PROPOSAL TESIS

PENERAPAN METODE *SELF ORGANIZING MAP* UNTUK PEMBOBOTAN FITUR PADA ALGORITMA *NAÏVE BAYES*

Oleh:

KRISMAR WULANDA

P31.2015.01847

PROGRAM MAGISTER TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS DIAN NUSWANTORO

SEMARANG

2017

PROPOSAL TESIS

PENERAPAN METODE *SELF ORGANIZING MAP* UNTUK PEMBOBOTAN FITUR PADA ALGORITMA NAÏVE BAYES

KRISMAR WULANDA

P31.2015.01847

PROGRAM MAGISTER TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS DIAN NUSWANTORO

SEMARANG

2017

UNIVERSITAS DIAN NUSWANTORO

PERSETUJUAN PROPOSAL TESIS

JUDUL : PENERAPAN METODE *SELF ORGANIZING MAP* UNTUK PEMBOBOTAN FITUR PADA ALGORITMA NAÏVE BAYES

NAMA : Krismar Wulanda

NPM : P31.2015.01847

Proposal tesis ini telah diperiksa dan disetujui,

Semarang, September 2017

Romi Satria Wahono, Ph.D	Catur Supriyanto, M.CS
Pembimbing Utama	Pembimbing Pembantu

DAFTAR ISI

PROPOSAL TESIS ..... **i**
PERSETUJUAN PROPOSAL TESIS ..... **ii**
DAFTAR ISI ..... **iii**
DAFTAR GAMBAR ..... **v**
DAFTAR TABEL ..... **vi**
BAB 1 PENDAHULUAN **1**
1.1 Latar Belakang Masalah 1
1.2 Identifikasi Masalah 4
1.3 Rumusan Masalah 4
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 5
1.6 Sistematika Penulisan 5
BAB 2 TINJAUAN PUSTAKA **7**
2.1 Pendahuluan 7
2.2 Metode Review 7
2.2.1 Rumusan Masalah 10
2.2.2 Strategi Pencarian 11
2.2.3 Seleksi Pencarian 12
2.2.4 Ekstrasi Data 14
2.2.5 Penilaian Kualitas Penelitian dan Sintesis Data 15
2.3 Publikasi Jurnal Ilmiah 15
2.4 Peneliti yang Paling Aktif 16
2.5 Metode Pendekatan yang Digunakan 17
2.5.1 Metode Manipulasi Pembobotan Fitur 18
2.5.2 Metode Manipulasi Seleksi Fitur 18
2.5.3 Metode Manipulasi Data 18
2.5.4 Metode Manipulasi Struktur 19
2.6 Metode yang Pernah Diusulkan 19
2.6.1 Metode Wu, Pan, Zhu dan Cai [7]. 21

2.6.2 Metode Jiang, Cai, Zhang dan Wang [31] 23
2.6.3 Metode Taheri, Yearwood, Mammadov dan Seifollahi [6] 25
2.6.4 Metode Wu dan Cai [22] 27
2.6.5 Metode Jiang, Li, Wang dan Zhang [9] 29
2.6.6 Metode Zhang, Jiang, Li dan Kong [27] 31
2.7 Dataset yang Sering Digunakan 38
2.8 Daftar Referensi Systematic Literature Review 38
BAB 3 METODE PENELITIAN 41
3.1 Perancangan Penelitian 41
3.2 Analisis Masalah dan Tinjauan Pustaka 42
3.3 Pengumpulan Dataset 42
3.4 Metode yang Diusulkan 43
3.5 Eksperimen dan Pengujian Metode 47
3.6 Evaluasi dan Validasi Hasil 47
3.6.1 Evaluasi Hasil 48
3.6.2 Validasi Hasil 49
BAB 4 HASIL DAN PEMBAHASAN 52
BAB 5 KESIMPULAN 53
DAFTAR REFERENSI 54

DAFTAR GAMBAR

Gambar 2.1 Tahapan Systematic Literature Review 9
Gambar 2.2 Peta Pikiran Rumusan Masalah 11
Gambar 2.3 Langkah-Langkah Pencarian Literatur Penelitian 14
Gambar 2.4 Jumlah Publikasi Tahun 2010 - 2016 16
Gambar 2.5 Jumlah Publikasi Paper Asumsi Independensi Fitur 16
Gambar 2.6 Peneliti yang Aktif dan Jumlah Publikasi 17
Gambar 2.7 Distribusi Metode Untuk Mengatasi Asumsi Independensi Fitur. 18
Gambar 2.8 Peta Pikiran Penelitian Asumsi Independensi Fitur 20
Gambar 2.9 Metode Wu, Pan, Zhu dan Cai [7] 22
Gambar 2.10 Metode Jiang, Cai, Zhang dan Wang [21] 24
Gambar 2.11 Metode Taheri, Yearwood, Mammadov dan Seifollahi [6] 26
Gambar 2.12 Metode Wu dan Cai [22] 28
Gambar 2.13 Dataset yang Sering Digunakan 38
Gambar 3.1 Tahapan Penelitian, Aktivitas, dan Relasi Bab 41
Gambar 3.2 Metode Pembobotan Self Organizing Map (SOMWNB) 45
Gambar 3.3 Kerangka Pemikiran Penelitian 46
Gambar 3.4 10-Fold Cross Validation 50

DAFTAR TABEL

Tabel 2.1 Kriteria PICOC 10
Tabel 2.2 Rumusan Masalah 10
Tabel 2.3 Kriteria Penelitian yang Diambil dan Tidak Diambil 13
Tabel 2.4 Ekstraksi Data Untuk Pertanyaan Penelitian (RQ) 15
Tabel 2.5 Rangkuman Metode yang Pernah Diusulkan 33
Tabel 2.6 Daftar Referensi Systematic Literature Review 39
Tabel 3.1 Karakteristik Dari Dataset 43
Tabel 3.2 Spesifikasi Komputer yang Digunakan 47
Tabel 3.3 Perhitungan Accuracy 48
Tabel 3.4 Nilai AUC dan Interpretasinya 49

BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Klasifikasi bayes dikemukakan pertama kali oleh Thomas Bayes pada tahun 1950. Teori Bayes didasarkan pada perhitungan probabilitas dari suatu variabel. Klasifikasi bayes menggunakan perhitungan statistika [1] untuk menghitung peluang. Klasifikasi bayes mampu memprediksi kelas berdasarkan probabilitas sesuai fitur yang diberikan dan menentukan kelas mana yang paling optimal.

Algoritma *Naïve Bayes* merupakan algoritma untuk klasifikasi dengan efisiensi komputasi dan akurasi yang baik, terutama untuk fitur dan jumlah data yang besar [2]. *Naive Bayes* juga merupakan algoritma klasifikasi yang utama pada *data mining* dan banyak diterapkan dalam masalah klasifikasi karena memiliki performa klasifikasi yang tinggi. Algoritma pengklasifikasi *Naïve Bayes* juga memiliki beberapa keunggulan seperti mudah digunakan dan hanya membutuhkan satu kali iterasi data *training* [3] tetapi *Naïve Bayes* membutuhkan pengetahuan awal untuk dapat mengambil keputusan.

Pada penelitian yang dilakukan Hall *et all* pada tahun 2012 [4] telah melakukan perbandingan algoritma klasifikasi antara Naïve Bayes dengan C45, *Decision Tree, Logistic Regression, Naïve Bayes* dan *Neural Network* untuk menunjukan algoritma yang lebih baik dalam proses klasifikasi. Hasil algoritma yang terbaik adalah Naïve Bayes dimana *Naïve Bayes* memiliki akurasi yang lebih baik dari pembanding algoritma lainnya. Maka dari itu pada penelitian ini akan menggunakan algoritma *Naïve Bayes*

Secara umum model *Naïve Bayes* dalam mendapatkan hasil klasifikasi didapatkan dengan cara menghitung nilai probabilitas tiap fitur secara indenpenden yang artinya nilai fitur yang satu tidak bergantung dengan nilai fitur yang lain atau menilai setiap fitur berkonstribusi secara independen dalam mendapatkan hasil klasifikasi. Perhitungan dengan menggunakan asumsi independen ini untuk menyederhanakan perhitungan terlepas dari keterkaitan yang mungkin ada antar fitur. Performa algoritma ini menurun ketika antar fitur saling memiliki keterkaitan atau dependen, realitanya independensi fitur sering kurang tepat sehingga performa

klasifikasi *Naïve Bayes* bisa menurun [5] sehingga menyebabkan hasil kurang efektif.

Ada beberapa metode pendekatan pada *Naïve Bayes* untuk meningkatkan kinerja algoritma ini diantaranya menggunakan metode manipulasi fitur, metode manipulasi data dan metode manipulasi struktur [6]. Metode pendekatan manipulasi fitur terbagi menjadi dua metode yaitu metode pembobotan fitur dan seleksi fitur. Penelitian menggunakan beberapa metode pendekatan ini mampu dengan efektif menurunkan asumsi independensi fitur pada algoritma *Naïve Bayes* [7]

Meningkatkan kinerja *Naïve Bayes* berdasarkan manipulasi fitur diterapkan dengan memberikan bobot yang berbeda ke tiap fitur (pembobotan fitur) dan memilih fitur yang terbaik (seleksi fitur). Adapun beberapa metode pendekatan yang menerapkan manipulasi fitur menggunakan pembobotan fitur adalah *deep feature weighting* [8], *attribute weighted using a local optimization* [5], *self adaptive attribute weighting* [6]. Beberapa metode pendekatan yang menggunakan manipulasi fitur berdasarkan seleksi fitur adalah *correlation based feature selection* (CFS), *feature selection using ant colony optimization* [9], *correlation based algorithm* [10], *relief attribute ranking algorithm* [11]

Meningkatkan kinerja *Naïve Bayes* berdasarkan manipulasi data diterapkan dengan memberikan bobot yang berbeda ke tiap data[7]. Beberapa metode pendekatan yang menggunakan manipulasi data adalah *combined neighourhood Naïve Bayes* (CNNB) [7], *local value diffence metric* [12], *instance weighted Naïve Bayes* (IWNB) [7].

Meningkatkan kinerja *Naïve Bayes* berdasarkan manipulasi struktur penerapannya dengan menambahkan atau memperluas struktur jaringan di algoritma *Naïve Bayes* (*strukture extension*) [13]. Beberapa metode pendekatan yang menggunakan manipulasi struktur adalah *Random One Dependence Estimator* (RODE) [13], *structure extended multinominal Naïve Bayes* (SEMNB) [14], *tree-augmented Naïve Bayes* (TAN) [15].

Menurut lee [16] metode pembobotan fitur lebih fleksibel daripada metode yang lain maka penelitian ini difokuskan pada pembobotan fitur. Pembobotan fitur menilai setiap fitur tidak mempunyai kedudukan yang sama, sebagian fitur

mempunyai kedudukan yang lebih penting daripada fitur lain sehingga tujuan memberikan bobot yang tidak sama disetiap fitur yaitu untuk mendapatkan satu kelompok fitur yang tidak sama kedudukannya [6] tanpa menghilangkan fitur yang tidak ada kaitannya.

Banyak penelitian yang telah dilakukan pada topik pembobotan fitur. Termasuk Wu dan Cai pada tahun 2011 melakukan penelitian tentang [17] metode *Differential Evolution Weighted Naive Ba*yes (DEWNB) yang memiliki performa lebih baik daripada algoritma *evolusioner* yang lain untuk menentukan bobot fitur. Metode ini memiliki kelebihan yaitu mampu mendapatkan bobot optimal secara otomatis karena menerapkan proses mutasi, crossover dan seleksi untuk bobot atribut. Selain itu metode DEWNB ini juga tidak memerlukan pengetahuan apriori untuk mendapatkan bobot atribut, akan tetapi metode ini memiliki kelemahan yaitu konvergensi prematur sehingga penentuan nilai bobot sering kurang tepat [17].

Wu, Pan, Zhu dan Caian [6] telah mengusulkan *Arificial Immune System Weight Naïve Bayes* (AIWNB) dimana metode ini mampu menyesuaikan bobot secara mandiri sehingga probabilitas tiap fitur bisa ditentukan lebih akurat. Metode ini mengikuti cara kerja imun, namun jika hasil mutasi dan pembangkitan antibodi secara acak yang dihasilkan oleh *death rate* tidak tepat maka menghasilkan solusi yang lokal optimum, sehingga pemberian bobot tiap fitur kurang maksimal[6].

Lungan Zhang**,** Jiang, Li dan Kong [18] juga telah mengusulkan metode *Gain Ratio Weight Naïve Bayes* (GRWNB). Umumnya metode ini bekerja dengan cara memberikan nilai setiap fitur dengan nilai nol atau bilangan bulat positif. Metode ini mengasumsikan bahwa semua fitur hanya memiliki dua nilai nol dan tidak nol. Menurut Lungan Zhang**,** Jiang, Li dan Kong [18] metode GRWNB merupakan metode yang sederhana dan memiliki waktu komputasi yang efesien, namun ketika satu kelas memiliki lebih banyak dokumen pelatihan daripada yang lain, hasil pemilihan bobot menjadi buruk.

Menurut kohonen di tahun 1998 [19] *self organizing map* merupakan metode yang efektif untuk visualisasi data berdimensi tinggi yang mampu mengubah hubungan statistik nonlinier dengan cara menerapkan pemetaan secara teratur dari distribusi berdimensi tinggi ke grid berdimensi lebih rendah. Pada penelitian yang

dilakukan Shieh dan Liao pada tahun 2012 [20] *self organizing map* merupakan metode yang efektif untuk data yang memiliki fitur banyak dan efektif pada pemberian bobot dalam mengklaster data. Sehingga pada penelitian ini akan menggunakan pemberian bobot dengan metode *self organizing map* (SOM). Metode *self organizing map* dimanfaatkan untuk memberikan bobot disetiap fitur. Metode *self organizing map* mampu mendapatkan bobot yang tepat dengan mencari bobot yang paling mendekati dengan *euclidean distance* dan meng*update* hasil pembobotan sampai mendapatkan bobot yang tepat, sesuai dengan nilai iterasi yang ditetapkan dengan mengurangi laju pembelajaran (α) [20]. Pada penelitian ini menggunakan 21 dataset dari UCI Dataset *Repository* [5] dan menerapkan 10 *fold cross validation* karena validasi ini sudah menjadi standar. Untuk mengevaluasi hasil penelitian akan menggunakan *Accuracy* (ACC) dan *Area Under Curve* (AUC)

1.2 Identifikasi Masalah

Berdasarkan latar belakang masalah yang diuraikan di atas, masalah penelitian (*Research Problem* (RP) yang diangkat pada penelitian ini adalah Algoritma *Naïve Bayes* merupakan algoritma untuk klasifikasi dengan efisiensi komputasi dan akurasi yang baik, terutama untuk fitur dan jumlah data yang besar namun kinerja algoritma ini menurun ketika dataset yang digunakan memiliki fitur saling berkaitan karena *Naïve Bayes* menggunakan asumsi independen fitur.

1.3 Rumusan Masalah

Berdasarkan latar belakang masalah dan identifikasi masalah, maka pada penelitian ini dibuat rumusan masalah (*Research Question* (RQ) adalah bagaimana pengaruh penerapan pembobotan fitur menggunakan *self organizing map* untuk menangani asumsi independen fitur terhadap akurasi algoritma *Naïve Bayes*?

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah mengembangkan metode pembobotan fitur menggunakan *Self Organizing Map* untuk menangani asumsi independensi fitur pada algoritma *Naïve Bayes* sehingga meningkatkan hasil akurasi.

1.5 Manfaat Penelitian

Manfaat yang didapatkan setelah tujuan tercapai dan rumusan masalah terpecahkan dalam penelitian ini adalah

1. Hasil penelitian ini diharapkan dapat digunakan untuk menangani asumsi independen fitur sehingga kinerja *Naïve Bayes* lebih meningkat.
2. Pengembangan teori yang berkaitan dengan penerapan pembobotan fitur menggunakan *Self Organizing Map* pada algoritma *Naïve Bayes*

1.6 Sistematika Penulisan

Pada tesis ini akan dibagi menjadi lima bab dan disetiap bab akan dibagi lagi menjadi beberapa subbab sesuai topik yang dibahas. Sistematika pada penulisan ini adalah:

Bab1 Pendahuluan

Pada bab ini berisi uraian tentang latar belakang masalah, identifikasi masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

Bab2 Tinjauan Pustaka

Pada bab ini membahas tinjauan studi yang berisi metode untuk menangani asumsi independensi fitur di algoritma *Naïve Bayes* yang sudah pernah digunakan dan usulan metode, selain itu pada bab ini juga akan dibahas mengenai tinjauan pustaka.

Bab3 Metode Penelitian

Pada bab ini menyajikan tahapan penelitian yang digunakan dalam melakukan penelitian ini, analisa masalah dan tinjauan pustaka, pengumpulan dataset, metode yang diusulkan, eksperimen dan pengujian metode, evaluasi dan validasi hasil juga dibahas dalam bab ini.

Bab 4 Hasil dan pembahasan

Pada bab ini akan dibahas mengenai hasil dari penelitian dan pembahasannya. Hasil pada bab ini akan menunjukkan ukuran dari performa metode yang diusulkan dibandingkan dengan metode yang lain.

Bab 5 Penutup

Pada bab ini menyajikan kesimpulan dari hasil penelitian, dan saran untuk penelitian lebih lanjut.

BAB 2
TINJAUAN PUSTAKA

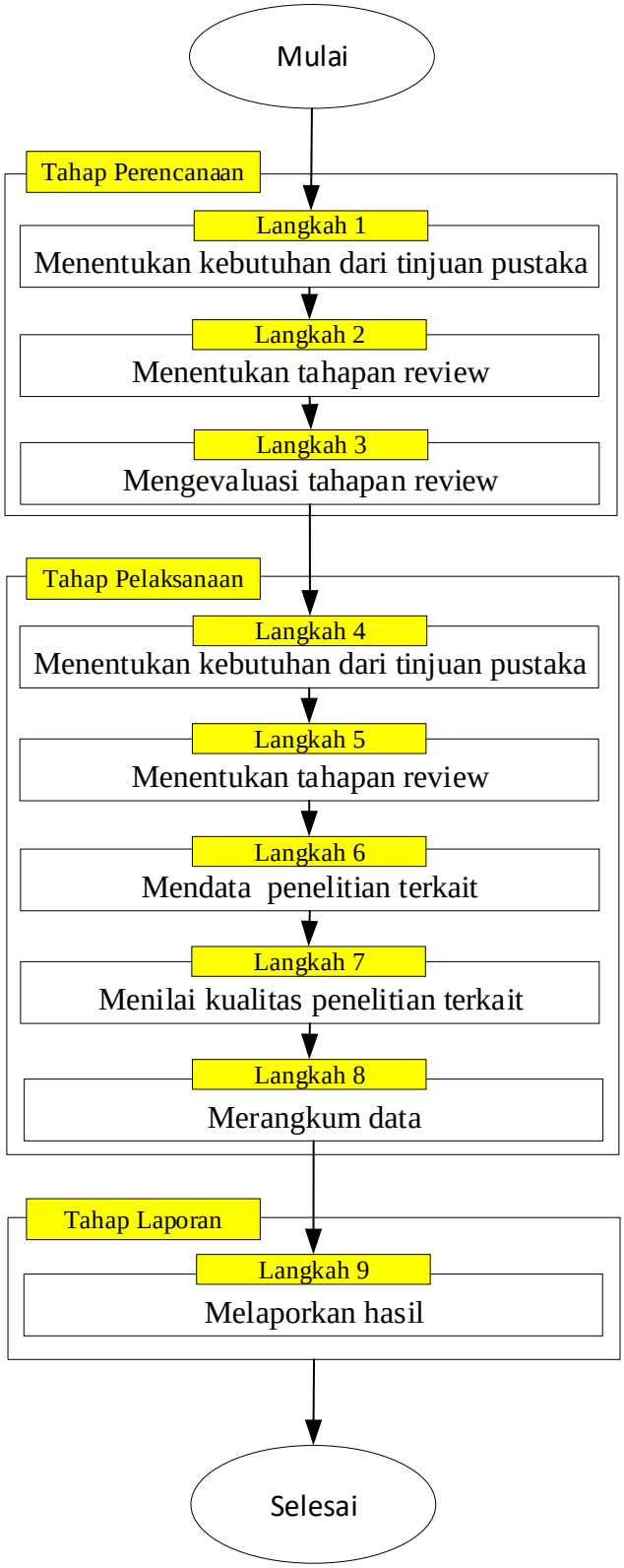
2.1 Pendahuluan

Sebelum melakukan penelitian lebih lanjut, diperlukan kajian terhadap penelitian terdahulu untuk mengetahui lebih lanjut mengenai metode yang sudah pernah dilakukan. Tinjauan pustaka ini dilakukan untuk mengetahui *state of the art* tentang penelitian pada algoritma *Naïve Bayes* dengan permasalahan asumsi independensi fitur. Ada sebanyak 24 jurnal mengenai algoritma *Naïve Bayes* dengan permasalahan asumsi independensi fitur yang dipublikasikan antara 1 Januari 2010 hingga 31 Desember 2016 akan diinvestigasi. Penelitian tentang algoritma *Naïve Bayes* dengan permasalahan asumsi independensi fitur telah banyak dilakukan dan penelitian tersebut sangat beragam dan kompleks, sehingga diperlukan sebuah gambaran yang komprehensif mengenai *state of the art* dari penelitian tersebut. Pada bab ini akan dibahas mengenai tinjauan pustaka secara sistematis, di mana hal tersebut akan dapat memberikan gambaran mengenai apa saja yang telah dipublikasikan, tetapi review ini bukanlah review yang lengkap mengenai seluruh jurnal yang telah diterbitkan. Metode review, sumber literatur, gaya dan perumusan pertanyaan pada bab ini terinspirasi oleh Wahono [21].

2.2 Metode Review

Pada penelitian ini akan menggunakan pendekatan sistematik untuk mereview penelitian tentang data dengan algoritma *Naïve Bayes* dengan permasalahan asumsi independensi fitur. *Systematic Literature Review* (SLR) merupakan sebuah proses untuk mengidentifikasi, menilai, dan menginterpretasikan semua penelitian yang tersedia dengan tujuan untuk memberikan jawaban untuk rumusan masalah (*Research Question* (RQ)) tertentu [22]. Dalam panduan yang telah dibuat oleh Kitchenham pada tahun 2007 [22], maka tinjauan pustaka ini akan disusun berdasarkan *Systematic Literature Review*.

Tahapan SLR dapat dilihat pada Gambar 2.1 dibagi menjadi tiga tahap, yaitu tahap perencanaan, kemudian tahap pelaksanaan, dan yang terakhir tahap laporan. Tahapan pertama dari tinjauan pustaka secara sistematis adalah menentukan kebutuhan dari dilakukannya tinjauan pustaka (langkah 1). Kebutuhan dari dilakukannya tinjauan pustaka telah dijabarkan pada pendahuluan pada bab ini. Kemudian langkah berikutnya yaitu merencanakan tahapan review. Dalam tahap merencanakan tahapan review ini diharapkan dapat mengurangi bias dari penelitian. Pada tahap ini akan di tentukan rumusan masalah, strategi pencarian, proses seleksi dari jurnal yang akan diambil termasuk kriteria kriteria jurnal yang akan direview dan yang tidak direview, menilai kualitas dari jurnal, merangkum data, dan yang terakhir melaporkan data. Tahapan review akan dijabarkan pada sub bab 2.2.1, 2.2.2, 2.2.3, 2.2.4, dan 2.2.5.



Gambar 2.1 Tahapan *Systematic Literature Review*

2.2.1 Rumusan Masalah

Rumusan masalah (*Research Question* (RQ)) ini dilakukan agar review terhadap penelitian terkini tetap fokus. Pada tahap ini akan menggunakan desain kriteria dari Kitchenham (2007) [22], yaitu *Population, Intervention, Comparison, Outcomes,* dan *Context* (PICOC). Pada penelitian ini akan menggunakan kriteria PICOC seperti yang dapat dilihat pada Tabel 2.1

Tabel 2.1 Kriteria PICOC

Population	*Naïve Bayes*
Intervention	Independensi fitur, pembobotan fitur
Comparison	-
Outcomes	Mengatasi masalah independensi fitur
Context	Dataset UCI

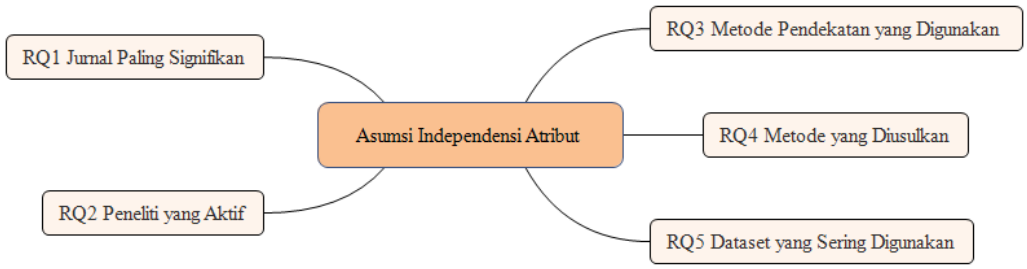
Untuk rumusan masalah pada tinjauan pustaka dapat dilihat pada Tabel 2.2. Rumusan masalah ini hanya ditujukan untuk tinjauan pustaka, tidak untuk penelitian utama. Peta pikiran rumusan masalah secara visual dapat dilihat pada Gambar 2.2

Tabel 2.2 Rumusan Masalah

	Rumusan masalah	Motivasi
RQ1	Jurnal mana yang paling banyak mempublikasikan penelitian tentang asumsi independensi fitur pada algoritma *Naïve Bayes*	Mengidentifikasi jurnal yang paling signifikan dalam masalah asumsi independensi fitur pada algoritma *Naïve Bayes*
RQ2	Siapa peneliti yang paling aktif dalam penelitian tentang asumsi independensi fitur pada algoritma *Naïve Bayes*	Mengidentifikasi peneliti yang paling aktif dalam penelitian tentang asumsi independensi fitur pada algoritma *Naïve Bayes*
RQ3	Metode pendekatan apa yang sering digunakan dalam penelitian	Mengidentifikasi metode pendekatan yang sering digunakan dalam penelitian asumsi

	Rumusan masalah	Motivasi
	tentang asumsi independensi fitur pada algoritma *Naïve Bayes*	independensi fitur pada algoritma *Naïve Bayes*
RQ4	Metode apa yang pernah diusulkan untuk menyelesaikan asumsi independensi fitur pada algoritma *Naïve Bayes*	Mengidentifikasi metode yang pernah diusulkan untuk menyelesaikan asumsi independensi fitur pada algoritma *Naïve Bayes*
RQ5	Dataset apa yang sering digunakan dalam penelitian menangani asumsi independen fitur pada algoritma *Naïve Bayes*	Mengidentifikasi dataset yang sering digunakan dalam penelitian menangani asumsi independen fitur pada algoritma *Naïve Bayes*

RQ3, RQ4 dan RQ5 akan digunakan untuk mendukung penelitian utama. Sedangkan RQ1 dan RQ2 akan digunakan untuk mendukung konteks penelitian.



Gambar 2.2 Peta Pikiran Rumusan Masalah

2.2.2 Strategi Pencarian

Pada tahap ke empat, mencari penelitian terkait, dibutuhkan beberapa aktifitas, antara lain memilih sumber pencarian, menentukan kata kunci pencarian, melakukan pencarian awal, mengevaluasi dan menyusun ulang kata kunci pencarian, dan mengelola hasil pencarian dari sumber pencarian berdasarkan kata kunci pencarian. Sebelum memulai pencarian, menentukan kata kunci harus

terlebih dahulu dilakukan untuk memperbesar kemungkinan menemukan penelitian terkait yang sesuai.

Pada penelitian ini akan digunakan sumber pencarian dari:

1. ScienceDirect (sciencedirect.com)
2. IEEE eXplore (ieeexplore.ieee.org)

Kata kunci pencarian akan disusun berdasarkan langkah langkah berikut:

1. Identifikasi kata kunci dari PICOC, terutama dari *population* dan *intervention*
2. Identifikasi kata kunci dari rumusan masalah
3. Identifikasi kata kunci dari judul, abstraksi dan kata kunci yang relevan
4. Identifikasi kata kunci dari sinonimnya
5. Mengkonstruksi kata kunci yang kompleks, yang terdiri dari beberapa kata kunci dengan menggunakan boolean AND dan OR.

Kata kunci pencarian yang digunakan adalah:

Naïve Bayes* *and (Indepen*OR irrele*) AND (attribut* OR featur*)*

Pada saat penyesuaian kata kunci pencarian dibentuk, kata kunci pencarian awal akan terus disimpan hingga hasil pencarian dapat secara signifikan menaikkan kecocokan pencarian. Pencarian dibatasi pada tahun 2010-2016. Publikasi yang diambil adalah publikasi berupa jurnal dan prosiding konferensi. Selain itu, pencarian juga dibatasi pada publikasi yang menggunakan bahasa inggris.

2.2.3 Seleksi Pencarian

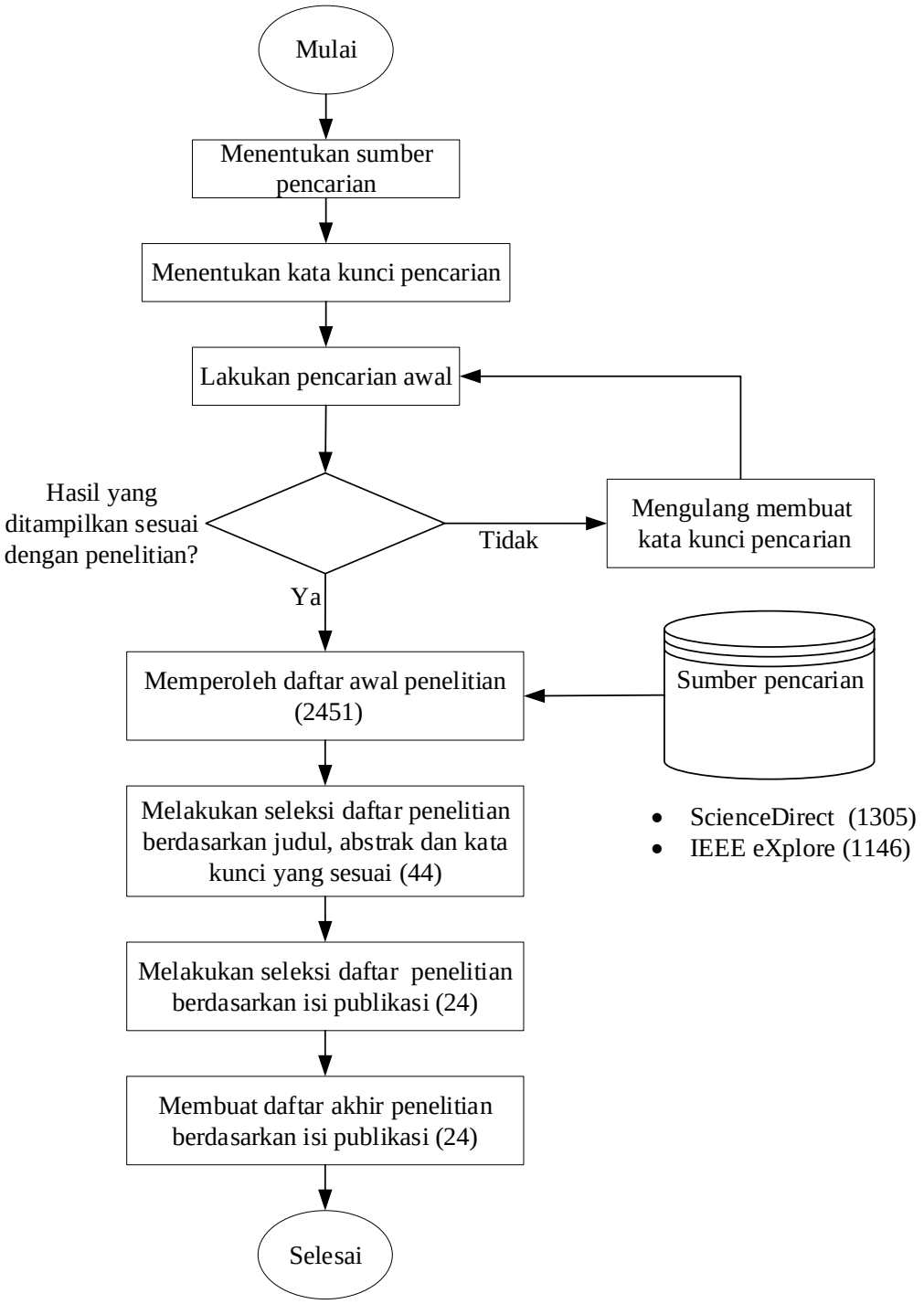
Pencarian terhadap publikasi yang ada terdapat beberapa kriteria yang akan digunakan dan juga ada beberapa kriteria yang tidak akan digunakan. Kriteria-kriteria tersebut dapat dilihat pada Tabel 2.3

Tabel 2.3 Kriteria Penelitian yang Diambil dan Tidak Diambil

Kriteria penelitian yang diambil	Penelitian yang membahas mengenai asumsi independensi fitur pada algoritma *Naïve Bayes*
	Untuk penelitian yang mempunyai dua tipe publikasi jurnal dan *conference*, maka akan diambil tipe publikasi yang jurnal
	Untuk penelitian yang duplikasi, akan diambil data terlengkap dan terbaru
Kriteria penelitian yang tidak diambil	Penelitian yang membahas asumsi independensi fitur pada algoritma *Naïve Bayes*, tetapi tidak mengusulkan metode untuk mengatasi asumsi independensi fitur pada algoritma *Naïve Bayes*
	Penelitian yang tidak menggunakan validasi yang kuat
	Penelitian tidak ditulis dalam bahasa inggris.

Software Mendeley (mendeley.com) digunakan untuk menyimpan dan mengelola hasil pencarian. Detail proses pencarian dan jumlah penelitian yang ditemukan pada setiap tahapnya dapat dilihat pada Gambar 2.3. Penelitian yang tidak menampilkan hasil penelitian tidak akan digunakan. Hasil pencarian awal terdapat 2.451 paper di bidang *computer science*, kemudian dilakukan seleksi berdasarkan judul, abstrak, dan kata kunci yang tepat didapatkan 44 paper.

Setelah melalui tahap seleksi dan menilai kualitas dari penelitian yang didapatkan, maka ada 24 penelitian yang akan direview. Tahap seleksi sudah memperhitungkan kriteria-kriteria yang terdapat pada Tabel 2.3. Selain itu penelitian yang sama juga tidak disertakan.



Gambar 2.3 Langkah-Langkah Pencarian *Literature* Penelitian

2.2.4 Ekstrasi Data

Ekstraksi terhadap publikasi penelitian terhadap asumsi independensi fitur pada algoritma *Naïve Bayes* diperlukan untuk mendapatkan data yang berhubungan dengan RQ pada Bab ini Tabel 2.4

Tabel 2.4 Ekstraksi Data Untuk Pertanyaan Penelitian (RQ)

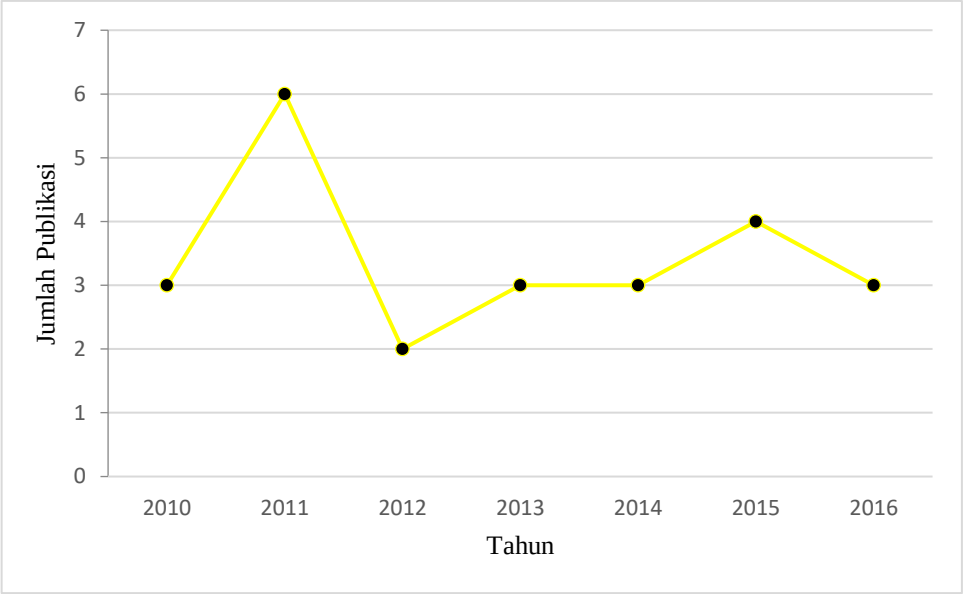
Ekstraksi Data	Pertanyaan penelitian
Peneliti dan tahun publikasi	RQ1, RQ2
Metode pendekatan yang digunakan	RQ3
Metode yang diusulkan	RQ4
Dataset yang sering digunakan	RQ5

2.2.5 Penilaian Kualitas Penelitian dan Sintesis Data

Penilaian kualitas penelitian dapat digunakan untuk membantu menginterpretasikan kualitas dari temuan dan untuk menentukan kekuatan kesimpulan yang diuraikan. Sedangkan tujuan dari sintesis data adalah untuk mengumpulkan bukti dari publikasi yang sudah didapatkan untuk menjawab pertanyaan penelitian. Sintesis data yang digunakan pada penelitian ini, secara umum akan berupa sintesis narasi. Beberapa tabel dan alat visual akan digunakan untuk menunjang penjelasan pada penelitian ini.

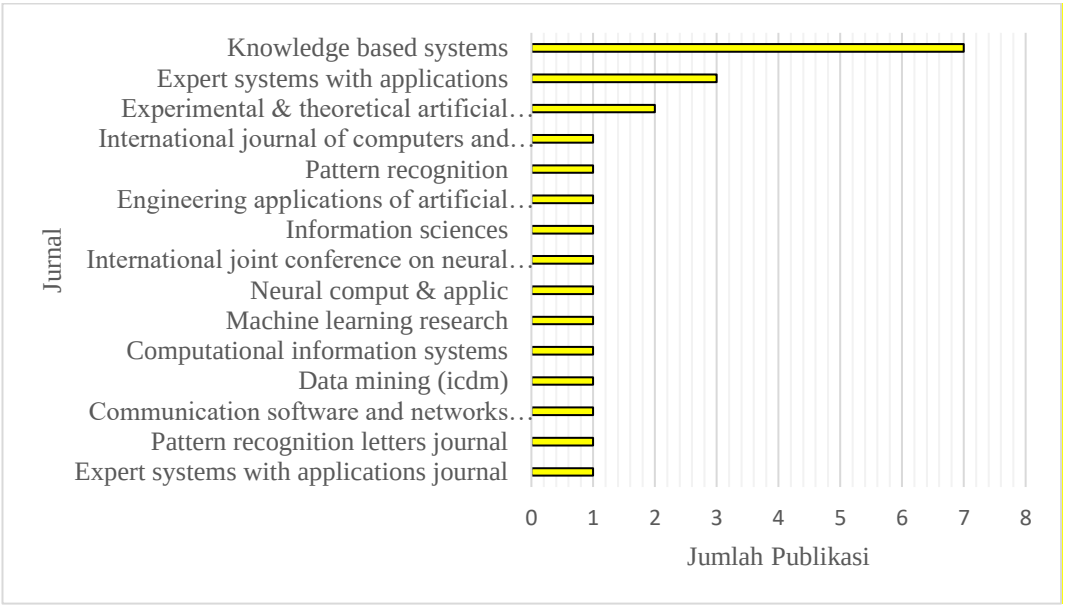
2.3 Publikasi Jurnal Ilmiah

Pada tinjauan pustaka ini akan menganalisa sebanyak 24 publikasi yang membahas tentang asumsi independensi fitur pada algoritma *Naïve Bayes*. Seperti yang sudah dijelaskan sebelumnya, tinjauan pustaka ini akan dibatasi pada jurnal yang dipublikasi tahun 2010 sampai 2016. Rentang waktu tersebut untuk melihat apakah penelitian pada asumsi independensi fitur pada algoritma *Naïve Bayes* masih relevan. Pada Gambar 2.4 dapat dilihat bahwa tren penelitian dari tahun 2010 sampai 2016 mengalami peningkatan, sehingga dapat disimpulkan bahwa penelitian tentang asumsi independensi fitur pada algoritma *Naïve Bayes* masih sangat relevan sampai saat ini.



Gambar 2.4 Jumlah Publikasi Tahun 2010 - 2016

Kemudian pada Gambar 2.5 dapat dilihat jurnal yang mempublikasikan paper tentang asumsi independensi fitur pada algoritma *Naïve Bayes*. Sebagai catatan, jurnal yang dimaksud adalah jurnal yang mempublikasikan paper yang telah terpilih.

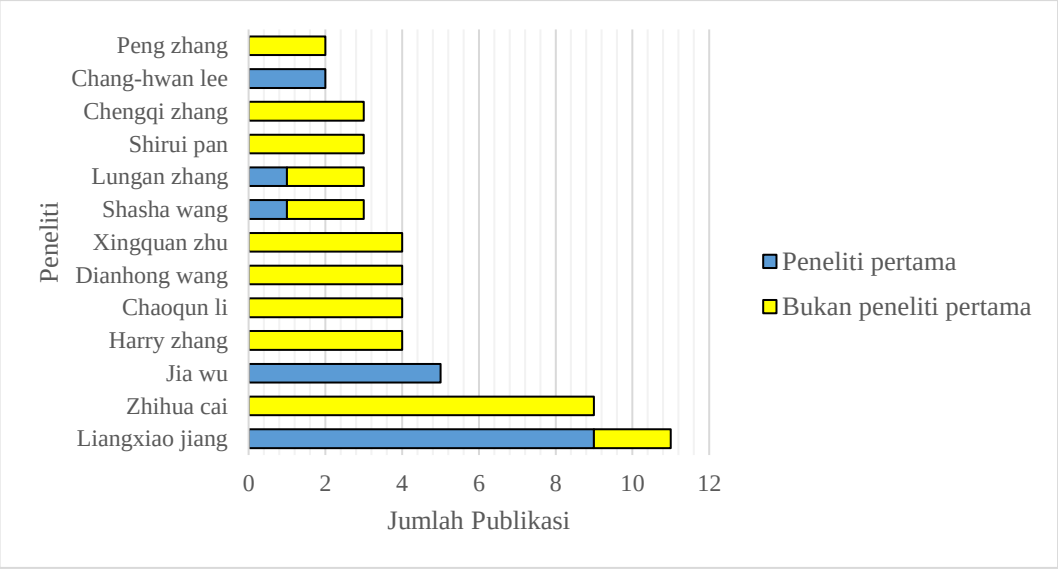


Gambar 2.5 Jumlah Publikasi Paper Asumsi Independensi Fitur

2.4 Peneliti yang Paling Aktif

Dari publikasi yang didapat, peneliti yang paling aktif dan berkontribusi pada penelitian tentang asumsi independensi fitur pada algoritma *Naïve Bayes* akan

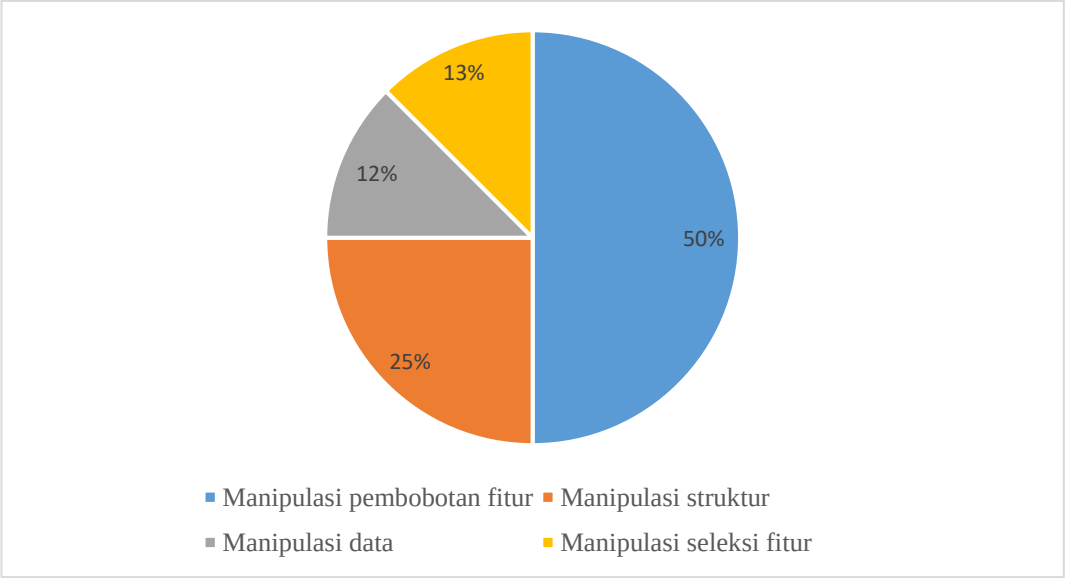
diinvestigasi dan diidentifikasi. Gambar 2.6 menunjukkan peneliti yang paling aktif pada penelitian tentang asumsi independensi fitur pada algoritma *Naïve Bayes*. Sebagai catatan, data yang terdapat pada Gambar 2.6 merupakan peneliti dengan jumlah publikasi lebih dari 1 kali.



Gambar 2.6 Peneliti yang Aktif dan Jumlah Publikasi

2.5 Metode Pendekatan yang Digunakan

Dari 24 penelitian yang telah dilakukan, ada beberapa metode pendekatan yang digunakan untuk mengatasi masalah asumsi independensi fitur pada algoritma *Naïve Bayes*. Metode pendekatan tersebut pada bab ini akan dibahas satu persatu untuk mengetahui lebih lanjut penelitian yang sudah dilakukan oleh peneliti sebelumnya. Metode pendekatan secara garis besar dapat dikelompokkan menjadi 4 besar, yaitu metode pendekatan yang menggunakan manipulasi pembobotan fitur, manipulasi seleksi fitur, manipulasi data, manipulasi struktur [23]. Pada Gambar 2.7 dapat dilihat distribusi dari metode untuk mengatasi masalah asumsi independensi fitur pada algoritma *Naïve Bayes*.



Gambar 2.7 Distribusi Metode Untuk Mengatasi Asumsi Independensi Fitur

2.5.1 Metode Manipulasi Pembobotan Fitur

Metode Manipulasi pembobotan fitur dipakai oleh beberapa peneliti diantaranya Lee *et al* (2011) [16], Wu *et al* (2011) [17], dan Zaidi *et al* (2014) [24]. Metode ini digunakan untuk menghitung nilai optimal atau bobot antar fitur [23] sehingga probabilitas masing masing fitur bisa ditentukan lebih akurat.

2.5.2 Metode Manipulasi Seleksi Fitur

Metode manipulasi seleksi fitur pernah digunakan oleh Deisy (2010) [25] dan Kannan (2010) [26]. Metode ini mengambil beberapa kombinasi fitur dengan didasarkan dari adanya pembobotan yang dilakukan terhadap fitur tersebut, dengan menggunakan parameter yang ditetapkan. Dari sejumlah fitur diambil fitur yang mempunyai bobot lebih tinggi menjadi pioritas untuk diklasifikasikan kedalam metode *Naïve Bayes*.

2.5.3 Metode Manipulasi Data

Metode Manipulasi data digunakan oleh Zhang *et al* (2016) [18], Farid *et al* (2014) [27] dan Jiang *et al*(2016) [7] yang di mana *Naïve Bayes* mengasumsikan bahwa semua fitur dari data independen satu sama lain [28] sehingga metode ini

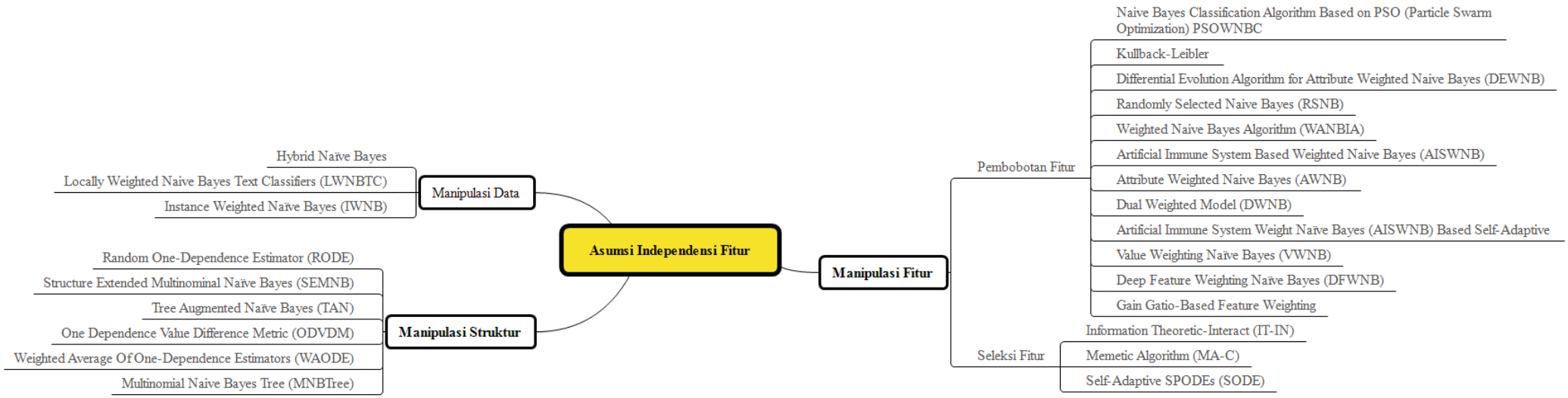
memberikan nilai setiap data sehingga mampu melemahkan asumsi independen di *Naïve Bayes*

2.5.4 Metode Manipulasi Struktur

Metode Manipulasi struktur digunakan oleh Lee *et al* (2011) [29], Jiang (2014) [13] dan Jiang *et al* (2016) [15]. Metode ini diterapkan dengan memperluas struktur jaringan *Naïve Bayes* dan menciptakan dependensi atibut yang baik dengan pemilihan fitur yang tepat yang dapat memperbaiki akurasi klasifikasi

2.6 Metode yang Pernah Diusulkan

Sejak tahun 2010, telah ada sebanyak 24 penelitian telah dilakukan untuk mengatasi asumsi independensi fitur pada algoritma *Naïve Bayes*. Peta pikiran dari penelitian tersebut dapat dilihat pada Gambar 2.8 Penelitian tersebut pada bab ini akan dibahas satu persatu untuk mengetahui lebih lanjut penelitian yang sudah dilakukan oleh peneliti sebelumnya. Pada bab ini juga akan digambarkan flow diagram secara umum dari masing-masing metode yang pernah diusulkan, sehingga jika ada blok alur yang tidak digunakan pada metode yang diusulkan, maka akan diberi tanda berupa kotak berarsir.

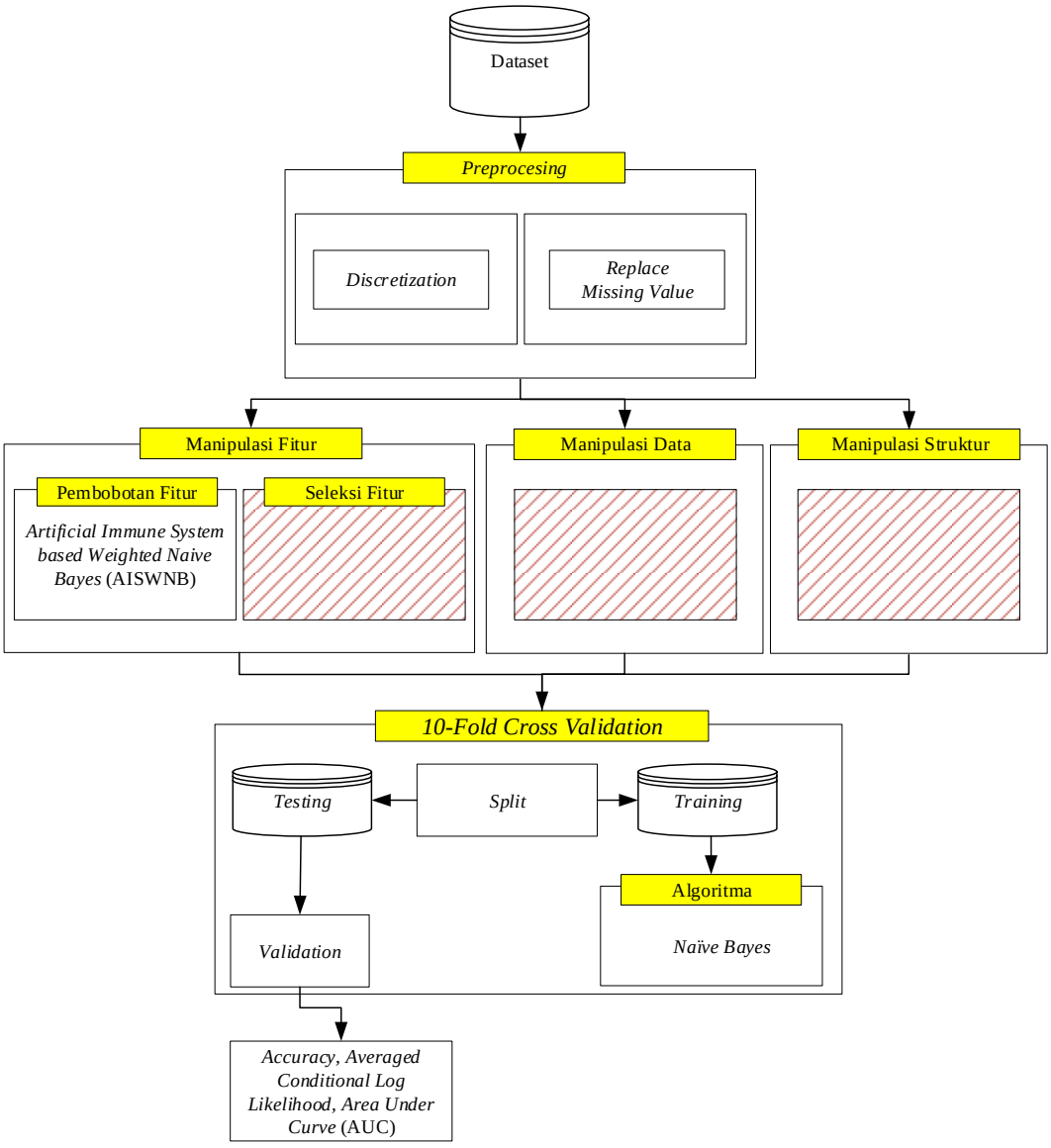


Gambar 2.8 Peta Pikiran Penelitian Asumsi Independensi Fitur

2.6.1 Metode Wu, Pan, Zhu dan Cai.

Pada tahun 2015, Wu, Pan, Zhu dan Cai [6] melakukan penelitian tentang menghitung bobot optimal antar fitur untuk mengurangi asumsi independensi fitur pada algoritma *Naïve Bayes* menggunakan *Arificial Immune System* disingkat AISWNB. Metode ini mampu menyesuaikan bobot secara mandiri sehingga probabilitas tiap fitur bisa ditentukan lebih akurat karena metode ini mengikuti cara kerja imun, yaitu melalui tahap penggandaan diri, pembelahan, mutasi dan *memory*. Metode ini bisa menentukan bobot yang tepat selama proses pembelajaran sehingga kinerja *Naïve Bayes* semakin baik.

Bagan metode penelitian ini bisa dilihat pada Gambar 2.9. Pada penelitian ini menggunakan 36 dataset dari UCI *repository* dan *preprocessing* menggunakan *discretization* dan *replace missing value*. Untuk mengukur performa dari metode yang diusulkan, pada penelitian tersebut menggunakan alat ukur *accuracy, averaged conditional log likelihood, area under curve* (AUC) dan metode validasi yang digunakan adalah *10-fold cross validation.* Berdasarkan hasil perhitungan dan perbandingan dengan metode lainnya seperti CFSWNB (pembobotan menggunakan korelasi, GRWNB (pembobotan dengan *gain ratio*), MIWNB (pembobotan menggunakan *mutual information*), ReFWNB (pembobotan menggunakan fitur *estimation*), TreeWNB (pembobotan berdasarkan tingkat bergantungan tiap fitur), SBC (pemilihan fitur berdasarkan *decision tree*) dan RMWNB (pembobotan fitur yang dipilih dengan *random* antara 0 dan 1), metode AISWNB memiliki performa yang lebih baik dalam hal konvergensi disebabkan bobot dipilih yang paling optimal.

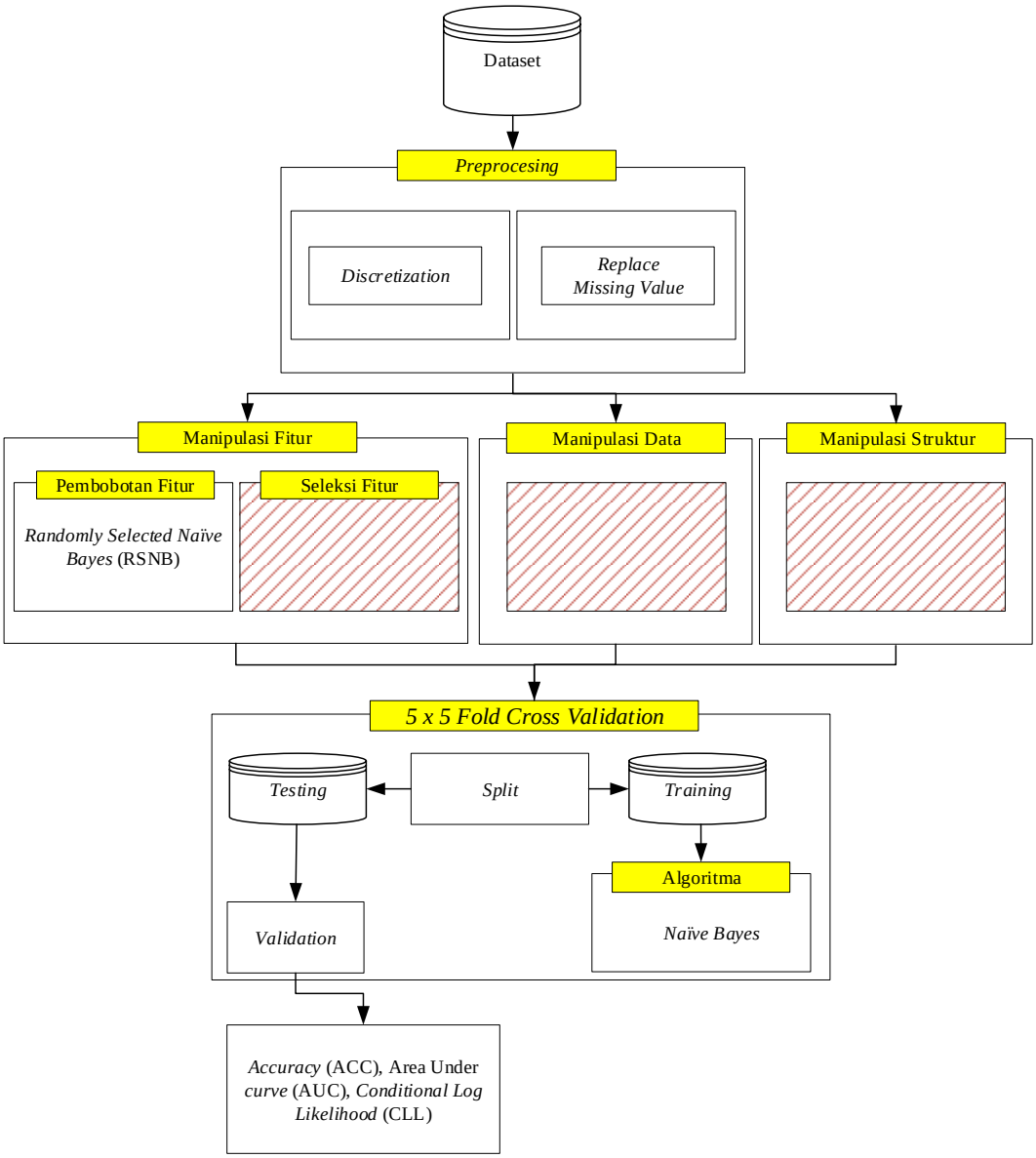


Gambar 2.9 Metode Wu, Pan, Zhu dan Cai [6]

2.6.2 Metode Jiang, Cai, Zhang dan Wang

Metode Jiang, Cai, Zhang dan Wang [30] yang diberi nama *Randomly Selected Naive Bayes* (RSNB) melakukan penelitian pada tahun 2012, metode ini diusulkan untuk mengatasi masalah indenpendensi fitur dengan metode pendekatan manipulasi fitur menggunakan pemilihan fitur secara *wrappers* atau menggunakan algoritma itu sendiri sebagai *black box* untuk mengevaluasi subset fitur yang dipilih secara acak dari fitur terbaik pada setiap iterasi.

Gambar 2.10 merupakan bagan metode penelitian yang dilakukan Jiang *et al*. Pada penelitian ini nilai yang hilang diganti dengan *mode* dan *means* dari data yang tersedia, tahapan berikutnya menggunakan *discretization* dengan metode *unsupervised ten bin discretization* pada aplikasi WEKA. Eksperimen ini menggunakan 36 dataset dari UCI *repository* dan ada beberapa fitur didataset yang dihapus secara manual karena dianggap tidak berguna contoh fitur nomor rumah sakit di dataset colic.ORIG. Metode validasi yang digunakan adalah 5 x 5 *fold cross validation* dan dievaluasi menggunakan *Accuracy* (ACC), *Area Under Curve* (AUC), *Conditional Log Likelihood* (CLL).

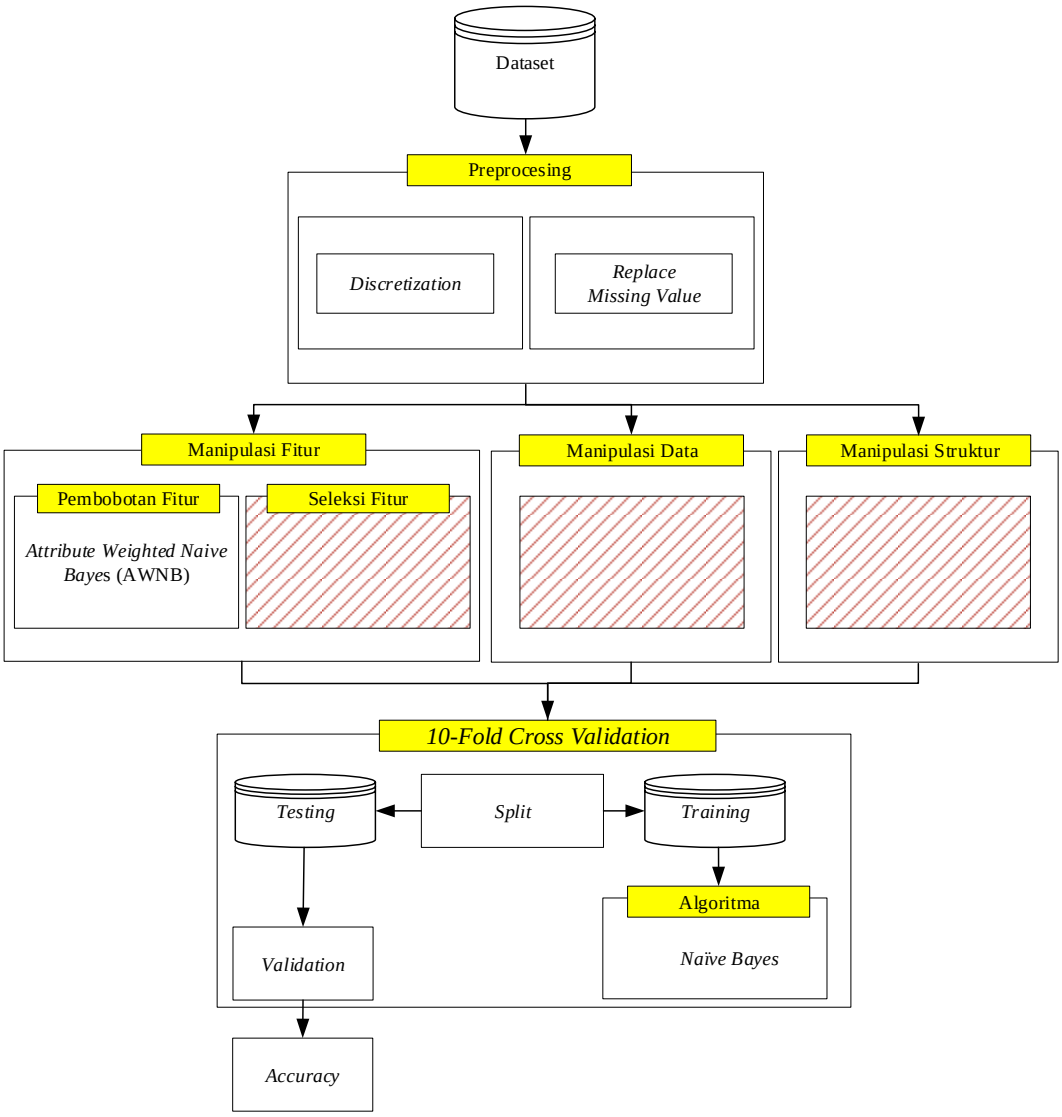


Gambar 2.10 Metode Jiang, Cai, Zhang dan Wang [30]

2.6.3 Metode Taheri, Yearwood, Mammadov dan Seifollahi

Pada tahun 2014, Taheri, Yearwood, Mammadov dan Seifollahi [5] melakukan penelitian untuk mengurangi independensi fitur dalam upaya meningkatkan kinerja pengklasifikasi *Naïve Bayes*. Metode pembobotan baru yang diberi nama *Attribute Weighted Naive Bayes* (AWNB) untuk pengelompokan bobot *Naïve Bayes*, di mana untuk setiap fitur menggunakan lebih dari satu bobot tergantung pada jumlah label kelas. Fungsi objektif yang terdiri dari bobot fitur berdasarkan struktur pengklasifikasi *Naïve Bayes* kemudian dimodelkan untuk mengoptimalkan bobot fitur. Fungsi objektif ini dioptimalkan dengan optimasi lokal dengan menggunakan metode *quasisecant*. Nilai awal dalam metode *quasisecant* ditentukan dengan nilai satu, yang berarti bahwa klasifikasi *Naïve Bayes* akan melakukan sejumlah percobaan pada beberapa data.

Pada penelitian ini menggunakan 16 dataset yang mana 11 dataset dari UCI *repository* dan 5 dataset dari LIBSVM. Pada eksperimen ini melakukan perbandingan pada *Naïve Bayes*, *Tree Augmented Naïve Bayes* (mencari hubungan antar fitur ditemukan dengan menggunakan struktur pohon), *Improved Naïve Bayes* (menggunakan probablilitas untuk menemukan korelasi fitur) dan *Atribut Weighted Naïve Bayes* (AWNB) Untuk mengukur performa dari metode yang diusulkan, pada penelitian tersebut menggunakan alat ukur *accuracy* dan metode validasi yang digunakan adalah 10-*fold cross validation.* Hasil eksperimen menunjukan metode yang diusulkan memiliki dampak positif pada akurasi *Naïve Bayes*. Bagan metode penelitian ini bisa dilihat pada Gambar 2.11.

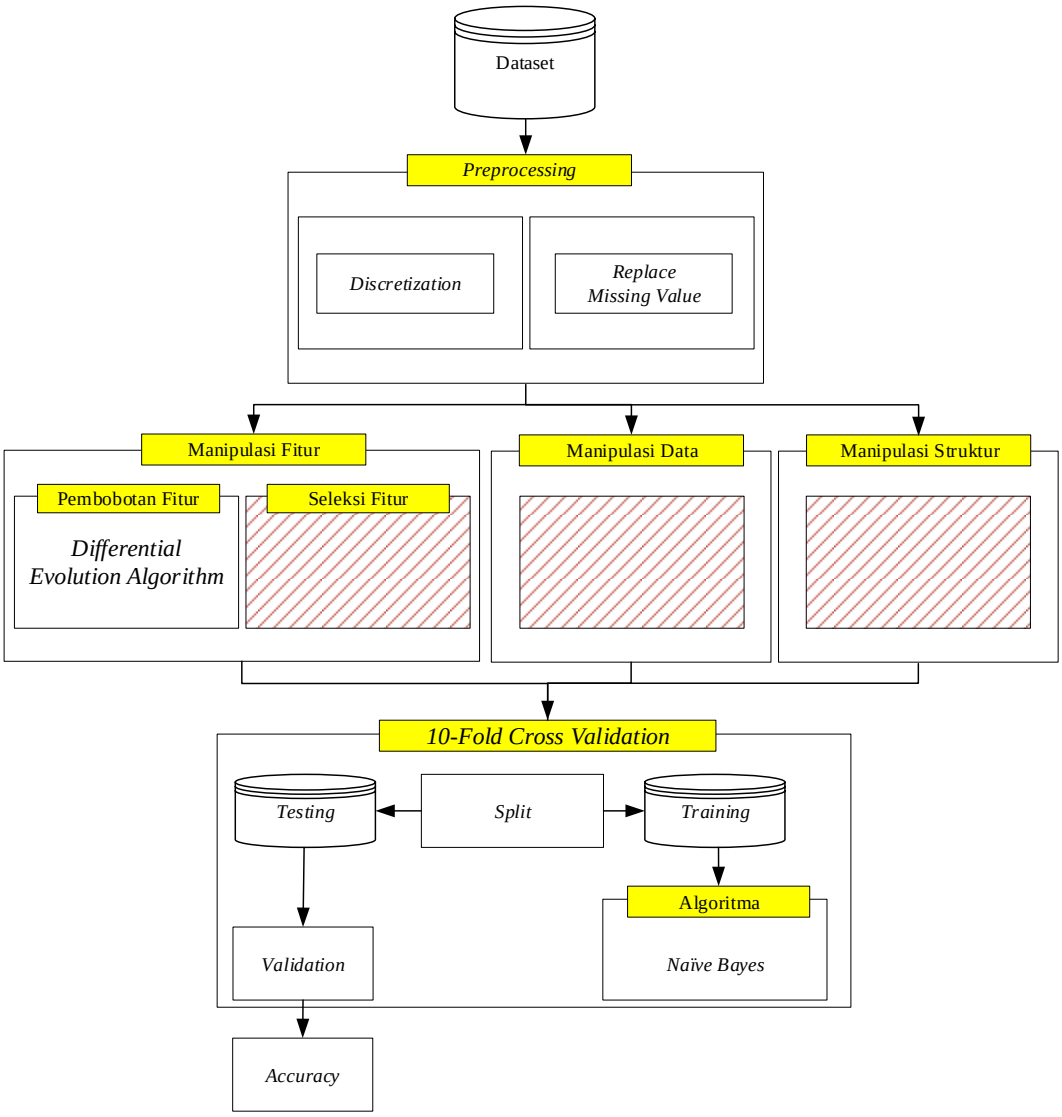


Gambar 2.11 Metode Taheri, Yearwood, Mammadov dan Seifollahi [5]

2.6.4 Metode Wu dan Cai

Wu dan Cai pada tahun 2011 [17] melakukan penelitian untuk memperbaiki kinerja *Naïve Bayes* dengan melemahkan asumsi independensi fitur. Wu dan Cai melakukan metode pendekatan berdasarkan manipulasi fitur yang menggunakan algoritma *Differntial Evolution* yang memiliki performa lebih baik daripada algoritma *evolusioner* yang lain untuk menentukan bobot fitur dan menggunakan bobot ini diperhitungan *Naïve Bayes*. Penelitian ini menggunakan 36 dataset dari UCI *repository*.

Tahapan pertama penelitian ini mengganti semua nilai fitur yang hilang menggunakan *replace missing value* dan mengubah nilai numerik menjadi diskrit menggunakan *discretization* dengan metode *sub optimal agglomerative clustering.* Eksperimen ini melakukan perbandingan antara DE-WNB (*Differntial Evolution*), *Tree*-WNB (menggunakan pembobotan dengan pohon keputusan), MI-WNB (menggunakan pembobotan dengan metode *mutual information*) dan CFS-WNB (pembobotan menngunakan *correlation based feature selection*). Untuk mengukur performa dari metode yang diusulkan, pada penelitian tersebut menggunakan alat ukur *accuracy* dan metode validasi yang digunakan adalah 10 *fold cross validation.* Dari hasil eksperimen metode DE-WNB memberikan hasil yang signifikan dari pada metode yang lain. Bagan metode penelitian ini bisa dilihat pada Gambar 2.12

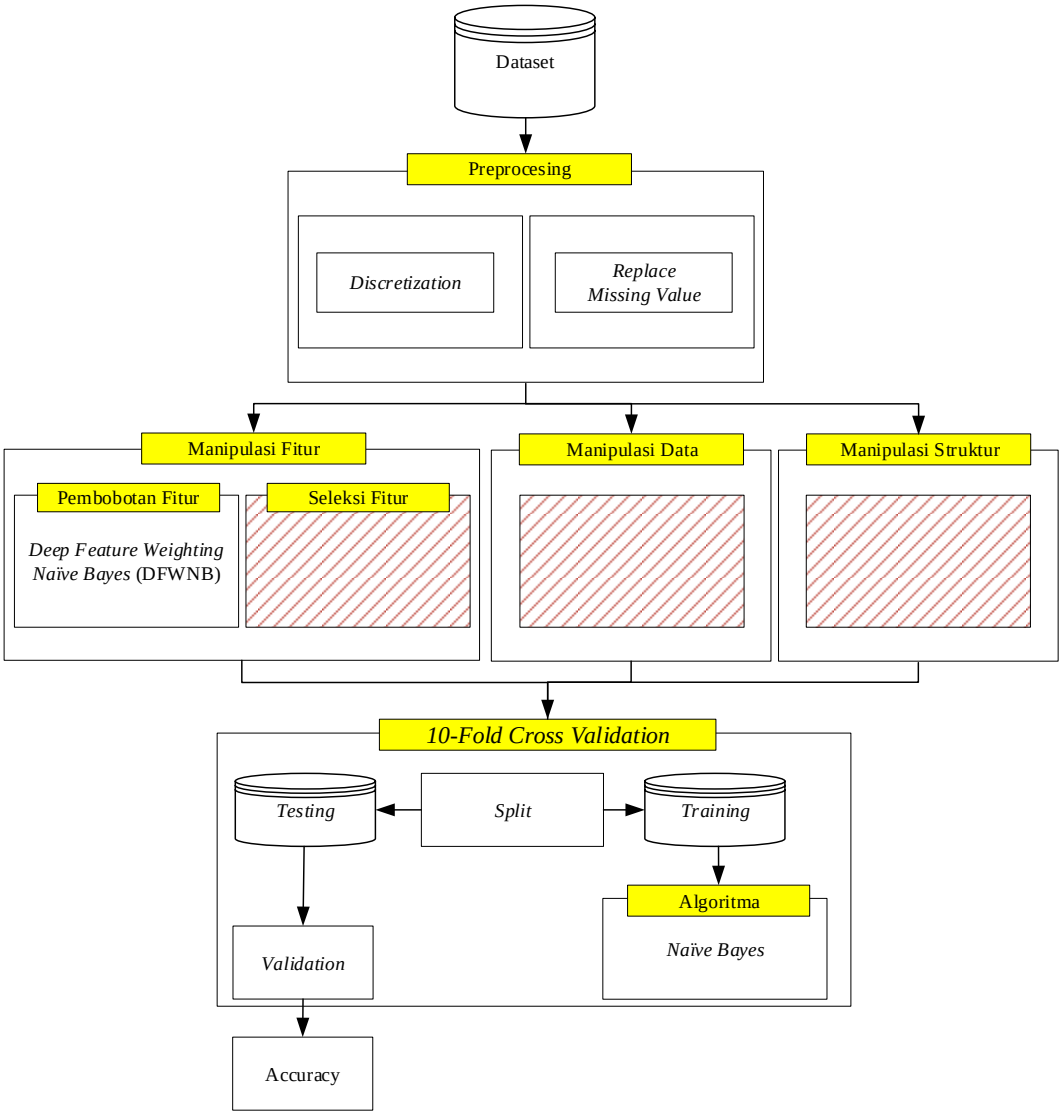


Gambar 2.12 Metode Wu dan Cai [17]

2.6.5 Metode Jiang, Li, Wang dan Zhang

Jiang, Li, Wang dan Zang yang pada tahun 2016 [8] melakukan penelitian untuk memperbaiki kinerja *Naïve Bayes* dengan melemahkan asumsi independensi fitur. Jiang, Li, Wang dan Zang melakukan metode pendekatan berdasarkan manipulasi fitur yang menggunakan algoritma *Deep Feature Weighted Naïve Bayes* yang memiliki kinerja lebih baik sederhana, efisien dan efektif untuk menentukan bobot fitur dan memasukan bobot ini diperhitungan perkiraan probabilitas *Naïve Bayes*. Penelitian ini menggunakan 36 dataset dari UCI *repository* dan secara manual menghapus tiga fitur yang tidak berguna di dataset "colic.ORIG", "splice", dan "zoo".

Tahapan pertama penelitian ini mengganti semua nilai fitur yang hilang menggunakan *replace missing value* dan mengubah nilai numerik menjadi diskrit menggunakan *discretization* dengan metode fayyad & irani's. Eksperimen ini melakukan perbandingan antara *Deep Feature Weighted Naïve Bayes* (DFW), *Naïve Bayes* standar, pembobotan menggunakan *Gain Ratio* (GRFWNB), pembobotan fitur menggunakan *Decision Tree* (DTFWNB), pembobotan fitur menggunakan ReliefF (RFWNB) dan pembobotan menggunakan *correlation based feature selection* (CFS-WNB). Untuk mengukur performa dari metode yang diusulkan, pada penelitian tersebut menggunakan alat ukur *accuracy* dan metode validasi yang digunakan adalah 10 *fold cross validation.* Dari hasil eksperimen metode *Deep Feature Weighted Naïve Bayes* (DFW), memberikan hasil yang mencapai peningkatan yang luar biasa dari pada metode yang lain. Bagan metode penelitian ini bisa dilihat pada Gambar 2.13

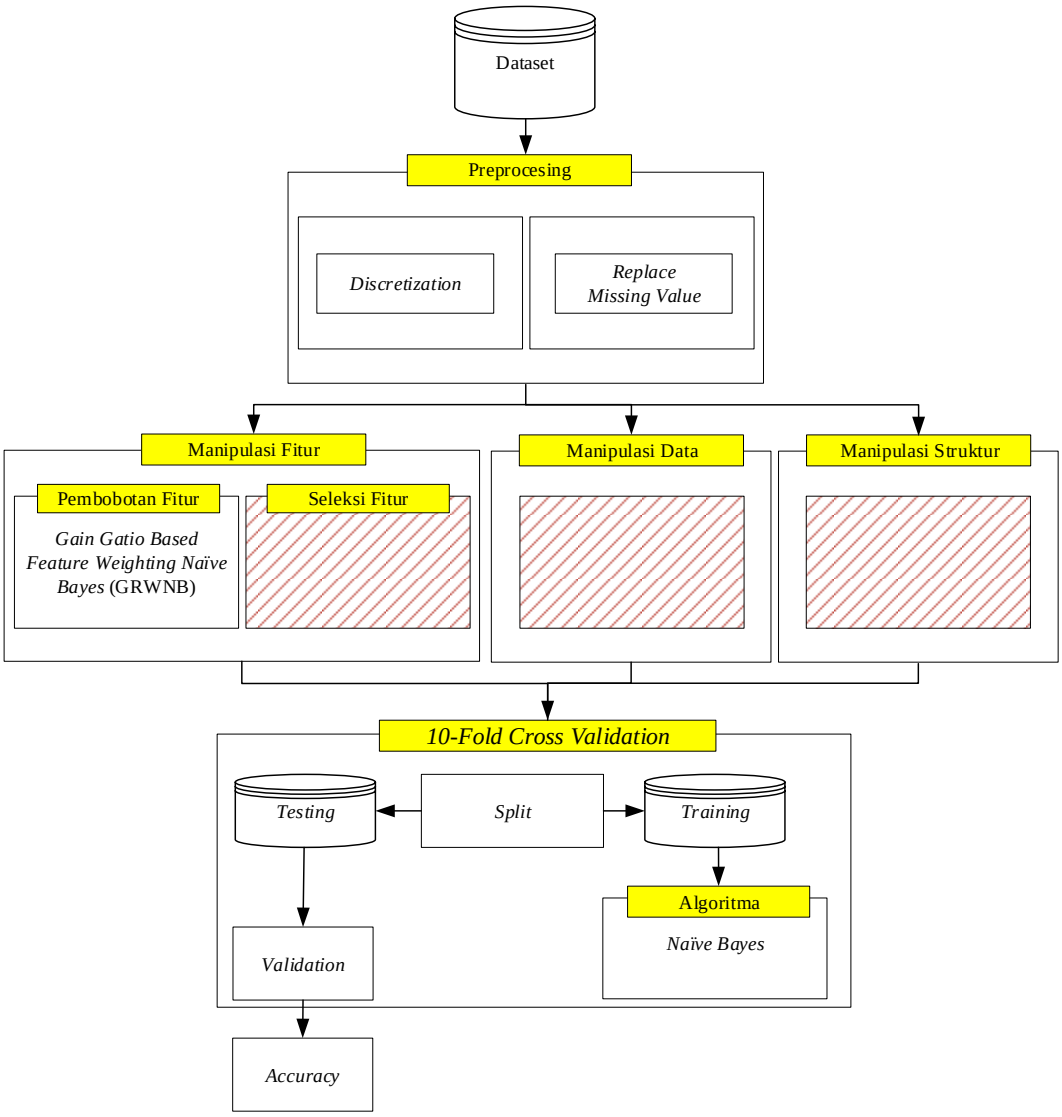


Gambar 2.13 Metode Jiang, Li, Wang dan Zang [8]

2.6.6 Metode Zhang, Jiang, Li dan Kong

Pada tahun 2016, Lungan Zhang**,** Jiang, Li dan Kong [18] melakukan penelitian tentang masalah asumsi independensi fitur pada algoritma *Naïve Bayes* dengan melakukan metode pendekatan berdasarkan pembobotan fitur menggunakan gain ratio kemudian proses *preprocessing discretization* dan *replace missing value* terlebih dahulu selanjutnya menerapkan metode *Gain Ratio Based Feature Weighting* (GRWNB) untuk mengatasi asumsi independensi fitur pada algoritma Naïve Bayes. Metode ini memberikan nilai setiap fitur dengan nilai nol atau bilangan bulat positif. Selain itu penelitian ini mengasumsikan bahwa semua fitur hanya memiliki dua nilai nol dan tidak nol. Bagan metode penelitian ini bisa dilihat pada Gambar 2.14

Penelitian ini membandingkan metode *Gain Ratio Based Feature Weighting* (GRWNB), pembobotan fitur dengan *Decision Tree*, *correlation* berdasarkan *feature selection* (CFS) dan menggunakan 15 dataset dari UCI *repository*. Untuk mengukur performa dari metode yang diusulkan, pada penelitian ini menggunakan alat ukur *Accuracy, Elapsed Test Time, Elapsed Training Time* dan metode validasi yang digunakan adalah 10-*fold cross validation.* Pada penelitian ini menunjukkan akurasi klasifikasi yang lebih tinggi dengan tetap mempertahankan kesederhanaan dan efesien.



Gambar 2.14 Zhang, Jiang, Li dan Kong [18]

Tabel 2.5 Rangkuman Metode yang Pernah Diusulkan

Peneliti	*Prepocessing*	Metode pendekatan				Algoritma	Evaluasi	Validasi
		Metode manipulasi pembobotan fitur	Metode manipulasi seleksi fitur	Metode manipulasi data	Metode manipulasi struktur			
Kannan dan Ramaraj [26]	*Replace Missing Value* dan *Discretization*	-	*Memetic algorithm* (MA-C)	-	-	*Naïve Bayes*	*Accuracy*	*10-fold cross validation*
Deisy *et al* [25]	*Replace Missing Value* dan *Discretization*	-	*Information theoretic-interact* (IT-IN)	-	-	*Naïve Bayes*, SVM dan ELM	*Accuracy*	*10-fold cross validation*
Jiang,Cai dan Wang [7]	*Replace Missing Value* dan *Discretization*	-	-	*Instance Weighted Naive Bayes* (IWNB) *and Combined Neighbourhood Naive Bayes* (CNNB)	-	*Naïve Bayes*	*Accuracy*	*10-fold cross validation*
Chaoqun dan Li [29]	-	-	-	-	*One Dependence Value Difference Metric* (ODVDM)	*Naïve Bayes*	*Accuracy*	*10-fold cross validation*

Peneliti	*Prepocessing*	Metode pendekatan				Algoritma	Evaluasi	Validasi
		Metode manipulasi pembobotan fitur	Metode manipulasi seleksi fitur	Metode manipulasi data	Metode manipulasi struktur			
Lin dan Yu [31]	-	*Naive Bayes Classification Algorithm Based on* PSO *(particle swarm optimization)* PSOWNBC	-	-	-	*Naïve Bayes*	*Accuracy*	*Leave One Out Cross Validation*
Jia, Wu dan Cai [17]	*Replace Missing Value* dan *Discretization*	*Differential Evolution Algorithm for Attribute Weighted Naive Bayes* (DEWNB)	-	-	-	*Naïve Bayes*	*Accuracy*	*10-fold cross-validation*
Lee, Gutierrez dan Dou [16]	*Discretization*	*Kullback-Leibler measure*	-	-	-	*Naïve Bayes*	*Accuracy*	*10-fold cross-validation*
Jiang [13]	*Replace Missing Value* dan *Discretization*	-	-	-	*Random One Dependence Estimators* (RODE)	*Naïve Bayes*	*Accuracy, Averaged Conditional Log Likelihood, Area Under Curve*	*10-fold cross validation*
Jiang *et al* [32]	*Replace Missing Value* dan *Discretization*	-	-	-	*Weighted Average of One Dependence*	*Naïve Bayes*	*Accuracy*	*10-fold cross validation*

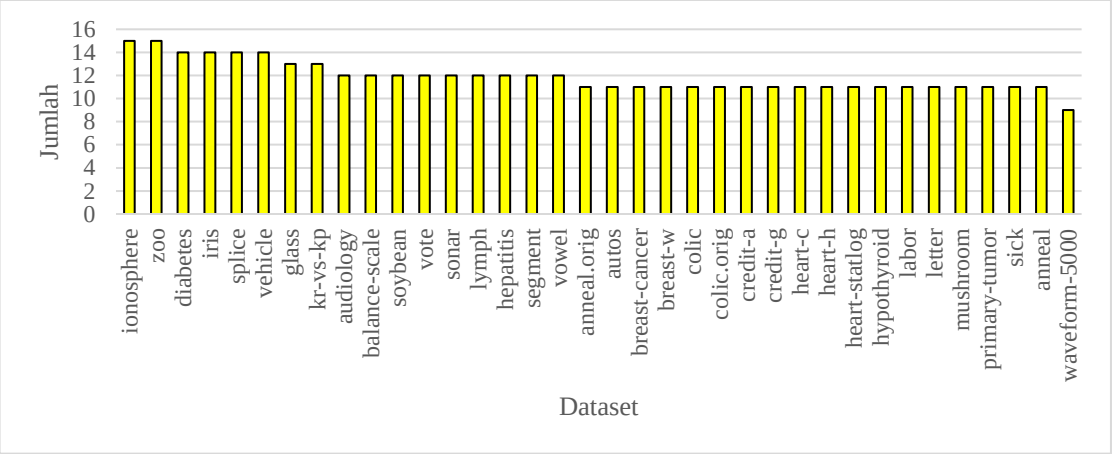
Peneliti	*Prepocessing*	Metode pendekatan				Algoritma	Evaluasi	Validasi
		Metode manipulasi pembobotan fitur	Metode manipulasi seleksi fitur	Metode manipulasi data	Metode manipulasi struktur			
					Estimators (WAODE)			
Jiang *et al* [30]	*Replace Missing Value* dan *Discretization*	*Randomly Selected Naive Bayes* (RSNB)	-	-	-	*Naïve Bayes*	*accuracy* (ACC), *area under curve* (AUC), *conditional log likelihood* (CLL)	*5 x 5-fold cross-validation*
Jiang *et al* [15]	*Replace Missing Value* dan *Discretization*	-	-	-	*Tree Augmented Naive Bayes* (TAN)	*Naïve Bayes*	*Conditional log likelihood*	*10-fold cross-validation*
Zaidi *et al* [24]	*Replace Missing Value* dan *Discretization*	*Weighted Naive Bayes Algorithm* (WANBIA)	-	-	-	*Naïve Bayes*	*Conditional log likelihood*, MSE	*2-fold cross-validation*
Jiang *et al* [28]	-	-	-	*Locally Weighted Naive Bayes Text Classifiers* (LWNBTC)	-	*Naïve Bayes*	*Accuracy*	*5-fold cross-validation*
Farid *et al* [27]	-	-	-	*Hybrid Naïve Bayes*	-	*Naïve Bayes*	*Accuracy*	*10-fold cross-validation*

Peneliti	*Prepocessing*	Metode pendekatan				Algoritma	Evaluasi	Validasi
		Metode manipulasi pembobotan fitur	Metode manipulasi seleksi fitur	Metode manipulasi data	Metode manipulasi struktur			
Taheri *et al* [5]	*Replace Missing Value* dan *Discretization*	*Attribute Weighted Naive Bayes* (AWNB)	-	-	-	*Naïve Bayes*	*Accuracy*	*10-fold cross-validation*
Wu *et al* [33]	*Replace Missing Value* dan *Discretization*	*Dual Weighted Model* (DWNB)	-	-	-	*Naïve Bayes*	*Accuracy*	*10-fold cross-validation*
Wu *et al* [23]	*Replace Missing Value* dan *Discretization*	*Artificial Immune System Based Weighted Naive Bayes* (AISWNB)	-	-	-	*Naïve Bayes*	*Accuracy*	*10-fold cross validation*
Wu *et al* [6]	*Replace Missing Value* dan *Discretization*	*Artificial Immune System weighting Naïve Bayes* (AISWNB) *Based Self-Adaptive*	-	-	-	*Naïve Bayes*	*Accuracy, Averaged Conditional Log Likelihood, Area Under Curve*	*10-fold cross-validation*
Jiang *et al* [14]	-	-	-	-	*Strukture Extended Multinomial Naive Bayes* (SEMNB)	*Naïve Bayes*	*Accuracy, Elapsed Test Time, Elapsed Training Time*	*10-fold cross-validation*

Peneliti	*Prepocessing*	Metode pendekatan				Algoritma	Evaluasi	Validasi
		Metode manipulasi pembobotan fitur	Metode manipulasi seleksi fitur	Metode manipulasi data	Metode manipulasi struktur			
Lee [34]	*Replace Missing Value* dan *Discretization*	*Value Weighting Naïve Bayes* (VWNB)	-	-	-	*Naïve Bayes*	*Accuracy*	*10-fold cross validation*
Wang, Jiang dan Li [35]	-	-	-	-	*Multinomial Naive Bayes Tree* (MNBTree)	*Naïve Bayes*	*Accuracy*	*10-fold cross validation*
Zhang *et al* [18]	-	*Gain Gatio Based Feature Weighting*	-	-	-	*Naïve Bayes*	*Accuracy, Elapsed Test Time, Elapsed Training Time*	*10-fold cross validation*
Wu *et al.*[36]	*Replace Missing Value* dan *Discretization*	-	*Self-Adaptive* SPODEs (SODE)	-	-	*Naïve Bayes*	*Accuracy, Elapsed Test Time, Elapsed Training Time*	*10-fold cross validation*
Jiang *et al.*[8]	*Replace Missing Value* dan *Discretization*	*Deep Feature Weighting Naïve Bayes* (DFWNB)	-	-	-	*Naïve Bayes*	*Accuracy* dan *kappa*	*10-fold cross validation*

2.7 Dataset yang Sering Digunakan

Pada Gambar 2.15 dapat dilihat jumlah penggunaan dataset ionosphere dan zoo yang sering dipakai dalam penelitian dengan asumsi independensi fitur pada algoritma *Naïve Bayes*. Selama kurun waktu tahun 2010 hingga 2016 sebanyak 100% peneliti menggunakan dataset *public*.



Gambar 2.15 Dataset yang Sering Digunakan

2.8 Daftar Referensi *Systematic Literature Review*

Daftar referensi dari *systematic literature review* ini dapat dilihat pada Tabel 2.6 daftar referensi terdiri dari 4 atribut (tahun, peneliti, metode dan publikasi) dan terdapat 24 publikasi penelitian dari tahun 2010 hingga 2016, dan pada Tabel 2.6 tersebut telah diurutkan berdasarkan tahun publikasi.

Tabel 2.6 Daftar Referensi *Systematic Literature Review*

Tahun	Peneliti	Metode	Publikasi
2010	Kannan dan Ramaraj	*Memetic algorithm*(MA-C)	Knowledge Based Systems
	Deisy *et al*	*Information theoretic interact* (IT-IN)	Expert Systems with Applications journal
	Jiang, Cai dan Wang	*Instance Weighted Naive Bayes* (IWNB) *and Combined Neighbourhood Naive Bayes* (CNNB)	International Journal of Computers and Applications
2011	Chaoqun dan Li	*One Dependence Value Difference Metric* (ODVDM)	Knowledge Based Systems
	Lin dan Yu	*Naive Bayes Classification Algorithm Based on* PSO *(particle swarm optimization)* PSOWNBC	Communication Software and Networks (ICCSN)
	Wu dan Cai	*Differential Evolution Algorithm for Attribute Weighted Naive Bayes* (DEWNB)	Computational Information Systems
	Lee, Gutierrez dan Dou	*Kullback-Leibler measure.*	Data Mining (ICDM)
	Jiang	*Random one dependence estimators* (RODE)	Pattern Recognition Letters journal
	Jiang *et al*	*Weighted Average of One Dependence Estimators* (WAODE)	Experimental & Theoretical Artificial Intelligence
2012	Jiang *et al*	*Randomly Selected Naive Bayes* (RSNB)	Expert Systems with Applications
	Jiang *et al*	*Tree Augmented Naive Bayes* (TAN)	Knowledge Based Systems
2013	Zaidi *et al*	*Weighted Naive Bayes Algorithm* (WANBIA)	Machine Learning Research
	Jiang *et al*	*Locally Weighted Naive Bayes Text Classifiers* (LWNBTC)	Experimental & Theoretical Artificial Intelligence
2014	Farid *et al*	*Hybrid Naïve Bayes*	Expert Systems with Applications
	Taheri *et al*	*Attribute Weighted Naive Bayes* (AWNB)	Neural Comput & Applic

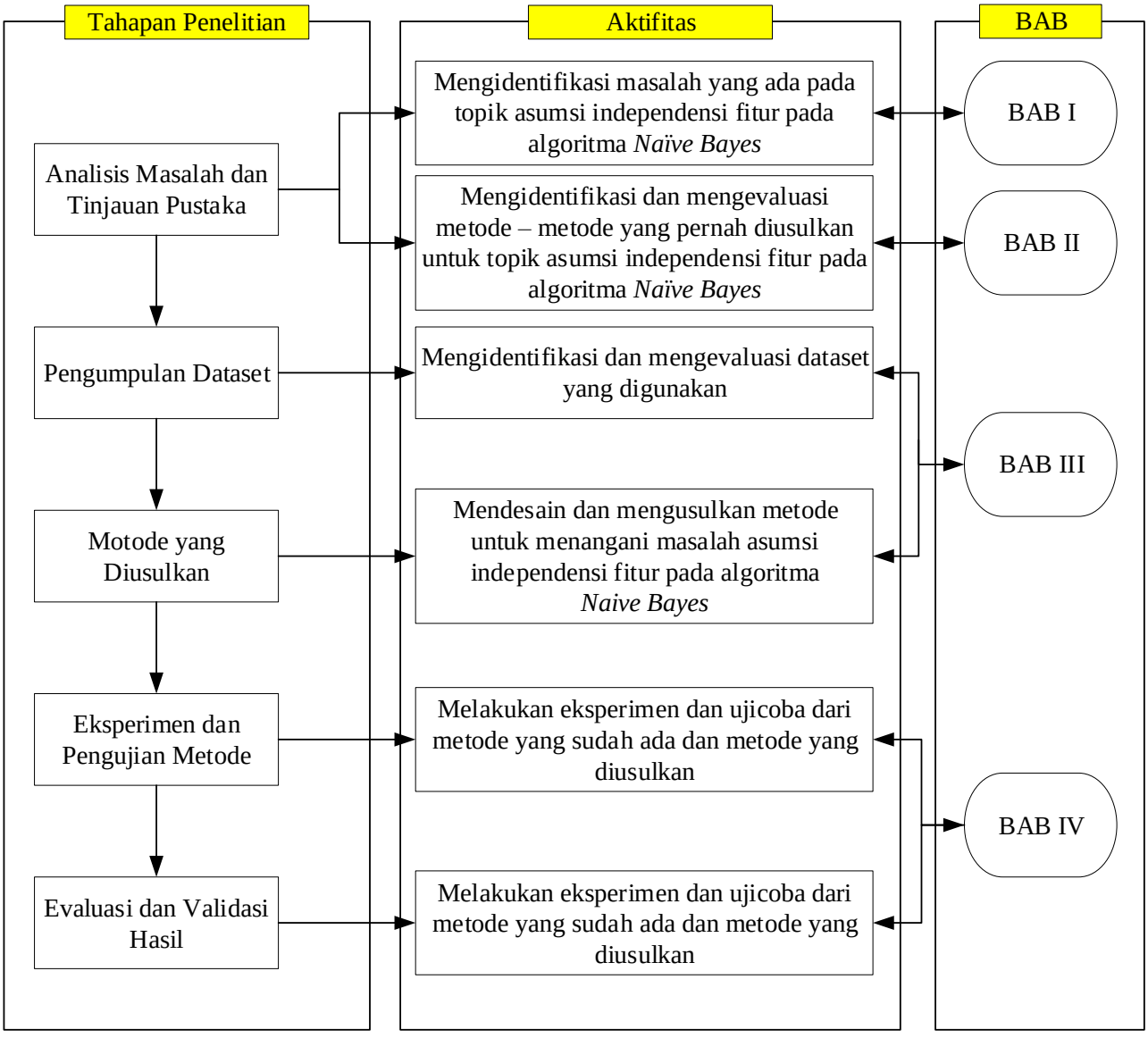
Tahun	Peneliti	Metode	Publikasi
	Wu, Pan, Zhu dan Cai	*Dual Weighted Model* (DWNB)	International Joint Conference on Neural Networks
2015	Wu, Cai, Zeng, dan Zhu	*Artificial Immune System based Weighted Naive Bayes* (AISWNB)	Knowledge Based Systems
	Wu, Pan, Zhu dan Cai	*Artificial Immune System Weighting Naïve Bayes* (AISWNB) *Based Self Adaptive*	Expert Systems with Applications
	Jiang, Wang, dan Zhang	*Structure Extended Multinomial Naive Bayes* (SEMNB)	Information Sciences
	Lee	*Value Weighting Naïve Bayes* (VWNB)	Knowledge Based Systems
	Wang, Jiang dan Li	*Multinomial Naive Bayes Tree* (MNBTree)	Knowledge and Information Systems
2016	Zhang *et al*	*Gain Gatio Based Feature Weighting*	Knowledge Based Systems
	Wu *et al.*	*Self Adaptive* SPODEs (SODE)	Pattern Recognition
	Jiang *et al.*	*Deep Feature Weighting Naïve Bayes* (DFWNB)	Engineering Applications of Artificial Intelligence

BAB 3

METODE PENELITIAN

3.1 Tahapan Penelitian

Pada penelitian ini menggunakan metode penelitian eksperimen. Metode penelitian eksperimen adalah uji coba yang dilakukan oleh peneliti sendiri untuk melakukan investigasi hubungan sebab akibat. Tahapan pada penelitian ini dapat dilihat pada Gambar 3.1, yang berisi sebagai berikut:



Gambar 3.1 Tahapan Penelitian, Aktivitas, dan Relasi Bab

3.2 Analisis Masalah dan Tinjauan Pustaka

Pada Gambar 3.1 dapat dilihat bahwa pada penelitian ini akan mencoba untuk mereview beberapa publikasi penelitian sebagai langkah awal. Metode review yang digunakan adalah metode *Systematic Literature Review* seperti yang diusulkan oleh Kitchenham [22]. *Systematic Literature Review* merupakan sebuah proses untuk identifikasi, menilai, dan menginterpretasikan dari semua penelitian dengan tujuan untuk menjawab dari pertanyaan penelitian tertentu. Metode-metode yang sudah pernah diusulkan untuk topik asumsi independensi fitur pada algoritma *Naïve Bayes* juga diidentifikasi berdasarkan *Systematic Literature Review*. Hasil dari *Systematic Literature Review* untuk topik asumsi independensi fitur pada algoritma *Naïve Bayes* telah dijelaskan pada Bab 2.

Hasil dari *Systematic Literature Review* pada Bab 2 dirangkum metode metode yang sudah pernah diusulkan oleh para peneliti untuk topik data dengan asumsi independensi fitur pada algoritma *Naïve Bayes*. Dari rangkuman tersebut dapat diidentifikasi kelebihan dan kekurangan pada metode yang sudah pernah diusulkan yang pada akhirnya dijadikan landasan permasalahan dan menjadi dasar untuk membuat metode usulan yang akan digunakan pada penelitian ini.

3.3 Pengumpulan Dataset

Dataset yang yang umum digunakan untuk asumsi independensi fitur pada algoritma *Naïve Bayes* adalah dataset dari UCI *Repository* dari tahun 2010 hingga 2016 terdapat 36 dataset yang sering digunakan peneliti untuk mengukur kemampuan kinerja metode yang diusulkannya. Adapun karakteristik dari dataset yang digunakan dalam penelitian ini menampilkan nama dataset, data, jumlah fitur, *missing* (Y/N), tipe data numerik Tabel 3.1. Dataset didapatkan dari UCI *machine learning repository*. Dalam penelitian ini menerapkan *unsupervised attribute filter replace missing value* untuk mengisi *missing value* yang terdapat didalam dataset dengan menggunakan software WEKA.

Tabel 3.1 Karakteristik Dari Dataset

Dataset	Data	Fitur	Kelas	*Missing*	Numerik
Ionosphere	351	35	2	N	Y
Zoo	101	18	7	N	Y
Glass	214	10	7	N	Y
Balance scale	625	5	3	N	Y
Sonar	208	61	2	N	Y
Lymph	148	19	4	N	Y
Hepatitis	155	20	2	Y	Y
Segment	2310	20	7	N	Y
Vowel	990	14	11	N	Y
Anneal.orig	898	39	6	Y	Y
Autos	205	26	7	Y	Y
Colic	368	23	2	Y	Y
Colic.orig	368	28	2	Y	Y
Credit-a	690	16	2	Y	Y
Credit-g	1000	21	2	N	Y
Hypothyroid	3772	30	4	Y	Y
Labor	57	17	2	Y	Y
Letter	20000	17	26	N	Y
Sick	3772	30	2	Y	Y
Anneal	898	39	6	Y	Y
Waveform-5000	1000	41	3	N	Y

3.4 Metode yang Diusulkan

Pada penelitian ini akan diusulkan metode dengan cara manipulasi fitur *weighting*, yaitu dengan menggunakan *self organizing map* yang mampu mendapatkan bobot yang tepat dengan mencari bobot yang paling mendekati dengan *euclidean distance* dan meng*update* hasil pembobotan sampai mendapatkan bobot yang tepat sesuai dengan nilai iterasi yang ditetapkan dengan mengurangi laju pembelajaran (α)[20]. Alur metode yang diusulkan dapat dilihat pada Gambar 3.2 adalah sebagai berikut:

1. Menyiapkan dataset
2. Data yang akan digunakan jika terdapat *missing value* harus menerapkan *unsupervised attribute filter replace missing values* untuk mengisi nilai yang kosong

3. Tentukan *decay rate*, minimal *learning rate* dan *learning rate* (faktor kali yang digunakan untuk mereduksi selisih antara data dan bobot pada saat proses *update* bobot) disimbolkan dengan α (*alpha*).
4. Menentukan inisialisasi bobot awal dengan nilai acak antara 0 sampai dengan 1 disetiap fitur sejumlah kelas (C)
5. Tentukan inisialisasi *node* bobot tiap (c) dengan rumus

$$d^2 = (Euclidean\ distance)^2 = \sum_{k=1}^{n} \left(i_{l,k} - w_{j,k}\,(t)\right)^2 \quad (3.1)$$

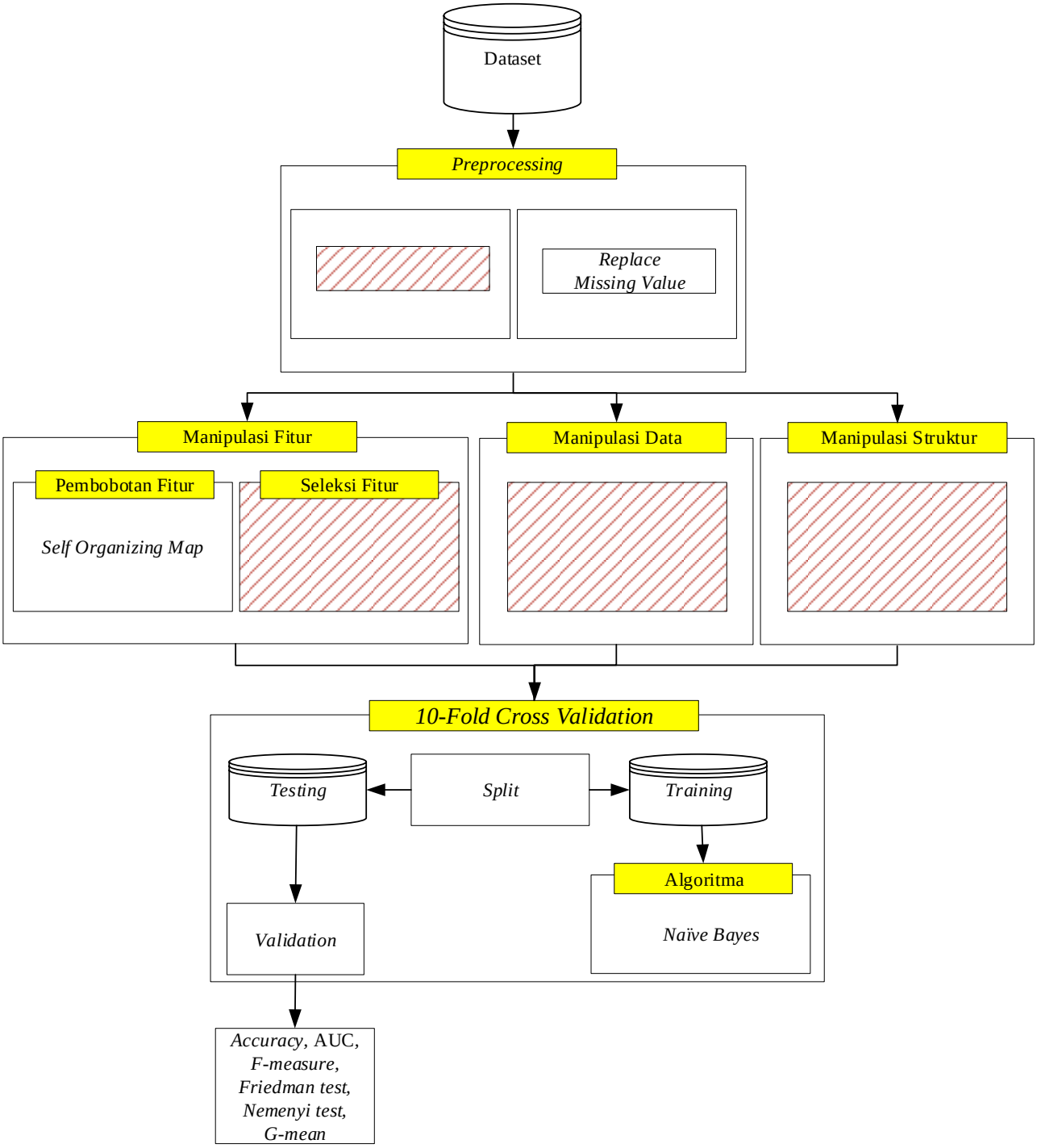
6. Cari jarak *euclidean* yang tebaik yang didapat dari hasil perhitungan proses ke 5
7. Kemudian *update* bobot sesuai jarak euclidean yang terbaik dengan rumus

$$Weight\ update: w_j(t) + \eta(t)\,(i_l - w_j(t)) \quad (3.2)$$

8. Mengurangi *learning rate* dengan rumus

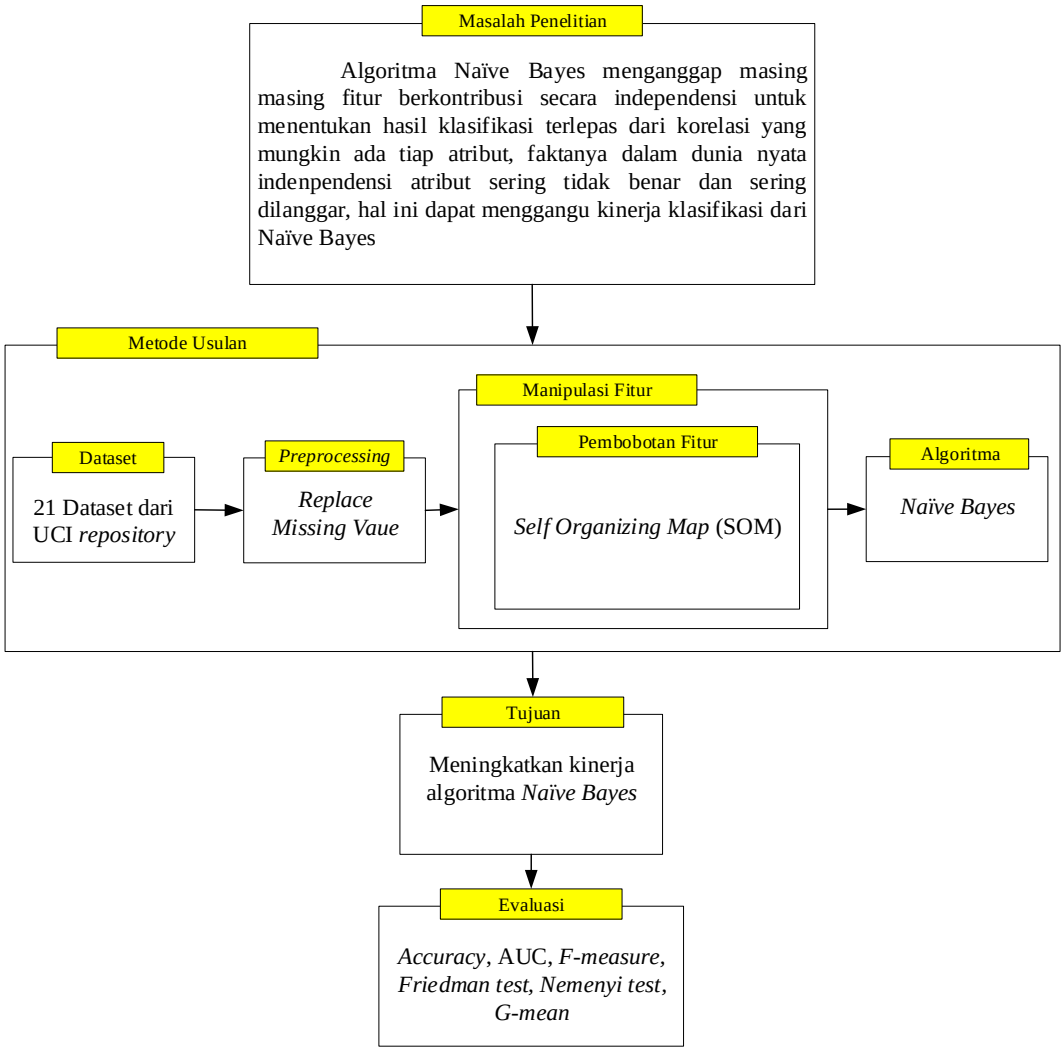
$$\alpha\ (alpha) = decay\ rate * \alpha\ (alpha) \quad (3.3)$$

9. Ulangi proses ke 5 hingga nilai *learning rate* kurang dari atau sama dengan minimal *learning rate*
10. *Update* bobot terakhir dihitung dengan data dan didapatkan bobot terbaik yang akan digunakan dalam perhitungan *Naïve Bayes*
11. Hitung *accuracy*, AUC, *F-measure, friedman test, nemenyi test* dan *G-mean*



Gambar 3.2 Metode Pembobotan *Self Organizing Map* (SOMWNB)

Kerangka pemikiran pada penelitian ini dapat dilihat pada Gambar 3.3 yang terdiri dari beberapa tahap. Masalah dalam penelitian ini adalah pada algoritma *Naïve Bayes* menganggap masing masing fitur berkontribusi secara independensi untuk menentukan hasil klasifikasi terlepas dari korelasi yang mungkin ada tiap fitur, faktanya dalam dunia nyata indenpendensi fitur sering tidak benar dan sering dilanggar, hal ini dapat menggangu kinerja klasifikasi dari *Naïve Bayes*. Untuk mengurangi asumsi independen ini akan diberikan bobot tiap fitur. Tujuan dari penelitian ini adalah untuk meningkatkan kinerja algoritma *Naïve Bayes* untuk mengatasi asumsi independen dalam perhitungannya. Untuk mengevaluasi kinerja dari metode yang diusulkan, digunakan *accuracy*, AUC, *F-measure*, *friedman test*, *nemenyi test* dan *G-mean*.



Gambar 3.3 Kerangka Pemikiran Penelitian

3.5 Eksperimen dan Pengujian Metode

Pada penelitian ini menggunakan beberapa metode sebagai pembanding antara lain metode *Deep Feature Weighted Naïve Bayes* dan *Gain Ratio Weighted Naïve Bayes* (GRWNB) di tahun 2016, metode *Attribute Weighted Naïve Bayes Using Mutual Information* di tahun 2014, *Differential Evolution Algorithm for Attribute Weighted Naive Bayes* (DEWNB) di tahun 2011

Tahapan eksperimen dalam penelitian ini adalah sebagai berikut:

1. Menyiapkan dataset yang akan digunakan
2. Mengukur performa metode *Deep Feature Weighted Naïve Bayes* [8]
3. Mengukur performa *Gain Ratio Weighted Naïve Bayes* (GRWNB) [18]
4. Mengukur performa *Attribute Weighted Naïve Bayes Using Mutual Information* [32]
5. Mengukur performa dari *Naïve Bayes* Standar [37]
6. Mengukur performa *Differential Evolution Algorithm For Attribute Weighted Naive Bayes* (DEWNB) [17]
7. Mengukur performa *Self Organizing Map Weighted Naïve Bayes* (SOMWNB)
8. Membandingkan hasil metode menggunakan *accuracy, AUC,* F-*measure, friedman test, nemenyi test* dan *G-mean.*

Dalam penelitian ini akan menggunakan alat bantu software berupa Microsoft Office Excel 2016, WEKA 3.8, Dev-C++. Spesifikasi komputer yang digunakan untuk melakukan penelitian dapat dilihat pada Tabel 3.2.

Tabel 3.2 Spesifikasi Komputer yang Digunakan

Prosesor	AMD A10-7400P, 10 Compute Core 4C+6G 2.50 GHz
Memori	8 GB
Harddisk	1 TB
Sistem Operasi	Windows 10 Home 64bit
Aplikasi	Microsoft Office Excel 2016, WEKA 3.8, Dev-C++

3.6 Evaluasi dan Validasi Hasil

Evaluasi dan validasi hasil dari eksperimen merupakan sebuah alat ukur yang dapat digunakan untuk menilai atau mengukur seberapa baik metode yang

diusulkan terhadap metode lainnya dan apakah metode yang diusulkan mempunyai perbedaan hasil yang signifikan terhadap metode lainnya.

3.6.1 Evaluasi Hasil

3.6.1.1 *Accuracy*

Accuracy merupakan salah satu indikator yang dipakai dalam penelitian ini untuk mengukur performa dari metode yang diusulkan. Nilai *accuracy* dihitung dengan mengambil persentase prediksi yang benar dari keseluruhan data [3]. Prediksi yang tepat berarti hasil di mana kelas hasil prediksi adalah sama dengan kelas dari data. Pada Tabel 3.3 dapat dilihat perhitungan dari nilai *accuracy*.

Tabel 3.3 Perhitungan *Accuracy*

Kelas	Kenyataan Benar	Kenyataan Salah
Prediksi Benar	TP (*True Positive*)	FP (*False Positive*)
Prediksi Salah	FN (*False Negative*)	TN (*True Negative*)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (3.4)$$

3.6.1.2 *Area Under Curve* (AUC)

Selain *accuracy*, dalam penelitian ini juga menggunakan indikator AUC untuk mengukur performa dari metode yang diusulkan. AUC atau area dibawah kurva merupakan suatu kurva yang menggambarkan probabilitas dengan variabel sensitivitas dan kekhususan (*specificity*) dengan nilai batas antara 0 hingga 1. Area dibawah kurva memberikan gambaran tentang keseluruhan pengukuran atas kesesuaian dari model yang digunakan. Model dengan akurasi sempurna akan memiliki luas 1[3]. Pada umumnya, algoritma yang memiliki nilai AUC diatas 0.6, mempunyai performa yang cukup efektif untuk mengatasi asumsi indenpendensi fitur di *Naïve Bayes*. Pada Tabel 3.4 dapat dilihat interpretasi dari masing masing nilai AUC.

Tabel 3.4 Nilai AUC dan Interpretasinya

Nilai AUC	Interpretasi
0.90 - 1.00	Klasifikasi sempurna
0.80 - 0.90	Klasifikasi baik
0.70 - 0.80	Klasifikasi cukup
0.60 – 0.70	Klasifikasi rendah
< 0.60	Klasifikasi jelek

3.6.1.3 F-*measure*

F-*measure* merupakan sebuah alat ukur untuk menguji akurasi. F-*measure* memperhitungkan *precision* dan *recall* dari sebuah pengujian untuk mendapatkan nilainya [38]. *Precision* merupakan jumlah prediksi yang benar dibagi dengan total yang diprediksi benar. Sedangkan *recall* merupakan jumlah prediksi yang benar dibagi total yang seharusnya benar.

$$f - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{3.5}$$

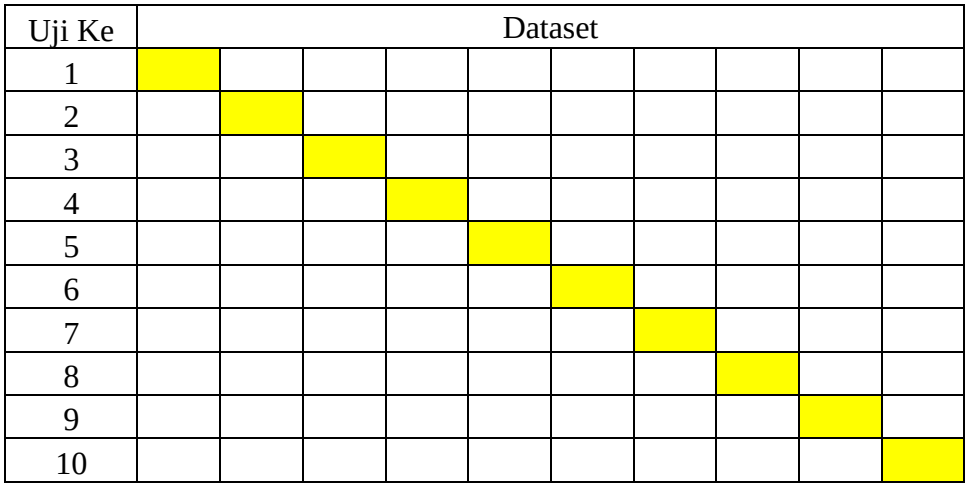
3.6.2 Validasi Hasil

Pada penelitian ini akan menggunakan metode validasi *10-fold cross validation.* Metode validasi ini akan membagi dataset menjadi 10 bagian yang sama dan proses pembelajaran akan berjalan 10 kali (

Uji Ke	Dataset									
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

Gambar 3.4). Setiap kali proses pembelajaran berlangsung, 1 bagian akan menjadi dataset *testing*, dan 9 bagian lainnya akan menjadi data *training*. Lalu nilai

rata-rata dan nilai deviasi hasil dari proses 10 kali pembelajaran akan dikalkulasi. Pada penelitian ini menggunakan *10-fold cross validation* karena sudah menjadi standard dari penelitian akhir akhir ini, dan beberapa penelitian juga didapatkan bahwa penggunaan stratifikasi dapat meningkatkan hasil yang lebih tidak beragam [39].



Gambar 3.4 10-*Fold Cross Validation*

Keterangan:

Kuning = k-subset (Data *testing*)

Putih = Data *training*

Evaluasi hasil metode usulan pada penelitian ini menggunakan *accuracy*, AUC, dan f-*measure*. Nilai *accuracy* dihitung dengan mengambil persentase prediksi yang benar dari keseluruhan data. Prediksi yang tepat berarti hasil di mana kelas hasil prediksi adalah sama dengan kelas dari data. AUC merupakan area dibawah kurva (*area under curve*), suatu kurva yang menggambarkan probabilitas dengan variabel sensitivitas dan kekhususan (*specificity*) dengan nilai batas antara 0 hingga 1. Area dibawah kurva memberikan gambaran tentang keseluruhan pengukuran atas kesesuaian dari model yang digunakan. F-*measure*

memperhitungkan *precision* dan *recall* dari sebuah pengujian untuk mendapatkan nilainya.

Tahap selanjutnya adalah dengan mengukur kinerja dari metode yang diusulkan dengan menggunakan uji *nonparametric*. Sesuai dengan usulan dari Demsar[40], karena penelitian ini menggunakan beberapa dataset dan membandingkan beberapa metode maka pengujian dilakukan dengan menggunakan uji *friedman*. Uji *friedman* digunakan untuk membandingkan n-sampel untuk mengetahui apakah ada perbedaan yang signifikan atau tidak pada metode yang diuji. Tetapi pada uji *friedman* tidak dapat menunjukkan metode mana yang mempunyai perbedaan yang signifikan, maka digunakan uji *post hoc* [41]. Pada penelitian ini akan menggunakan uji *nemenyi post hoc* untuk mengetahui metode mana yang mempunyai perbedaan yang signifikan. Selain itu untuk mendapatkan kecenderungan nilai dari hasil kinerja masing masing metode, pada penelitian ini akan menggunakan *geometric mean* (G-*mean*).

BAB 4

HASIL DAN PEMBAHASAN

(*Sengaja dikosongkan*)

BAB 5
KESIMPULAN

(*Sengaja dikosongkan*)

DAFTAR REFERENSI

[1] C. C. Aggarwal, *Data Mining*. Springer London, 2015.

[2] L. Koc, T. A. Mazzuchi, and S. Sarkani, "A Network Intrusion Detection System Based on a Hidden Naïve Bayes Multiclass Classifier," *Expert Syst. Appl.*, vol. 39, no. 18, pp. 13492–13500, 2012.

[3] J. P. Jiawei Han, Michelin Kamber, *Data Mining Concepts and Techniques*. Elsevier Inc., 2012.

[4] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, "A Systematic Literature Review on Fault Prediction Performance in Software Engineering," *Softw. Eng.*, vol. 38, no. 6, pp. 1276–1304, 2012.

[5] S. Taheri, J. Yearwood, and M. Mammadov, "Attribute Weighted Naive Bayes Classifier Using a Local Optimization," *Neural Comput Applic*, pp. 995–1002, 2014.

[6] J. Wu, S. Pan, X. Zhu, Z. Cai, P. Zhang, and C. Zhang, "Self Adaptive Attribute Weighting for Naive Bayes Classification," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1487–1502, 2015.

[7] L. Jiang, Z. Cai, and D. Wang, "Improving Naive Bayes for Classification," *Int. J. Comput. Appl.*, vol. 7074, no. April, 2016.

[8] L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep Feature Weighting for Naive Bayes and Its Application to Text Classification," *Eng. Appl. Artif. Intell.*, vol. 52, pp. 26–39, 2016.

[9] A. Zakerolhosseini, "Unsupervised Probabilistic Feature Selection using Ant Colony Optimization Optimation," *Expert Syst. Appl.*, 2016.

[10] M. A. Hall, "Feature Selection for Discrete and Numeric Class Machine," *Comput. Sci.*, pp. 1–16, 1999.

[11] H. Zhang, "Learning Weighted Naive Bayes with Accurate Ranking," *Data Min.*, pp. 4–7, 2004.

[12] C. Li, L. Jiang, and H. Li, “Local Value Difference Metric,” *Pattern Recognit. Lett.*, vol. 49, pp. 62–68, 2014.

[13] L. Jiang, “Random One Dependence Estimators,” *Pattern Recognit. Lett.*, vol. 32, no. 3, pp. 532–539, 2011.

[14] L. Jiang, S. Wang, C. Li, and L. Zhang, “Structure Extended Multinomial Naive Bayes,” *Inf. Sci. (Ny).*, 2015.

[15] L. Jiang, Z. Cai, D. Wang, and H. Zhang, “Improving Tree Augmented Naive Bayes for Class Probability Estimation,” *Knowledge-Based Syst.*, vol. 26, pp. 239–245, 2012.

[16] C. Lee, F. Gutierrez, and D. Dou, “Calculating Feature Weights in Naive Bayes with Kullback-Leibler Measure,” *Data Min.*, 2011.

[17] J. Wu and Z. Cai, “Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB),” *Comput. Inf. Syst.*, vol. 5, pp. 1672–1679, 2011.

[18] L. Zhang, L. Jiang, C. Li, and G. Kong, “Two FeatureWeighting Approaches for Naive Bayes Text Classifier,” *Knowledge-Based Syst.*, 2016.

[19] T. Kohonen, “The Self Organizing Map,” *Neurocomputing*, vol. 21, no. May, pp. 1–6, 1998.

[20] S. Shieh and I. Liao, “Expert Systems With Applications A New approach for Data Clustering and Visualization Using Self Organizing Maps,” *Expert Syst. Appl.*, vol. 39, no. 15, pp. 11924–11933, 2012.

[21] R. S. Wahono, “A Systematic Literature Review of Software Defect Prediction : Research Trends , Datasets , Methods and Frameworks,” *J. Softw. Eng.*, vol. 1, no. 1, pp. 1–16, 2015.

[22] B. Kitchenham and S. Charters, *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. 2007.

[23] J. Wu, Z. Cai, S. Zeng, and X. Zhu, “Artificial Immune System for Attribute Weighted Naive Bayes Classification,” *Comput. Intell. Syst.*, no. 61075063,

2011.

[24] N. A. Zaidi, M. J. Carman, and G. I. Webb, “Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting,” *Mach. Learn. Res.*, vol. 14, pp. 1947–1988, 2013.

[25] C. Deisy, S. Baskar, N. Ramraj, J. S. Koori, and P. Jeevanandam, “A Novel Information Theoretic Interact Algorithm (IT-IN) for Feature Selection Using Three Machine Learning Algorithms,” *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7589–7597, 2010.

[26] S. S. Kannan and N. Ramaraj, “A Novel Hybrid Feature Selection Via Symmetrical Uncertainty Ranking Based Local Memetic Search Algorithm,” *Knowledge-Based Syst.*, vol. 23, no. 6, pp. 580–585, 2010.

[27] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, “Hybrid Decision Tree and Naïve Bayes Classifiers for Multi-Class Classification Tasks,” *Expert Syst. Appl.*, vol. 41, no. 4 PART 2, pp. 1937–1946, 2014.

[28] L. Jiang, Z. Cai, H. Zhang, and D. Wang, “A Locally Weighted Learning Approach,” *Exp. Theor. Naive Bayes text Classif.*, no. November 2013, pp. 37–41, 2012.

[29] C. Li and H. Li, “One Dependence Value Difference Metric,” *Knowledge-Based Syst.*, vol. 24, no. 5, pp. 589–594, 2011.

[30] L. Jiang, Z. Cai, H. Zhang, and D. Wang, “Not So Greedy : Randomly Selected Naive Bayes,” *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11022–11028, 2012.

[31] J. Lin, “Weighted Naive Bayes Classification Algorithm Based on Particle Swarm Optimization,” *Commun. Softw. Networks*, pp. 444–447, 2011.

[32] L. Jiang, H. Zhang, Z. Cai, and D. Wang, “Weighted Average of One Dependence Estimators,” *Exp. Theor. Artif. Intell.*, no. December 2014, pp. 37–41, 2011.

[33] J. Wu, S. Pan, Z. Cai, X. Zhu, and C. Zhang, “Dual Instance and Attribute Weighting for Naive Bayes Classification,” *Comput. Intell. Syst.*, 2014.

[34] C. Lee, “A Gradient Approach for Value Weighted Classification Learning in Naive Bayes,” *Knowledge-Based Syst.*, vol. 85, pp. 71–79, 2015.

[35] S. Wang, L. Jiang, and C. Li, “Adapting Naive Bayes Tree for Text Classification,” *Knowl. Inf. Syst.*, pp. 77–89, 2015.

[36] J. Wu, S. Pan, X. Zhu, P. Zhang, and C. Zhang, “SODE : Self Adaptive One Dependence Estimators for Classification,” *Pattern Recognit.*, vol. 51, pp. 358–377, 2016.

[37] C. S. Division, M. Park, P. Langley, and P. Smyth, “Bayesian Network Classifiers,” *Mach. Learn.*, vol. 163, pp. 131–163, 1997.

[38] N. Ye, *Data Mining*. CRC Press, 2014.

[39] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Third Edition*. Elsevier, 2011.

[40] J. Demšar, “Statistical Comparisons of Classifiers Over Multiple Data Sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[41] G. W. Corder and D. I. Foreman, *Nonparametric Statistics for Non-Statisticians: A step by step approach*. 2009.