar juga. Tujuan dari *data mining* yaitu untuk melakukan klasifikasi, estimasi, prediksi, *association rule*, *clustering*, deskripsi dan visualisasi. *Data mining* juga dapat dikatakan sebagai suatu proses untuk menggali informasi (*knowledge*) dari sejumlah data yang besar. Data mining merupakan bagian dari proses *Knowledge Discovery in Database* (KDD) [HAN-06]. Proses KDD dapat dilihat pada gambar 2.1.



Gambar 2.1 Proses *Knowledge Discovery in Database* (KDD) [KDD-13]

Tahapan-tahapan dari proses *Knowledge Discovery in Database* (KDD) sebagai berikut [KUS-09] :

1. Data *Selection*

   Data *selection* atau pemilihan data merupakan sekumpulan data operasional yang perlu dilakukan sebelum tahap penggalian informasi. Data hasil seleksi yang digunakan untuk proses *data mining* kemudian disimpan dalam suatu berkas yang terpisah dari basis data operasional.

2. *Preprocessing*

   Preprocessing merupakan operasi dasar seperti penghapusan *noise*, sehingga pada proses tersebut terjadinya proses pembuangan duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan pada data. Pada proses ini dapat dilakukan proses *enrichment* data untuk memperkaya dan menambah data dengan informasi lain yang relevan dan diperlukan.

3. *Transformation*

   Merupakan suatu proses mencari fitur-fitur, yang berguna untuk mempresentasikan data sehingga sesuai untuk proses *data mining*. Proses ini merupakan proses kreatif dan bergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data Mining*

   Proses *data mining* adalah proses mencari pola atau informasi dalam data terpilih dengan menggunakan teknik atau metode tertentu. Pemilihan metode yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Evaluation*

   Tahap *evaluation* ini merupakan bagian dari proses KDD yang mencakup pemeriksaan kesesuaian pola atau informasi yang dihasilkan dengan fakta atau hipotesa yang ada sebelumnya. Model *data mining* dibuat berdasarkan salah satu dari dua jenis pembelajaran *supervised* dan *unsupervised*. Fungsi pembelajaran *supervised* digunakan untuk memprediksi suatu nilai. Sedangkan fungsi pembelajaran *unsupervised* digunakan untuk mencari struktur intrinsik atau relasi dalam suatu data yang tidak memerlukan *class* atau label sebelum proses pembelajaran.

Berdasarkan tugas yang dapat dilakukan, *data mining* dibagi menjadi beberapa kelompok, yaitu [KUS-09]:

a. Klasifikasi

   Dalam klasifikasi terdapat targe variabel kategori.

b. Pengklusteran

Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.

c. Asosiasi

Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu.

2.3.2. Metode Klasifikasi *Data Mining*

Model *data mining* akan memeriksa sejumlah data yang besar, setiap data berisi informasi tentang variabel target atau sasaran dan satu set input atau variabel prediktor. Dalam penentuan variabel prediktor tersebut metode klasifikasi sangat sesuai sebagai teknik penggalian informasi pada data. Klasifikasi dapat digunakan untuk menemukan model atau fungsi yang membedakan kelas data. Klasifikasi bertujuan untuk memprediksi kelas dari suatu objek yang labelnya belum diketahui.

Klasifikasi merupakan suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan-aturan tersebut digunakan pada data-data baru untuk diklasifikasi. Teknik ini menggunakan *supervised induction*, yang memanfaatkan kumpulan pengujian dari record yang terklasifikasi untuk menentukan kelas-kelas tambahan [KUS-07].

Klasifikasi adalah proses untuk menyatakan suatu objek ke dalam salah satu kategori yang sudah didefinisikan sebelumnya. Terdapat beberapa tahap klasifikasi, antara lain:

1. Pembangunan model

   Dalam tahap ini, dibuat suatu model untuk menyelesaikan masalah klasifikasi data. Model ini dibangun berdasarkan *training set.*

2. Penerapan model

   Model yang telah dibangun pada tahap sebelumnya, digunakan untuk menentukan *attribute* atau *class* dari sebuah data baru yang belum diketahui *attribute*-nya.

3. Evaluasi

   Hasil dari tahap sebelumnya dievaluasi mengunakan parameter terukur untuk menentukan penerimaan model klasifikasi data yang telah dibuat. Terdapat beberapa metode klasifikasi, antara lain *decision tree, bayesian, fuzzy, neural network, support vector machine (SVM)* dan *k-nearest neighbor* (KNN).

2.4 *K-Nearest Neighbor* (KNN)

2.4.1 Definisi *K-Nearest Neighbor* (KNN)

K – Nearest Neighbor (KNN) merupakan metode yang biasa digunakan pada klasifikasi data. Algoritma KNN adalah sebuah metode untuk mengklasifikasikan objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. *K-Nearest Neighbor* adalah suatu metode yang menggunakan algoritma *supervised* dengan hasil dari *query instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini ialah mengklasifikasikan objek baru berdasarkan atribut dan *training sample* [AKT-12].

Metode KNN merupakan metode yang sederhana, mudah diimplementasikan, dan efektif jika data training besar. Selain itu, metode KNN juga memiliki beberapa kelemahan seperti berikut [PAR-10]:

a. Biaya komputasi yang tinggi karena perlu menghitung jarak setiap data training.
b. Perlu menentukan nilai k parameter, jumlah tetangga terdekat.
c. Menggunakan perhitungan jarak yang belum diketahui pasti fungsi jarak yang digunakan.

2.4.2. Proses *K-Nearest Neighbor* (KNN)

Prinsip kerja algoritma *K-Nearest Neighbor* (KNN) yaitu mencari jarak terdekat antara data yang dievaluasi dengan k tetangga terdekat dalam *data training*. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan variabel [HAN-06].

2.4.2.1. *Interval Scaled Variable*

Interval Scaled Variable merupakan ukuran-ukuran kontinu dari skala linier. Ukuran tersebut berupa ukuran jarak, yang umum digunakan adalah jarak *euclidean*. Perhitungan untuk menghitung jarak (*Euclidean*) dengan persamaan [HAN-06] :

$$d\,(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(X_{1i} - X_{2i})^2} \quad \text{..... (2-1)}$$

Dimana X_1 merupakan nilai variabel data *testing* dan X_2 merupakan nilai variabel dari data *training*. Data kontinu dapat digunakan rumusan normalisasi atau standarisasi sebelum melakukan klasifikasi. Normalisasi bertujuan untuk mencegah atribut yang memiliki rentang terlalu besar dengan atribut yang bernilai kecil. Perhitungan min-max normalisasi dapat digunakan untuk mengubah atribut A dengan nilai v menjadi v' dalam *range* [0,1]. Perhitungan min-max normalisasi dengan persamaan [HAN-06] :

$$v' = \frac{v - \min_A}{\max_A - \min_A} \quad \text{.....(2-2)}$$

dimana :

v' = nilai normalisasi

v = nilai data yang akan dinormalisasi

$\min_A$ = nilai terendah (minimal) data pada atribut A

$\max_A$ = nilai tertinggi (maximal) data pada atribut A

2.4.2.2. Tahapan Proses *K-Nearest Neighbor* (KNN)

Tahapan dalam teknik K-NN dapat dilakukan dengan [JIA-06]:

1. Menentukan parameter k (jumlah tetangga paling dekat).
2. Menghitung jarak dengan menggunakan *Euclidean Distance* masing-masing obyek terhadap data sampel yang diberikan.

3. Mengurutkan obyek-obyek tersebut kedalam kelompok yang mempunyai jarak terkecil
4. Mengumpulkan kategori Y (klasifikasi *nearest neighbor*)

2.5. *Modified K-Nearest Neighbor* (MKNN)

2.5.1. Definisi *Modified K-Nearest Neighbor* (MKNN)

Tujuan utama yang menjadi dasar modifikasi pada metode KNN ini adalah menentukan kelas label dari *query instance* ke dalam k *data training* yang telah divalidasi. Setelah itu, weighted KNN akan dilakukan pada setiap data uji.

Modified K-Nearest Neighbor (MKNN) yaitu menempatkan label kelas data sesuai dengan k yang telah divalidasi dan sudah di tetapkan dengan perhitungan *K-Nearest Neighbor* (KNN).

2.5.2. Proses Modified K – Nearest Neighbor (MKNN)

Secara garis besar terdapat dua proses utama dalam metode MKNN, yaitu [PAR-10]:

1. **Validitas Data Training**

   Dalam metode MKNN setiap data training harus divalidasi terlebih dahulu. Validitas setiap data tergantung pada setiap tetangganya. Setelah dihitung validitas tiap data maka nilai validitas tersebut akan digunakan sebagai informasi lebih mengenai data tersebut.

   Validitas digunakan untuk menghitung jumlah titik dengan label yang sama untuk data tersebut. Persamaan yang digunakan untuk menghitung validitas dari setiap data adalah sebagai berikut.

$$Validitas\ (x) = \frac{1}{k} \sum_{i=1}^{k} S(label(x), (label(N_i(x))) \quad \ldots\ldots (2\text{-}3)$$

   dimana :

   k : jumlah titik terdekat

   Label (x) : kelas x

   Label $N_i(X)$: label kelas titik terdekat x

Fungsi S digunakan untuk menghitung kesamaan antara titik x dan data ke-i dari tetangga terdekat. Persamaan untuk mendefinisikan fungsi S dalam persamaan (2-4).

$$S(a,b) = \begin{cases} 1 & a=b \\ 0 & a \neq b \end{cases} \quad \text{..... (2-4)}$$

dimana :

a = kelas a pada data training

b = kelas lain selain a pada data training

melalui persamaan (2-4) ditunjukkan bahwa a dan b adalah label kelas kategori suatu data latih. S akan bernilai 1, jika label kategori a sama dengan label kategori b. S akan bernilai 0, jika label kategori a tidak sama dengan label kategori b.

2. *Weight Voting*

Weight voting KNN adalah salah satu variasi metode KNN yang menggunakan k tetangga terdekat, terlepas dari kelas data dan hasil perhitungan dari jarak dari masing-masing data. Dalam metode KNN, digunakan rumus *Weight* untuk masing-masing tetangga dengan persamaan sebagai berikut ini [PAR-10].

$$W(i) = \frac{1}{d+\alpha} \quad \text{..... (2-5)}$$

dimana d adalah jarak Euclidian dan α merupakan nilai *regulator smoothing*, dalam penelitian ini menggunakan α = 0,5. Weight *voting* ini kemudian dijumlahkan untuk setiap kelasnya, dan kelas dengan jumlah terbesar yang akan dipilih menjadi sebuah keputusan.

Dalam metode MKNN, masing-masing k tetangga terdekat dihitung dengan menggunakan persamaan (2-6). Kemudian, nilai validitas dari tiap data yang telah dihitung sebelumnya dikalikan dengan hasil *weight voting* berdasarkan jarak. Sehingga dalam metode MKNN, didapatkan persamaan *weight voting* sebagai berikut [PAR-10].

$$W(i) = Validitas(i) \times \frac{1}{d+\alpha} \quad \text{..... (2-6)}$$

Dimana :

W(i) : *Weight voting*

Validitas (i) : Nilai Validitas

d : Jarak Euclidean

Teknik *weight voting* ini berpengaruh terhadap data yang mempunyai nilai validitas lebih tinggi dan paling dekat dengan data. Selain itu, dengan mengalikan validitas dengan jarak data dapat mengatasi kelemahan dari setiap data yang mempunyai jarak dengan *weight* yang memiliki banyak masalah dalam *outlier*. Sehingga, metode MKNN secara signifikan akan lebih kuat daripada metode KNN yang hanya di dasarkan hanya pada jarak [PAR-10].

2.5.3. Tahapan Proses Modified K – Nearest Neighbor (MKNN)

Beberapa tahap dari algoritma *modified k-nearest neighbor* (MKNN), yaitu :

1. Menentukan nilai k tetangga terdekat
2. Menghitung validitas data latih
3. Menghitung jarak
4. Menghitung *weighted voting* (pembobotan)
5. Menentukan kelas dari data uji tersebut.

2.6. Akurasi Sistem

Akurasi adalah derajat kedekatan pengukuran terhadap nilai sebenarnya. Akurasi mencakup tidak hanya kesalahan acak, tetapi bisa juga yang disebabkan oleh kesalahan sistematik yang tidak terkoreksi [MUT-04].

Pada penelitian ini menggunakan tabel *Confusion* yang merupakan sebuah matriks yang berisi tentang informasi mengenai hasil klasifikasi oleh sistem dan hasil yang sebenarnya. Tabel *confusion* ditunjukkan pada tabel 2.5.

Tabel 2.5 Tabel Matriks *Confusion* 2x2

		Predicted Class	
		+	-
Actual class	+	*True Positive* (TP)	*False Positive* (FP)
	-	*False Negative* (FN)	*True Negative* (TN)

Tabel 2.6 Tabel Matriks *Confusion* 3x3

Actual/Predicted	A	B	C
A	AA	AB	AC
B	BA	BB	BC
C	CA	CB	CC

Keterangan :

1. True Positive (TP) menunjukkan bahwa data yang termasuk dalam hasil pengelompokkan oleh sistem memang merupakan anggota klasifikasi.
2. False Positive (FP) menunjukkan bahwa data yang termasuk dalam hasil pengelompokkan oleh sistem ternyata seharusnya bukan merupakan anggota klasifikasi.
3. False Negative (FN) menunjukkan bahwa data yang tidak termasuk dalam hasil pengelompokkan oleh sistem ternyata seharusnya merupakan anggota klasifikasi.
4. True Negative (TN) menunjukkan bahwa data yang tidak termasuk dalam hasil pengelompokkan oleh sistem ternyata seharusnya bukan merupakan anggota klasifikasi.
5. A,B,C merupakan permisalan sebuah contoh kelas.

Untuk menghitung nilai dari TP, FP, FN dan TN dengan menggunakan matriks confusion 3x3 pada persamaan 2.7 sampai 2.10 [KUS-08]:

TP = AA + BB + CC ……..…….……(2-7)

FP = (AB+AC) + (BA+BC) + (CA+CB) …… (2-8)

FN = (BA+CA) + (AB+CB) + (AC+BC) ……..……. (2-9)

TN = (BB+CC) + (AA+CC) + (AA+BB) ……... (2-10)

Apabila nilai dari TP, FP, FN dan TN didapatkan, maka dilakukan perhitungan akurasi sistem. Untuk menghitung akurasi sistem ditunjukkan pada persamaan 2.11 [KUS-08].

$$Akurasi\ Sistem = \frac{TP+TN}{TP+FP+FN+TN} x\ 100\% \text{.....(2-11)}$$