# Prediksi Penyakit Menggunakan *Genetic Algorithm (GA)* dan *Naive Bayes* Untuk Data Berdimensi Tinggi

# Prediction of Disease Using Genetic Algorithm (GA) and Naive Bayes For Data High Dimension

**Dwi Nugroho[1], Fhira Nhita S.T., M.T.[2], Danang Trantoro M, S.Si., M.T[3]**
[1,2,3]Ilmu Komputasi, Fakultas Informatika, Universitas Telkom
[1]mayriskhaii@gmail.com,[2]farid.alchair@gmail.com, [3]dto.lecturer@gmail.com

**ABSTRAK**

Sebuah sistem yang mampu menganalisis dan mengidentifikasi seseorang terkena suatu penyakit akan sangat membantu pada dunia medis, hal ini dikarenakan tingkat kematian setiap harinya selalu bertambah. Faktor kematian salah satunya adalah kurangnya penanganan dini pada pasien yang telah terkena suatu penyakit. Hal ini dapat terjadi karena pasien tidak mengetahui bahwa dirinya mengidap penyakit yang mematikan. Adapun lima penyakit yang digolongkan mematikan adalah Kanker, Jantung, Diabetes, AIDS, dan TBC.

Oleh karena itu, pada tugas akhir ini dibangun sistem yang mampu memprediksi seorang pasien apakah terjangkit penyakit atau tidak. Sistem ini sangat membantu dalam penanganan dini pasien yang terkena penyakit. Data penyakit Kanker, Jantung, Diabetes, AIDS, dan TBC bersumber dari website Kent Ridge Bio-medical Data Set Repository akan digunakan untuk membangun sistem ini, yang mana data tersebut adalah data berdimensi tinggi untuk setiap penyakit. Dimana data tersebut memiliki ribuan atribut yang akan dibagi menjadi dua data, yaitu data traning dan data testing selanjutnya dilakukan reduksi dengan Genetic Algorithm (GA) dan klasifikasi dengan Naïve Bayes Classifier.

Dengan prediksi menggunakan model tersebut, didapatkan hasil yang akan menunjukan seorang pasien yang terkena penyakit atau tidak. Selanjutnya dilakukan uji akurasi menggunakan data testing untuk mendapatkan hasil akurasi yang valid. Sehingga hasil akhir yang didapat menunjukkan metode crossvalidation lebih baik dengan nilai akurasi dari data colon tumor 88.89% dan leukemia 100% dibandingkan metode percentage split dengan akurasi dari data colon tumor 78.95% dan leukemia 77.27%.

**Kata Kunci** : *Genetic Algorithm, Naïve Bayes, Evolutionary Datamining*.

***ABSTRACT***

*A system that able to analyze and identify a person whether infected by a disease will extremely helping in medical world, because the increasingly death rates each day. One of the factors of death is the like of early treatment for the patients that are already infected by disease. This happens because the patients aren't realized that they're infected by deadly diseases. There are five diseases that are categorized as deadly which are cancer, heart attack, diabetes, AIDS, and TBC.*

*By that reason, on this final project will be built a system that able to predict whether a patient is infected by a disease or not. This system will help on the early treatment for patients that are caught by diseases. Data of cancer, heart attack, diabetes, AIDS, and TBC diseases a sourced from Kent Ridge Bio-medical Data Set Repository website will be used to build this system, in which these data are high dimensional data for each disease. These data has thousands of attributes that will be divided into two data, which are training data and testing data. Both data will be reduced by genetic algorithm (GA) method and classified by naïve bayes classifier method.*

*By predicting using this model, the result will show whether a patient have disease or not. Next is accuracy test using testing data to obtain the valid accuracy result. Therefore, from the result it shows that crossvalidation method is better with 88.89% accuracy for colon tumor data and 100% accuracy for leukemia data, compared to precentage split method with 78.95% accuracy for colon tumor data and 77.27% accuracy for leukemia data.*

*Keywords: Genetic Algorithm, Naïve Bayes, Evolutionary Datamining*.

# 1. Pendahuluan

## 1.1 Latar Belakang

Tingkat kematian masyarakat di dunia setiap harinya meningkat secara signifikan, salah satu faktor kematian pada masyarakat adalah penyakit yang menyerang masyarakat. Terdapat lima penyakit yang diklaim sebagai pembunuh teratas di dunia adalah Kanker, Jantung, Diabetes, AIDS, dan TBC. Kelima penyakit tersebut telah banyak memakan korban. Penanganan dini sebenarnya dapat dilakukan untuk menekan tingkat kematian masyarakat di dunia, salah satunya adalah dengan mencari informasi apakah seseorang mengidap penyakit tersebut. Dengan mengumpulkan data pasien yang berupa informasi tentang pasien yang mengidap penyakit merupakan langkah awal dalam menekan tingkat kematian. Namun dengan data berdimensi tinggi yang memiliki banyak atribut (dimensionality) tidak bisa menghasilkan informasi yang presisi. Oleh sebab itu atribut tersebut perlu direduksi supaya mengurangi data yang kurang informatif sehingga mampu membentuk model prediksi suatu penyakit, salah satu metode yang digunakan adalah evolutionary data mining.

Evolutionary Data Mining adalah penggabungan antara Algoritma Evolutionary dan Algoritma Data Mining. Evolutionary data mining dapat menangani masalah pada data dimensi tinggi, dan dapat memangkas waktu pengerjaan data dimensi tinggi, sehingga informasi akan lebih cepat didapatkan. Dalam tugas akhir ini Evolutionary Algorithms yang akan digunakan adalah Genetic Algorithm. Genetic Algorithm dipilih karena dapat mereduksi atribut pada data dimensi tinggi. Sehingga data yang awalnya memiliki banyak atribut direduksi menjadi beberapa atribut yang lebih sedikit, tanpa mengurangi informasi dari data tersebut. Algoritma yang digabungkan adalah Naïve Bayes dan Genetica Algorithm. Studi yang membahas tentang evolutionary data mining yaitu evolutionary data mining: an overview of genetic – based algorithms [1], kumpulan data medis [2] dan Heart Disease Prediction Sistem using Associative Classification and Genetic Algorithm. Pada penulisan tugas akhir ini paper yang digunakan adalah "Prediction of Heart Disease Using Genetic Algorithm for Selection of Optimal Reduced Set of Algorithm". Paper ini dipilih karena telah membuktikan bahwa Algoritma Genetika mampu mereduksi atribut pada data dimensi tinggi serta mampu membangun sistem yang dapat melakukan pengklasifikasian dengan Naïve Bayes, data dimensi tinggi yang digunakan adalah data pasien pengidap penyakit jantung UCI Repository, dan didapatkan hasil akurasi sebesar 96,5% [3].

Dalam tugas akhir ini dibangun suatu sistem yang dapat menganalisa dan memprediksi seseorang pasien apakah mengidap penyakit atau tidak, dengan berdasarkan data dari pasien yang kemudian dianalisa menggunakan evolutionary data mining. Adapun data yang digunakan dalam tugas akhir ini merupakan data berdimensi tinggi yaitu data pasien yang terkena penyakit yang diambil dari Kent Ridge Bio-medical Data Set Repository. Dengan model prediksi tersebut diharapkan dapat membatu permasalahan didunia medis.

## 1.2 Perumusan Masalah

Dalam tugas akhir ini, rumusan masalah yang dibahas adalah:

1. Bagaimana mengimplementasi *Naïve Bayes* dan *Genetic Algorithm (GA)* untuk data penyakit berdimensi tinggi?
2. Bagaimana cara kerja *Genetic Algorithm (GA)* untuk memilih atribut paling optimal pada data penyakit berdimensi tinggi*?*
3. Bagaimana membangun model yang mampu memprediksi pasien yang terjangkit penyakit dengan *Naïve Bayes?*

Batasan masalah yang digunakan dalam penelitian tugas akhir ini adalah sebagai berikut:

1. Dataset yang digunakan merupakan data penyakit berdimensi tinggi.
2. Dataset bertipe numerik.
3. Dataset yang digunakan yaitu beberapa data penyakit yang berasal dari *Kent Ridge Bio-medical Data Set Repository*

## 1.3 Tujuan

Adapun tujuan dari dilakukannya tugas akhir ini yaitu:

1. Mengimplementasikan *naïve bayes* dan *genetic algorithm (GA)* untuk data penyakit berdimensi tinggi.
2. Mengetahui cara kerja *genetic algorithm (GA)* untuk mendapatkan atribut yang paling optimal terhadap data penyakit berdimensi tinggi.
3. Membangun model yang mampu memprediksi pasien yang terkena penyakit dengan menggunakan *naïve bayes.*

## 2. Kajian Pustaka

### 2.1 Data Mining

Data mining adalah bagian dari proses penemuan pengetahuan terbesar yang meliputi tugas preprocessing seperti data extraction, data cleaning, data fusion, data reduction, dan feature construction, dan juga meliputi tugas post processing seperti pattern and model interpretation, hypothesis confirmation, dan sebagainya [3]. Proses data mining ini cenderung berulang dan interaktif.

### 2.2 Evolutionary Algorithms (EAs)

Evolutionary Algorithm (EAs) adalah algoritma-algoritma optimasi yang berbasis evolusi biologi yang ada di dunia nyata [14]. Dalam teori evolusi, suatu individu dalam sebuah populasi akan saling berkompetisi untuk dapat bertahan hidup di suatu daerah yang memiliki sumber daya terbatas. Tingkat adaptasi pada setiap individu dapat menentukan individu mana yang akan tetap bertahan hidup dan individu mana yang akan musnah.

#### 2.2.1. Genetic Algorithm (GA)

Genetic Algorithm (GA) adalah salah satu algoritma EAs. GA pertama kali dipublikasikan oleh John Holland (1975) di Amerika Serikat[14]. Pada saat itu, GA memiliki bentuk yang sangat sederhana sehingga disebut Simple GA. Ciri utama dari Algoritma ini adalah menitikberatkan pada rekombinasi (crossover) [15]. GA dapat digunakan sebagai tools pencarian yang optimal untuk memilih subset dari beberapa atribut [12], GA memiliki banyak keturunan yang dapat menjelajahi ruang solusi pada waktu bersamaan [17].

### 2.3 *Principal Component Analysis* (PCA)

PCA adalah sebuah teknik untuk membangun variable-variable baru yang merupakan kombinasi linear dari variable-variable asli. PCA adalah pengolahan data yang populer dan merupakan salah satu teknik reduksi dimensi untuk tipe data numerik. Fokus kerja PCA adalah 'meringkas' data, bukan mengelompokkan data seperti *clustering.* Principal Component Analysis (PCA) adalah suatu analisis yang menjelaskan struktur varian-kovarian dari suatu himpunan variabel yang melalui beberapa kombinasi linear dari variable – variabel tersebut. Konsep dasar matematika yang digunakan pada PCA meliputi standar deviasi, variansi, matriks kovariansi, matriks korelasi, vektor eigen, dan nilai eigen [10].

#### 2.3.1 Algoritma PCA

Algoritma PCA terdiri dari metode kovariansi dan korelasi. Metode kovariansi biasanya digunakan pada data yang memiliki satuan ukuran yang sama, misalkan data tinggi dengan satuan cm. Jika data tidak memiliki satuan ukuran yang sama maka digunakan metode korelasi [11]. Pada tugas akhir ini data memiliki satuan yaitu Ha, sehingga metode PCA yang digunakan adalah metode kovariansi.

##### 2.3.1.1 Metode kovariansi

Pada referensi [10] dijelaskan proses kerja PCA dengan menggunakan metode kovariansi dengan langkah-langkah sebagai berikut :

1. Mengurangi setiap record dari data asli *X* dengan *mean* $\overline{X}$ sehingga menghasilkan data hasil standardisasi yaitu *DataAdjust* dengan *mean* = 0.
2. Menentukan matriks kovariansi *S* dari *DataAdjust*.
3. Menghitung nilai eigen *L* dan vektor eigen *U* dari matriks *S.* Vektor eigen bersifat *orthonormal* sehingga :
   $$[illegible] = [illegible]^{-1} \quad \textbf{(2.1)}$$
4. Mengurutkan nilai eigen secara terurut menurun. Hubungan nilai eigen, vektor eigen, dan matriks kovariansi dinyatakan dengan persamaan :
   $$[illegible]SU = l \quad \textbf{(2.2)}$$
5. Mengubah *p* dimensi yang berkorelasi $x_1$ , $x_2$ , ..., $x_p$ menjadi *p* dimensi baru yang tidak berkorelasi $z_1$ , $z_2$ , ..., $z_p$ , dengan persamaan :
   $$l = [illegible]|x - [illegible] \quad \textbf{(2.3)}$$
6. Dimensi baru yang dihasilkan dari kombinasi linier dimensi asli tersebut dinamakan *principal component* (PC) dari $X$ . Setiap PC memiliki *mean*=0 dan variansi= $l_i$ . Nilai PC pada setiap record disebut *z-scores* yang dinyatakan dengan persamaan :

PC ke-i = ⬚ = ⬚ − ⬚ **(2.4)**

### 2.4 *Cross Validation*

Cross validation merupakan suatu metode evaluasi, dimana pada metode ini data yang digunakan dalam jumlah yang sama untuk training dan tepat satu kali untuk testing [8]. Crossvalidation memiliki beberapa pendekatan, pada tugas akhir ini pendekatan yang digunakan adalah Hold – Out. Hold – Out merupakan metode untuk memecah dataset menjadi dua bagian terpisah, yaitu set data latih dan set data uji.

### 2.5 *Naive Bayes Classifier*

#### 2.5.1 Teorema Bayes

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasarkan aturan Bayes dengan asumsi independensi (ketidaktergantungan) yang kuat. Independensi yang kuat dalam Bayes (terutama Naive Bayes) pada feature adalah bahwa sebuah feature pada sebuah data tidak berkaitan dengan nada atau tidaknya feature lain dalam data yang sama[9].

Prediksi Bayes didasarkan pada teorema Bayes dengan formula umum sebagai berikut:

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)} \quad \textbf{(2.5)}$$

Dimana:

P{H|E) = Probabilitas akhir bersyarat (conditional probability) suatu hipotesis H terjadi jika diberikan bukti (evidence) E terjadi.

P(E|H) = Probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis H

P(H) = Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun.

P(E) = Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis atau bukti yang lainnya.

#### 2.5.2 Naive Bayes Untuk Klasifikasi

Klasifikasi bayes mengasumsikan bahwa suatu *feature* tidak berpengaruh dengan adanya *feature* lain. Sebagai contoh, buah dapat disebut jeruk jika memiliki ciri-ciri berwarna orange, bulat, dan berdiameter 5 cm. *Feature* ini tidak bergantung satu sama lain atau pada saat adanya *feature* lain. Klasifikasi Naive Bayes menganggap *feature* ini *independent* [10].

Hipotesis dalam teorema Bayes merupakan label kelas yang menjadi target dalam pemetaan dalam klasifikasi, dan bukti merupakan feature-feature yang menjadi masukan dalam model klasifikasi. Jika E adalah masukan yang berisi feature dan H adalah label kelas, maka Naïve Bayes dituliskan dengan P(H|E) [9]. Notasi tersebut berarti probabilitas label kelas H didapatkan setelah feature-feature X diamati.

Formulasi Naive Bayes untuk klasifikasi:

$$P(H \mid E) = \frac{P(H)\prod_{i=1}^{q} P(E_i \mid H)}{P(E)} \quad (2.3)$$

Dimana:

$P(E_i \mid H)$ = Probabilitas data dengan *feature* E pada kelas H.

$P(H)$ = Probabilitas awal dengan kelas H

$\prod_{i=1}^{q} P(E_i \mid H)$ = Probabilitas independen kelas H dari semua *feature* dalam X

Umumnya bayes mudah dihitung untuk feature bertipe kategorikal. Perlakuan untuk data numerik akan sedikit berbeda dengan data kategorikal. Salah satunya adalah dengan

mengasumsikan bentuk tertentu dari distribusi menggunakan data *training*. Distribusi Gaussian biasanya dipilih untuk mempresentasikan *conditional probability feature continuous* pada

sebuah kelas P (X_i |Y). Distribusi Gaussian dikarakteristikan dengan dua parameter: rata-rata (μ) dan variansi (), x adalah nilai feature pada data yang akan diprediksi [11]. Persamaan distribusi gauss [9] adalah:

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp - \frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \quad (2.4)$$

Dimana:

μ = rata-rata

σ = standar deviasi

 = data ke i

## 3 Metodologi dan Desain Sistem

### 3.1 Deskripsi Sistem

Pada tugas akhir ini, perancangan sistem akan digunakan untuk memprediksi penyakit dimana pada data penyakit tersebut berdimensi tinggi. Pre-processing (normalisasi) digunakan untuk mengubah data asli menjadi data bernilai antara 0 – 0.9. Data yang sudah dinormalisasi dibagi menjadi dua kategori yaitu training dan testing, yang akan menjadi inputan sistem. Tahap selanjutnya adalah generate populasi yang ada di genetic algorithm dan dibantu dengan naive bayes sebagai pembangun model, keluaran ang diharapkan dari sistem adalah mendapatkan hasill akurasi dan atribut yang digunakan.

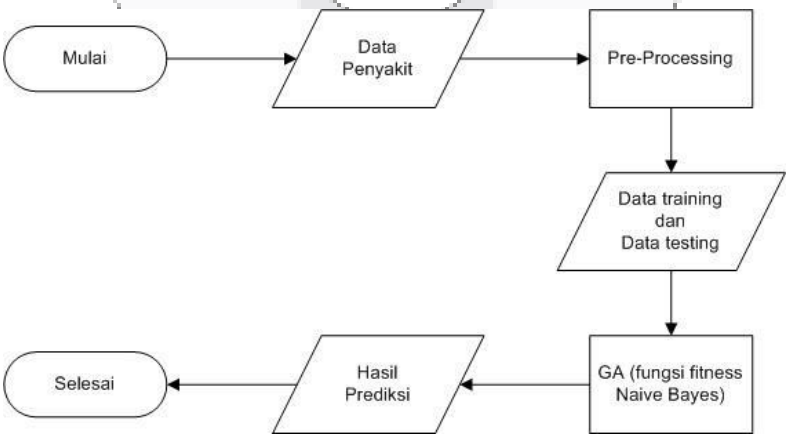
### 3.2 Dataset

Data yang digunakan merupakan data penyakit berdimensi tinggi berupa ekspresi gen yang diperoleh dari Kent Ridge Bio-medical Data Set Repository [11]. Data yang berada dalam repository tersebut meliputi gene expression data, protein profiling data dan genomic sequence data yang sudah pernah dipublikasikan dalam berbagai jurnal yaitu colon tumor dan Leukemia

### 3.3 Skenario Pengujian

Sistem prediksi yang akan dibangun secara umum terbagi dalam 3 tahap proses utama. Tahapan pertama adalah Pre-processing. Tahap kedua penulis menggunakan algoritma genetika untuk mereduksi data berdimensi tinggi. Tahap terakhir adalah mencari nilai fitness pada metode algoritma genetika dengan menggunakan metode naïve bayes,

Adapun diagram alur untuk sistem yang akan dibangun adalah sebagai berikut:




## 4. Pengujian dan Analasis

### 4.1 Metode Percentage Split

Pada bagian ini akan menampilkan hasil pengujian berdasarkan skenario yang telah dilakukan. Maka dapat dianalisa bagaimana pengaruh dari Ukuran Populasi (UkPop), Probabilitas mutasi (Pm) dan Probabilitas crossover (Pc).

Berdasarkan analisis yang sudah dilakukan dengan parameter ukuran populasi sebagai acuan perbandingan untuk menentukan akurasi dan atribut yang optimal, maka hasil prediksi dari kedua penyakit tersebut adalah sebagai berikut:

*Tabel 4.1.1 Hasil Prediksi untuk Penyakit Colon tumor dengan Percentage Split*

| Data Set | Skenario | Max Generasi | Pc | Pm | Ukuran Populasi | Atribut Optimal | Akurasi *Training* | Akurasi *Testing* |
|---|---|---|---|---|---|---|---|---|
| *Colon tumor* | 1 | 20 | 0.6 | 0.05 | 50 | 1006 | 37.21 % | 84.21 % |
| | | | 0.6 | 0.1 | | 1013 | 44.19 % | 78.95 % |
| | | | 0.8 | 0.05 | | 983 | 46.51 % | 78.95 % |
| | | | 0.8 | 0.1 | | 997 | 44.19 % | 78.95 % |
| | 2 | 10 | 0.6 | 0.05 | 100 | 975 | 39.53 % | 78.95 % |
| | | | 0.6 | 0.1 | | 984 | 41.86 % | 73.68 % |
| | | | 0.8 | 0.05 | | 1020 | 37.21 % | 78.95 % |
| | | | 0.8 | 0.1 | | 992 | 41.86 % | 78.95 % |
| | 3 | 5 | 0.6 | 0.05 | 200 | 960 | 37.21 % | 78.95 % |
| | | | 0.6 | 0.1 | | 944 | 41.86 % | 78.95 % |
| | | | 0.8 | 0.05 | | 992 | 51.16 % | 78.95 % |
| | | | 0.8 | 0.1 | | 976 | 41.85 % | 78.95 % |

Berdasarkan tabel diatas pada skenario satu dengan ukuran populasi 50 didapatkan atribut optimal sebesar 1006 dengan akurasi testing 84.21% dan akurasi training 37.21% dengan metode percentage split, yang mana Pc = 0.6 dan Pm = 0.05, dengan maks generasi 20. Pada skenario dua dengan ukuran populasi 100 didapatkan atribut optimal sebesar 975 dengan akurasi testing 78.95% dan akurasi training 39.53% yang mana Pc = 0.6 dan Pm = 0.05, dengan maks generasi 10. Pada skenario tiga dengan maks generasi dan ukuran populasi 200 didapatkan atribut optimal sebesar 944 dengan akurasi testing 78.95% yang mana Pc = 0.6 dan Pm = 0.1 dan akurasi training 41.86%. Dari ke tiga skenario tersebut dikarenakan nilai akurasi memiliki presentase yang sama maka atribut optimal yang dipilih adalah 1006 atribut dengan Pc = 0.6 dan Pm 0.5. Pada analisis ini dapat dilihat bahwa parameter Pc dan Pm juga berpengaruh terhadap nilai akurasi.

Dari hasil yang didapat, menunjukan nilai presentase dari akurasi training lebih kecil dibandingkan akurasi testing, hal itu disebabkan dari data sampel seperti data ekstrim, korelasi antar data. Hasil tersebut dapat dianalisis bahwa skenario ini menunjukan hasil prediksi dengan baik.

Tabel 4.1.2 Hasil Prediksi untuk Penyakit Leukemia dengan Percentage Split

| Data Set | Skenario | Max Generasi | Pc | Pm | Ukuran Populasi | Atribut Optimal | Akurasi *Training* | Akurasi *Testing* |
|---|---|---|---|---|---|---|---|---|
| *Leukemia* | 1 | 20 | 0.6 | 0.05 | 50 | 3582 | 90 % | 68.18 % |
| | | | 0.6 | 0.1 | | 3519 | 86 % | 68.18 % |
| | | | 0.8 | 0.05 | | 3537 | 92 % | 72.73 % |
| | | | 0.8 | 0.1 | | 3466 | 84 % | 68.18 % |
| | 2 | 10 | 0.6 | 0.05 | 100 | 3535 | 86 % | 68.18 % |
| | | | 0.6 | 0.1 | | 3561 | 88 % | 68.18 % |
| | | | 0.8 | 0.05 | | 3509 | 86 % | 72.73 % |
| | | | 0.8 | 0.1 | | 3483 | 84 % | 77.27 % |
| | 3 | 5 | 0.6 | 0.05 | 200 | 3607 | 88 % | 72.73 % |
| | | | 0.6 | 0.1 | | 3550 | 92 % | 68.18 % |
| | | | 0.8 | 0.05 | | 3589 | 92 % | 77.27 % |
| | | | 0.8 | 0.1 | | 3579 | 84 % | 72.73 % |

Berdasarkan tabel diatas pada skenario satu dengan ukuran populasi 50 didapatkan atribut optimal sebesar 3537 dengan akurasi testing 72.73% dan akurasi training 92% dengan metode percentage split, yang mana Pc = 0.8 dan Pm = 0.05, dengan maks generasi 20. Pada skenario dua dengan ukuran populasi 100 didapatkan atribut optimal sebesar 3483 dengan akurasi testing 77.27% dan akurasi training 84% yang mana Pc = 0.8 dan Pm = 0.1 dengan maks generasi 10. Pada skenario tiga dengan ukuran populasi 200 dan maks generasi 5 didapatkan atribut optimal sebesar 3589 dengan akurasi testing 77.27% dan akurasi treaining 92% yang mana Pc = 0.8 dan Pm = 0.1. Dari skenario dua dan skenario tiga memiliki nilai akurasi dengan presentase yang sama maka atribut optimal yang

dipilih adalah 3483 atribut dengan Pc = 0.8 dan Pm 0.1. Pada analisis ini dapat dilihat bahwa parameter Pc dan Pm juga berpengaruh terhadap nilai akurasi.

Dari hasil tabel diatas pada kolom akurasi testing dan training menunjukan presentase nilai akurasi training lebih besar dibandingkan presentasi nilai akurasi testing, dapat disimpulkan bahwa model yang sudah dibangun berdasarkan parameter tertentu tidak semua baik untuk setiap data.

### 4.2 Metode Crossvalidation dengan Hold-Out

Pada skenario ini akan menampilkan hasil pengujian dari metode hold-out dimana pembagian data training dan data pengujian secara acak oleh sistem dengan proporsi 70% dan 30%. Berdasarkan analisis yang sudah dilakukan dengan parameter ukuran populas sebagai acuan perbandingan dan parameter Pc dan Pm sebagai pembanding kedua, untuk menetukan akurasi dan atribut yang optimal. Hasil prediksi dari kedua penyakit sebagai berikut:

Tabel 4.2.2.1 Hasil Prediksi untuk Penyakit Colon tumor dengan Hold-Out

| Data Set | Skenario | Max Generasi | Pc | Pm | Ukuran Populasi | Atribut Optimal | Akurasi *Training* | Akurasi *Testing* |
|---|---|---|---|---|---|---|---|---|
| *Colon tumor* | 1 | 20 | 0.6 | 0.05 | 50 | 964 | 35 % | 72.22 % |
| | | | 0.6 | 0.1 | | 968 | 31 % | 72.22 % |
| | | | 0.8 | 0.05 | | 992 | 40 % | 83.33 % |
| | | | 0.8 | 0.1 | | 984 | 35 % | 83.33 % |
| | 2 | 10 | 0.6 | 0.05 | 100 | 1031 | 28 % | 83.33 % |
| | | | 0.6 | 0.1 | | 1020 | 31 % | 88.89 % |
| | | | 0.8 | 0.05 | | 1013 | 43 % | 83.33 % |
| | | | 0.8 | 0.1 | | 1008 | 33 % | 83.33 % |
| | 3 | 5 | 0.6 | 0.05 | 200 | 973 | 32 % | 83.33 % |
| | | | 0.6 | 0.1 | | 1000 | 34 % | 83.33 % |
| | | | 0.8 | 0.05 | | 1002 | 34 % | 88.89 % |
| | | | 0.8 | 0.1 | | 975 | 28 % | 83.33 % |

Dari hasil tabel diatas pada skenario satu dengan ukuran populasi 50 mendapatkan akurasi sebesar 83.33% dengan 984 atribut, pada skenario ke dua dengan ukuran populasi 100 mendapatkan akurasi 88.89% dengan 1020 atribut dan pada skenario ke tiga mendapatkan akurasi sebesar 88.89% dengan 1002 atribut. Dapat disimpulkan bahwa dari data penyakit colon tumor hasil prediksi yang terbaik yaitu pada skenario ke tiga dengan ukuran populasi 200, Pc = 0.8 dan Pm = 0.05. Dari parameter tersebut mendapatkan akurasi sebesar 88.89% dan 1002 atribut yang digunakan.

Tabel 4.2.2.2 Hasil Prediksi untuk Penyakit Leukemia dengan Hold-Out

| Data Set | Skenario | Max Generasi | Pc | Pm | Ukuran Populasi | Atribut Optimal | Akurasi *Training* | Akurasi *Testing* |
|---|---|---|---|---|---|---|---|---|
| *Leukemia* | 1 | 20 | 0.6 | 0.05 | 50 | 3509 | 51 % | 100 % |
| | | | 0.6 | 0.1 | | 3436 | 51 % | 100 % |
| | | | 0.8 | 0.05 | | 3468 | 51 % | 100 % |
| | | | 0.8 | 0.1 | | 3492 | 51 % | 100 % |
| | 2 | 10 | 0.6 | 0.05 | 100 | 3485 | 51 % | 100 % |
| | | | 0.6 | 0.1 | | 3450 | 51 % | 100 % |
| | | | 0.8 | 0.05 | | 3438 | 51 % | 100 % |
| | | | 0.8 | 0.1 | | 3478 | 51 % | 100 % |
| | 3 | 5 | 0.6 | 0.05 | 200 | 3493 | 51 % | 100 % |
| | | | 0.6 | 0.1 | | 3499 | 51 % | 100 % |
| | | | 0.8 | 0.05 | | 3458 | 51 % | 100 % |
| | | | 0.8 | 0.1 | | 3461 | 51 % | 100 % |

Dari hasil tabel diatas pada skenario satu dengan ukuran populasi 50 mendapatkan akurasi sebesar 100% dengan 3436 atribut, pada skenario ke dua dengan ukuran populasi 100 mendapatkan akurasi 100% dengan 3438 atribut dan pada skenario ke tiga mendapatkan akurasi sebesar 100% dengan 3493 atribut. Dapat disimpulkan bahwa dari data penyakit leukemia hasil prediksi yang terbaik yaitu pada skenario ke satu dengan ukuran populasi

50, Pc = 0.6 dan Pm = 0.1. Dari parameter tersebut mendapatkan akurasi sebesar 100% dan 3436 atribut yang digunakan.

### 4.3 Metode Percentage Split dengan PCA

Pada skenario ini data dipreprocessing terlebih dahulu dengan metode PCA. Data colon tumor sebesar 62x2000 direduksi menjadi 62x60 dan untuk data leukemia sebesar 72x7129 direduksi menjadi 72x70.

Tabel 4.2.3.1 Hasil Prediksi untuk Penyakit Colon tumor dengan PCA

| Data Set | Skenario | Max Generasi | Pc | Pm | Ukuran Populasi | Atribut Optimal | Akurasi *Training* | Akurasi *Testing* |
|---|---|---|---|---|---|---|---|---|
| *Colon tumor* | 1 | 20 | 0.6 | 0.05 | 50 | 31 | 97.67% | 94.74% |
| | | | 0.6 | 0.1 | | 26 | 90.70% | 84.21% |
| | | | 0.8 | 0.05 | | 28 | 83.72% | 94.74% |
| | | | 0.8 | 0.1 | | 26 | 86.05% | 89.47% |
| | 2 | 10 | 0.6 | 0.05 | 100 | 36 | 93.02% | 94.74% |
| | | | 0.6 | 0.1 | | 29 | 90.70% | 89.47% |
| | | | 0.8 | 0.05 | | 22 | 86.05% | 89.47% |
| | | | 0.8 | 0.1 | | 26 | 90.70% | 89.47% |
| | 3 | 5 | 0.6 | 0.05 | 200 | 24 | 93.02% | 89.47% |
| | | | 0.6 | 0.1 | | 28 | 90.70% | 94.74% |
| | | | 0.8 | 0.05 | | 32 | 90.70% | 89.47% |
| | | | 0.8 | 0.1 | | 28 | 88.37% | 89.47% |

Berdasarkan tabel 4.2.3.1 pada skenario satu dengan ukuran populasi 50 mendapatkan akurasi sebesar 94.74% dengan 28 atribut, pada skenario ke dua dengan ukuran populasi 100 mendapatkan akurasi 94.74% dengan 36 atribut dan pada skenario ke tiga mendapatkan akurasi sebesar 94.74% dengan 28 atribut. Dapat disimpulkan bahwa dari data penyakit colon tumor hasil prediksi yang terbaik yaitu pada skenario ke satu dan tiga. Dengan ukuran populasi 200, Pc = 0.8 dan Pm = 0.05 untuk skenario tiga sedangkan untuk skenario satu dengan ukuran populasi 50, Pc = 0.8 dan Pm = 0.05. Dari parameter tersebut mendapatkan akurasi sebesar 94.74% dan 28 atribut yang digunakan.

Tabel 4.2.3.2 Hasil Prediksi untuk Penyakit Leukemia dengan PCA

| Data Set | Skenario | Max Generasi | Pc | Pm | Ukuran Populasi | Atribut Optimal | Akurasi *Training* | Akurasi *Testing* |
|---|---|---|---|---|---|---|---|---|
| *Leukemia* | 1 | 20 | 0.6 | 0.05 | 50 | 39 | 86% | 68.18% |
| | | | 0.6 | 0.1 | | 34 | 88% | 77.27% |
| | | | 0.8 | 0.05 | | 29 | 84% | 72.73% |
| | | | 0.8 | 0.1 | | 27 | 76% | 72.73.% |
| | 2 | 10 | 0.6 | 0.05 | 100 | 36 | 86% | 72.73% |
| | | | 0.6 | 0.1 | | 28 | 88% | 68.18% |
| | | | 0.8 | 0.05 | | 32 | 88% | 77.27% |
| | | | 0.8 | 0.1 | | 37 | 86% | 81.82% |
| | 3 | 5 | 0.6 | 0.05 | 200 | 31 | 76% | 68.18% |
| | | | 0.6 | 0.1 | | 33 | 88% | 72.73% |
| | | | 0.8 | 0.05 | | 31 | 82% | 72.73% |
| | | | 0.8 | 0.1 | | 26 | 84% | 72.73% |

Dari hasil tabel diatas pada skenario satu dengan ukuran populasi 50 mendapatkan akurasi sebesar 77.27% dengan 34 atribut, pada skenario ke dua dengan ukuran populasi 100 mendapatkan akurasi 81.82% dengan 37 atribut dan pada skenario ke tiga mendapatkan akurasi sebesar 72.73% dengan 26 atribut. Dapat disimpulkan bahwa dari data penyakit leukemia hasil prediksi yang terbaik yaitu pada skenario ke satu dengan ukuran populasi 100, Pc = 0.8 dan Pm = 0.1. Dari parameter tersebut mendapatkan akurasi sebesar 81.82% dan 37 atribut yang digunakan.

Berdasarkan hasil dari tiga skenario yang di uji menunjukan bahwa crossvalidation merupakan metode terbaik untuk memprediksi penyakit leukemia. Pada data colon tumor, metode terbaik yang digunakan untuk memprediksi

adalah PCA. Hal ini dibuktikan dari hasil akurasi testing lebih besar dari pada akurasi training. Sedangkan metode dipilih berdasarkan akurasi terbesar dari hasil tersebut.

## 5. Kesimpulan dan Saran

### 5.1 Kesimpulan

Berdasarkan analisis dan pengujian yang telah dilakukan dapat diberikan berapa kesimpulan, yaitu:

1. Dari pengujian yang telah dilakukan diperoleh akurasi sebesar 94.74% untuk penyakit colon tumor dengan atribut optimal sebesar 28. Sedangkan untuk penyakit leukemia diperoleh akurasi sebesar 100% dengan atribut optimal sebesar 3436.
2. Kondisi optimal pada implementasi sistem dipilih berdasarkan akurasi paling besar dengan memuat informasi atribut optimal dan metode yang digunakan. Jika terdapat hasil akurasi yang sama, maka atribut optimal dipilih dari sedikitnya jumlah atribut.
3. Berdasarkan tiga skenario yang telah dilakukan diperoleh model terbaik untuk data penyakit colon tumor dengan pengujian menggunakan algoritma PCA tanpa crossvalidation. Sedangkan untuk penyakit leukemia model terbaik diperoleh saat pengujian menggunakan algoritma crossvalidation tanpa PCA.

### 5.2 Saran

Setelah proses prediksi penyakit ini, penulis menemukan beberapa saran yang dapat dilakukan, yaitu:

1. Untuk meningkatkan akurasi yang dihasilkan, maka perlu dilakukan penambahan record pada data.
2. Data yang digunakan, seharusnya dicheck terlebih dahulu apakah data tersebut memiliki pencilan atau data ektrim

## Daftar Pustaka

[1] Collard, M., CNRS, S. F., & Francisi, D. (2001). Evolutionary data mining: an overview of genetic based algorithm. Institute of Electrical and Electronics Engineers (IEEE).

[2] Kumar, G.R, Ramachandra, D.G.A. & Nagamani, K., 2014. An Efficient eature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets. International Journal of Advabced Research in Computer Science and Software Engineering.

[3] Ratnakar, S., Rajeswari, K., & Jacob, R. (2013). Prediction of Heart Diease Using Genetic Algorithm for Selection of Optimal Reduced Set of Attributes. International Journal of Advanced Computing Engineeing and Networking.

[4] Prasetyo Eko, Data Mining, Mengolah Data Menjadi Informasi Menggunakan Matlab, Yogyakarta: ANDI, (2014)

[5] Etin, "Kecerdasan Buatan: Bab & Algoritma Genetika," [Online]. Available: http://lecturer.eepisits.edu/~entin/Kecerdasan%20Buatan/Buku/Bab%207%20Algoritma%20Genetika.pdf. [Diakses 03 3 2016].

[6] Suyanto, S. M. (2008). Evolutionary Computing. Bandung: Informatika Bandung.

[7] Suyanto, Soft Computing: Membangun Mesin Ber-IQ Tinggi, Bandung: Informatika, 2008.

[8] Prasetyo, E. (2009). Data Mining Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta: Andi.

[9] Suyanto, S. M. (2008). Soft Computing. Bandung: Informatika.

[10] Han, J., Kamber, M., & Pei, J. (2012). Data mining Conpects and Techniques. Singapore: Markono Print Media Ptc Ltd.

[11] Shantanam, T. & Padmavathi, M.S., 2015. Aplication of K-Means and Genetic Algorithms for Dimensional Reduction by Intergrating SVM for Diabetes Diagnosis. ScienceDirect.

[12] Susetyoko, R. & Purwantini, E. Teknik Reduksi Dimensi Menggunakan Komponen Utama Data Partisi Pada Pengklasifikasian Data berdimensi Tinggo dengan Ukuran Sampel Kecil. Surabaya: Politeknik Elektronika Negeri Surabaya.

[13] Fhira Nhita, "Analisis Principal Components Analysis (PCA) pada Unsupervised Learning untuk Data Berdimensi Tinggi". Laporan Tugas Akhir, Jurusan Teknik Informatika, STT Telkom.

[14] Jackson, J.Edward. A User's Guide to Principal Components. New York, (1991)