**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

# ANALISA PERBANDINGAN ALGORITMA KLASIFIKASI *SUPPORT VECTOR MACHINE,DECESSION TREE* DAN *NAIVE BAYES*

Agus Heri Yunial
Prodi Teknik Informatika, Fakultas Teknik, Universitas Pamulang
e-mail : dosen02525@unpam.ac.id

**ABSTRAK**

Data bisa hanya menjadi sebatas sekumpulan fakta atau statistik bagi seseorang namun bisa juga menjadi informasi yang sangat berguna jika diolah dan digali demi mendapatkan suatu informasi baru yang bisa digunakan sebagai analisa dan persiapan yang akan terjadi kedepannya, dimana pengolahan data tersebut dinamakan data mining. Klasifikasi adalah salah satu algoritma pada data mining yang menelompokan suatau data kedalam kriteria atau kategori tertentu dengan membaca data sebelumnya yang sudah ada. Beberapa algoritma klasivikasi yang sering digunakan diantaranta adalah *Support Vector Machine, Decession Tree,* dan *Naive Bayes*. Dimana pada penelitian ini akan dilakukan analisa terhadap hasil dari nilai akurasi ketiga algortima tersebut dalam mengklasifikasikan suatu data. Data yang digunakan berupa dataset dari uci *Machine Learning* diantaranya *Chronic Kidney Disease, facebook large,* dan *Breast Cancer.* Hasil dari penelitian ini didapatkan nilai akurasi rata-rata menggunakan algoritma decision tree adalah sebesar 87.43 % dan merupakan nilai akurasi tertinggi dari kedua algoritma lainnya. Algoritma Support Vector Machine adalah sebesar 87.16 %, dan nilai akurasi rata-rata menggunakan algoritma Naive Bayes adalah sebesar 84.92 %.
**Kata kunci:** Data Mining, Klasifikasi, *Support Vector Machine, Decession Tree, Naive Bayes*

***ABSTRACT***

Data can only be limited to a collection of facts or statistics for someone, but it can also be very useful information if it is processed and extracted in order to obtain new information that can be used as analysis and preparation for what will happen in the future, where data processing is called data mining. Classification is one of the algorithms in data mining that classifies data into certain criteria or categories by reading the previous data that already exists. Some of the classification algorithms that are often used include Support Vector Machine, Decession Tree, and Naive Bayes. Where in this study an analysis will be carried out on the results of the accuracy value of the three algorithms in classifying a data. The data used in the form of a dataset from uci *Machine Learning* including Chronic Kidney Disease, Facebook Large, and Breast Cancer. The results of this study show that the average accuracy value using the decision tree algorithm is 87.43% and is the highest accuracy value of the other two algorithms.The Support Vector Machine algorithm is 87.16%, and the average accuracy value using the Naive Bayes algorithm is 84.92%..

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

***Keywords :*** *Data Mining, clasification, Support Vector Machine, Decession Tree, Naive Bayes*

## 1. PENDAHULUAN

Data adalah sekumpulan fakta yang terjadi yang masih berupa bahan mentah dan perlu diolah untuk bisa menjadi suatu informasi yang bermanfaat. Banyak informasi yang dapat diperoleh dari pengolahan data tersebut yang bisa digunakan untuk mendukung berbagai keputusan yang akan dibuat kedepannya dalam meminimalisir reskio yang akan terjadi. Informasi tersebut juga dapat digunakan dalam menentukan kriteria atau kategori dari data tersebut. Bisa jiga dunakaan untuk mengelompokan data-data yang berbeda kedalam klaster-klaster tertentu. Pengolahan data demikian dinamakan data mining, yaitu proses pengolahan data untuk bisa didapatkan informasi yang baru yang berguna dan bisa dimanfaatkan.

Pada data mining ada beberapa algoritma, diantaranya adalah klasifikasi. Klasifikasi adalah algoritma pada data mining yang menggolongkan atau menentukan suatu kriteria dari suatu data yang didasarkan pada data sebelumnya yang sudah dipelajari. Pada klasifikasi terdapat beberapa algoritma yang sering digunakan diantaranya adalah *Support Vector Machine, Decession Tree,* dan *Naive Bayes*. Perlu diketahui bahwa dalam algoritma klasifikasi nilai akurasi merupakan salah satu poin penting dalam menilai keakuraatan dari suatu algoritma klasifikasi itu sendiri. Semakin tinggi nali akurasi dari algoritma tersebut, semakin baik juga kemampuan algoritma tersebut dalam mengklasifikasikan suatu data. Oleh karena itu pada penelitian kali ini akan dilakukan perbandingan antara nilai akurasi dari algoritma klasifikasi *Support Vector Machine, Decession Tree, dan Naive Bayes*. Pada penelitian kali ini, data yang digunakan adalah berupa 2 buah dataset dari UCI *Machine Learning* yaitu data *Chronic Kidney Disease* dan *Breast Cancer*. Data yang diambil memiliki perbedaan masing-masing. Pada data *Chronic Kidney Disease* terdapat 400 *instance* atau record data dan 25 atribut sedangkan pada data *Breast Cancer* terdapat 286 instances dan 9 atribut. Dari kedua data tersebut akan dibandingkan keakurasian dari algoritma klasifikasi dalam mengolah data dengan perbandingan atribut dan instances yang beragam.

Pada penelitian lain ada juga yang membandingkan beberapa algortima klasifikasi dengan menggunakan beberapa dataset. Pada penelitian Ardea dan Fahrurozi yang dilakukan pada tahun 2019 , dilakukan perbandingan algoritma Naïve Bayes, K-Nearest Neighbor, Decision Tree dan Random Forest dalam mengklasifikasikan data penyakit jantung koroner. Metode uji yang digunakan adalah *cross validation* . Data yang digunakan adalah data penyakit jantung koroner yang berasal dari lokasi Cleveland Clinic Foundation yang telah diadopsi oleh instansi Hungarian Institute of Cardiology. Data tersebut terdiri dari 300 instances atau record dan 14 atribut atau variabel. Hasil penelitian mereka adalah didapatkan klasifikasi dengan algoritma Random Forest memiliki nilai akurasi sebesar 85,668 %, Naïve Bayes dan Decision Tree memiliki nilai akurasi sebesar 80.33 % sedangkan K-Nearest Neighbor memiliki nilai akurasi sebesar 69.67 % [1].

Penelitian yang dilakukan oleh Ardiyansyah dan kawan-kawan pada tahun 2018 yang membandingkan algoritma decision tree, Naïve bayes, k-nearest neighbour, ID3, dan CHAID. Pada penelitian mereka, data yang digunakan adalah dataset dari UCI *Machine Learning* yaitu dataset blogger. Data yang digunakan

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

terdiri dari 100 instances dan 6 atribut. Metode uji yang dilakukan adalah dengan menggunakan *cross validation* . Hasil dari penelitian mereka adalah algoritma k-nearest neighbour memiliki nilai akurasi tertinggi sebesar 85%, ID3 sebesar 82.00%, CHAID sebesar 75.00%, *Naive Bayes* sebesar 71,00% dan Decision Tree sebesar 68.00% [2].

Penelitian yang dilakukan oleh Dian Prajarini pada tahun 2016 yang membandingkan algoritma *Decession Tree, Naive Bayes*, k-nearest neighbour and suport vector machine. Data yang digunakan adalah dataset dari UCI yaitu dataset penyakit kulit. Data tersebut memiliki 366 instances dan 35 atribut. Metode uji yang digunakan adalah persentase split 50, 60, 70, 80, dan 90 %. Hasil rata-rata dari penelitian tersebut adalah didapat nilai akurasi tertinggi oleh *Support Vector Machine* dan *Naive Bayes* sebesar 98.1%, K-Nearest Neighbor sebesar 95.3%, sedangkan Decision Tree sebesar 94.7%. [3]

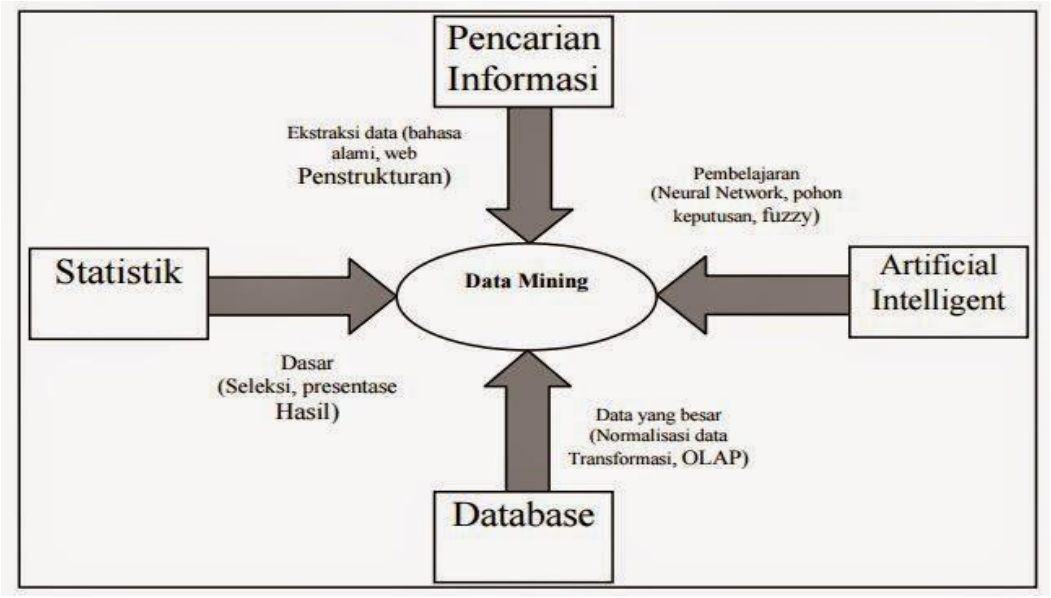
Pada penelitian ini, untuk menguji keakurasian dari algoritma *Support Vector Machine, Decession Tree, dan Naive Bayes* akan dilakukan metode pengujian *cross validation* -10 dengan menggunakan aplikasi weka versi 3.9.4.

## 2. LANDASAN TEORI

### 2.1 Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan dan *Machine Learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [4].

Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah maupun terlebih dahulu. Gambar 2.1 Menunjukan bahwa data mining memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (artificial intelligent), machine learning, statistic, database, dan jug informasi retrieval [5].




Gambar 2.1 Bidang Ilmu Data Mining

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

## 2.2 Klasifikasi

Pengukuran kinerja klasifikasi dengan menggunakan rumus akurasi dari matriks konfusi, sebagai berikut [6]:

Akurasi = (Jumlah benar / Jumlah data uji) x 100%

Klasifikasi merupakan proses menemukan sebuah model (fungsi) yang mendeskripsikan dan membedakan kelas data. Model tersebut diperoleh dari analisis data pelatihan (training data). Proses klasifikasi bisa dilakukan setelah analisis relevansi yang bertujuan untuk menentukan atribut yang relevan dengan proses tersebut. Ketepatan prediksi dari suatu pengklasifikasi dapat dituangkan dalam tabel kontingensi (contingency table) atau confusion matrix (Flach) seperti pada tabel berikut:

Tabel 2.1 Tabel confusion matrix

| | Prediksi (+) | Prediksi (-) |
|---|---|---|
| Aktual (+) | TP<br>(True Positives) | FN<br>(False Negatives) |
| Aktual (-) | FP<br>(False Positives) | TN<br>(True Negatives) |

Keterangan:
TP: Jumlah kasus positif yang diklasifikasi sebagai positif
FP: Jumlah kasus negatif yang diklasifikasi sebagai positif
TN: Jumlah kasus negatif yang diklasifikasi sebagai negatif
FN: Jumlah kasus positif yang diklasifikasi sebagai negatif
TP dan TN memberikan informasi ketika classifier benar, sedangkan FP dan FN memberitahu ketika classifier salah.

Kinerja klasifikasi bisa dievaluasi dengan memperhatikan ukuran-ukuran sebagai berikut [7]:

1) Akurasi
   Akurasi adalah ukuran untuk mengukur ketepatan prediksi pengklasifikasi pada kelas tertentu. Rumus untuk menghitung akurasi klasifikasi sesuai persamaan berikut:
   $$\text{Akurasi} = \frac{\text{Jumlah klasifikasi benar}}{\text{jumlah total data}} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (2.1)$$
2) Precision
   Precision adalah ukuran untuk mengukur ketepatan prediksi pengklasifikasi pada kelas tertentu. Rumus untuk menghitung precision klasifikasi sesuai persamaan berikut:
   $$\text{Precision} = \frac{TP}{TP+FP} \quad (2.2)$$
3) Sensitivitas
   Sensitivitas adalah ukuran untuk mengukur berapa banyak data dari kelas tertentu yang dapat diprediksikan secara benar. Rumus untuk menghitung sensitivitas klasifikasi sesuai persamaan berikut:
   $$\text{Sensitivitas} = \frac{TP}{TP+FN} \quad (2.3)$$

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

4) Spesifitas
   Spesifitas adalah ukuran untuk mengukur berapa banyak data dari kelas tertentu yang dapat diprediksikan secara benar. Rumus untuk menghitung spesifitas klasifikasi sesuai persamaan berikut:
   Spesifitas = $\frac{TP}{TN+FP}$ (2.4)

Parameter yang digunakan untuk membandingkan kinerja dari beberapa algoritma klasifikasi adalah (Fitri, 2014):

1) Test Mode: Mendefinisikan mode tes yang digunakan adalah cross-validation test dan percentage split test mode untuk teknik evaluasi.
2) Time to build model: merupakan istilah untuk menerangkan berapa waktu yang
   dibutuhkan untuk membangun model klasifikasi untuk masing-masing algoritma
3) Correctly classified instances: berapa banyak baris data yang terklasifikasikan dengan benar.
4) Incorrectly classified instances: berapa banyak baris data yang terklasifikasikan tidak benar.

## 2.3 Suppor Vector Machine

Algoritma *Support Vector Machine* atau sering disingkat SVM, adalah algoritma klasifikasi yang cenderung baru dikembangkan, karena dimulai pada tahun 1992 oleh Vladimir Vapnik dan rekannya, Bernhard Boser dan Isabelle Guyon yang dikembangkan dari teori Structural Risk Minimaization (SRM) dengan menggunakan trik kernel untuk memetakan sampel pelatihan dari ruang input ke ruang fitur dimensi tinggi [8].

SVM adalah sebuah algoritma yang bekerja menggunakan pemetaan nonlinear untuk mengubah data pelatihan asli ke dimensi yang lebih tinggi. Dalam dimensi yang baru, kemudian akan mencari linear optimal pemisah hyperplane (yaitu, "decision boundary" yang memisahkan tupel dari satu kelas dengan kelas lainnya) [9].

## 2.4 Decision Tree

Decision trees adalah struktur flowchart yang menyerupai trees (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas [10].

Algoritma C4.5 merupakan bagian dari kelompok algoritma decision trees dan merupakan kategori 10 algoritma yang paling popular. Algoritma C4.5 diperkenalkan oleh J. Ross Quinlan diakhir tahun 1970 hingga awal tahun 1980-an. J. Ross Quinlan seorang peneliti dibidang mesin pembelajaran yang merupakan pengembangan dari algoritma ID3 (Interative Dichotomiser), algoritma tersebut digunakan untuk membentuk pohon keputusan [11].

## *2.5 Naive Bayes*

*Naive Bayes* merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (aturan bayes) dengan asumsi

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

independensi ( ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam Naive Bayes model yang digunakan adalah model fitur independen [12].

## 3. Metode Penelitan

### 3.1. Analisa Kebutuhan

Dalam penelitian ini digunakan spesifikasi *hardware* dan *software* sebagai alat bantu penelitian yaitu :

a. Perangkat Keras
   - Laptop dengan spesifikasi seperti *Processor* : Core I5, Ram : 8 GB dan *Hardisk* 1 TB

b. Perangkat Lunak
   - Sistem Operasi *Windows* 10
   - *Microsoft Office* 2016
   - Aplikasi Weka versi 3.9.4

c. Dataset

Dataset yang digunakan pada penelitian ini adalah dataset yang dapat didownload situs UCI *Machine Learning* dengan alamat (https://archive.ics.uci.edu/ml/index.php). Ada 2 buah dataset yang digunkan, diantaranya:

- Dataset *Chronic Kidney Disease*

  Data ini berisi tentang gejala orang yang mengidap penyakit ginjal kronis yang dikategorikan sebagai ckd (*Chronic Kidney Disease*). Dataset ini terdidi dari 400 instances atau record dan 25 atribut yang terdidi dari:

**Tabel 3.1 Tabel atribut dataset *Chronic Kidney Disease***

| No | Nama Atribut | Tipe | Keterangan |
|---|---|---|---|
| 1 | Age | Numerik | age in years |
| 2 | Blood Pressure | Numerik | bp in mm/Hg |
| 3 | Specific Gravity | Alphanumerik | sg - (1.005,1.010,1.015,1.020,1.025) |
| 4 | Albumin | Alphanumerik | al - (0,1,2,3,4,5) |
| 5 | Sugar | Alphanumerik | su - (0,1,2,3,4,5) |
| 6 | Red Blood Cells | Alphanumerik | rbc - (normal,abnormal) |
| 7 | Pus Cell | Alphanumerik | pc - (normal,abnormal) |
| 8 | Pus Cell clumps | Alphanumerik | pcc - (present,notpresent) |
| 9 | Bacteria | Alphanumerik | ba - (present,notpresent) |
| 10 | Blood Glucose Random | Numerik | bgr in mgs/dl |

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

| | | | |
|---|---|---|---|
| 11 | Blood Urea | Numerik | bu in mgs/dl |

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

| | | | |
|---|---|---|---|
| 12 | Serum Creatinine | Numerik | sc in mgs/dl |
| 13 | Sodium | Numerik | sod in mEq/L |
| 14 | Potassium | Numerik | pot in mEq/L |
| 15 | Hemoglobin | Numerik | hemo in gms |
| 16 | Packed Cell Volume | Numerik | |
| 17 | White Blood Cell Count | Numerik | wc in cells/cumm |
| 18 | Red Blood Cell Count | Numerik | rc in millions/cmm |
| 19 | Hypertension | Alphanumerik | htn - (yes,no) |
| 20 | Diabetes Mellitus | Alphanumerik | dm - (yes,no) |
| 21 | Coronary Artery Disease | Alphanumerik | cad - (yes,no) |
| 22 | Appetite | Alphanumerik | appet - (good,poor) |
| 23 | Pedal Edema | Alphanumerik | pe - (yes,no) |
| 24 | Anemia | Alphanumerik | ane - (yes,no) |
| 25 | Class | Kategori | class - (ckd,notckd) |

- Dataset *Breast Cancer*

  Dataset ini berisi gejala orang yang mengidap kanker payudara. Dataset ini memiliki 286 instances atau record dan 10 atribut yang terdiri dari:

**Tabel 3.2 Tabel atribut dataset *Breast Cancer***

| No | Nama Atribut | Tipe | Keterangan |
|---|---|---|---|
| 1 | Class | Alphanumerik | no-recurrence-events, recurrence-events |
| 2 | Age | Alphanumerik | 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99. |
| 3 | Menopause | Alphanumerik | lt40, ge40, premeno. |
| 4 | Tumor-size | Alphanumerik | 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59. |

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

| | | | |
|---|---|---|---|
| 5 | Inv-nodes | Alphanumerik | 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39 |
| 6 | Node-caps | Alphanumerik | yes, no. |
| 7 | Deg-malig | Numerik | 1, 2, 3. |
| 8 | Breast | Alphanumerik | left, right. |
| 9 | Breast-quad | Alphanumerik | left-up, left-low, right-up, right-low, central. |
| 10 | Irradiat | Kategori | yes, no. |

### 3.2. Teknik Analisis

#### 3.2.1. Kapasitas data pengujian

Dataset yang digunakan pada penelitian ini ada 3 buah dataset seperti pada tabel dibawah ini.

**Tabel 3.4 Tabel Data Set**

| No | Dataset | Instances/Record | Atribut | Size |
|---|---|---|---|---|
| 1 | *Chronic Kidney Disease* | 400 | 25 | 44 kb |
| 2 | *Breast Cancer* | 286 | 10 | 19 kb |

#### 3.2.2. Hasil Pengukuran Dan Pembanding

Pada penelitian ini akan dibandingkan hasil nilai akurasi dari algoritma *Support Vector Machine*, *Decision Trees*, dan *Naive Bayes* dengan 2 buah dataset. Mode pengujian yang dilakukan pada penelitian ini adalah *Cross validation 10*, yang artinya akan dilakukan pengolahan data sebanyak 10 kali dan mengulangi (men-iterasi) experimennya sebanyak 10 kali juga. Hasil nilai akurasi akan ditulis pada tabel perbandingan berikut:

**Tabel 3.5 Form Perbandingan Nilai Akurasi**

| ALGORITMA | DATA SET | |
|---|---|---|
| | ***CHRONIC KIDNEY DISEASE*** | ***BREAST CANCER*** |
| *Support Vector Machine* | | |
| *Decision Tree* | | |
| *Naive Bayes* | | |

## 4. HASIL DAN PEMBAHASAN

### 4.1 Hasil

#### *4.1.1* Hasil Akurasi Algoritma *Support Vector Machine*

Pada algoritma kalsifikasi *Support Vector Machine* yang dilakukan pengukuran akurasi menggunakan aplikasi Weka dengan metode *cross validation* dan diperoleh nilai akurasi sebagai berikut:

*a.* *Support Vector Machine* dengan dataset *Breast Cancer*

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

Setelah dilakukan pengolahan data mining pada aplikasi Weka dengan algoritma *Support Vector Machine* diperoleh nilai akurasi sebesar 76.57 % seperti terlihat pada gambar dibawah ini.

```
=== Summary ===

Correctly Classified Instances         219               76.5734 %
Incorrectly Classified Instances        67               23.4266 %
Kappa statistic                          0.2241
Mean absolute error                      0.2343
Root mean squared error                  0.484
Relative absolute error                 64.4385 %
Root relative squared error            113.6824 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC
                 0.922     0.735     0.801      0.922     0.857      0.2
                 0.265     0.078     0.514      0.265     0.350      0.2
Weighted Avg.    0.766     0.579     0.733      0.766     0.736      0.2

=== Confusion Matrix ===

   a   b   <-- classified as
 201  17 |   a = no
  50  18 |   b = yes
```

Gambar 4.1 Hasil akurasi algoritma SVM pada dataset *Breast Cancer*

Dari Gambar di atas kita dapat melihat bahwa dari 286 *instances*, sebanyak 219 *instances* termasuk Correctly Classified, dan 67 instances termasuk Incorrectly Classified. *Confusion matrix* yang diperoleh jika dibuatkan tabel bisa terlihat pada tabel berikut:

**Tabel 4.1 Tabel *confusion matrix Support Vector Machine***

| | Prediksi (yes) | Prediksi (no) |
|---|---|---|
| Aktual (yes) | 201 | 17 |
| Aktual (no) | 50 | 18 |

Dari tabel di atas bisa diketahui bahwa:
TP = 201
FN = 17
FP = 50
TN = 18
Nilai akurasi = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{201+18}{286} = \frac{219}{286} = 0.765734$
Jadi nilai akurasi yang diperoleh sebesar 76.57 %.

*b.* *Support Vector Machine* dengan dataset *Chronic Kidney Disease*
Setelah dilakukan pengolahan data mining pada aplikasi Weka dengan algoritma *Support Vector Machine* diperoleh nilai akurasi sebesar 97.75%

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

seperti terlihat pada gambar dibawah ini

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

```
=== Summary ===

Correctly Classified Instances         391               97.75    %
Incorrectly Classified Instances         9                2.25    %
Kappa statistic                          0.9526
Mean absolute error                      0.0225
Root mean squared error                  0.15
Relative absolute error                  4.7982 %
Root relative squared error             30.9838 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC
                 0.964     0.000     1.000       0.964   0.982      0.954
                 1.000     0.036     0.943       1.000   0.971      0.954
Weighted Avg.    0.978     0.013     0.979       0.978   0.978      0.954

=== Confusion Matrix ===

   a   b   <-- classified as
 241   9 |   a = ckd
   0 150 |   b = notckd
```

Gambar 4.2 Hasil akurasi algoritma SVM pada dataset *Chronic Kidney Disease*

Dari Gambar di atas kita dapat melihat bahwa dari 400 *instances*, sebanyak 319 *instances* termasuk Correctly Classified, dan 9 instances termasuk Incorrectly Classified. *Confusion matrix* yang diperoleh jika dibuatkan tabel bisa terlihat pada tabel berikut:

**Tabel 4.2 Tabel confusion matrix Support Vector Machine**

| | Prediksi (yes) | Prediksi (no) |
|---|---|---|
| Aktual (yes) | 241 | 9 |
| Aktual (no) | 0 | 150 |

Dari tabel di atas bisa diketahui bahwa:
TP = 241
FN = 9
FP = 0
TN = 150
Nilai akurasi = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{241+150}{400} = \frac{391}{400} = 0.9775$
Jadi nilai akurasi yang diperoleh sebesar 97.75 %.

### *4.1.2* Hasil Akurasi Algoritma *Decision Tree*

Pada algoritma kalsifikasi *Decision Tree* yang dilakukan pengukuran akurasi menggunakan aplikasi Weka dengan metode *cross validation* dan diperoleh nilai akurasi sebagai berikut:

*a.* *Decision Tree* dengan dataset *Breast Cancer*

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

Setelah dilakukan pengolahan data mining pada aplikasi Weka dengan algoritma *Decision Tree* diperoleh nilai akurasi sebesar 75.87 % seperti terlihat pada gambar dibawah ini.

```
=== Summary ===

Correctly Classified Instances         217               75.8741 %
Incorrectly Classified Instances        69               24.1259 %
Kappa statistic                          0.201
Mean absolute error                      0.3041
Root mean squared error                  0.4167
Relative absolute error                 83.6574 %
Root relative squared error             97.8831 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC A
                 0.917     0.750     0.797       0.917   0.853      0.217    0.628     0.798
                 0.250     0.083     0.486       0.250   0.330      0.217    0.628     0.446
Weighted Avg.    0.759     0.591     0.723       0.759   0.729      0.217    0.628     0.714

=== Confusion Matrix ===

   a   b   <-- classified as
 200  18 |   a = no
  51  17 |   b = yes
```

Gambar 4.3 Hasil akurasi algoritma DT pada dataset *Breast Cancer*

Dari Gambar di atas kita dapat melihat bahwa dari 286 *instances*, sebanyak 217 *instances* termasuk Correctly Classified, dan 69 instances termasuk Incorrectly Classified. *Confusion matrix* yang diperoleh jika dibuatkan tabel bisa terlihat pada tabel berikut:

**Tabel 4.3 Tabel confusion matrix Decision Tree**

| | Prediksi (yes) | Prediksi (no) |
|---|---|---|
| Aktual (yes) | 200 | 18 |
| Aktual (no) | 51 | 17 |

Dari tabel di atas bisa diketahui bahwa:
TP = 200
FN = 18
FP = 51
TN = 17
Nilai akurasi = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{200+17}{286} = \frac{217}{286} = 0.7587$
Jadi nilai akurasi yang diperoleh sebesar 75.87 %.

*b.* *Decision Tree* dengan dataset *Chronic Kidney Disease*

Setelah dilakukan pengolahan data mining pada aplikasi Weka dengan algoritma *Decision Tree* diperoleh nilai akurasi sebesar 99 % seperti terlihat pada gambar dibawah ini

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

```
=== Summary ===

Correctly Classified Instances         396               99      %
Incorrectly Classified Instances         4                1      %
Kappa statistic                          0.9786
Mean absolute error                      0.0225
Root mean squared error                  0.0807
Relative absolute error                  4.7995 %
Root relative squared error             16.6603 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Ar
                 0.996     0.020     0.988       0.996   0.992      0.979    0.999     1.000
                 0.980     0.004     0.993       0.980   0.987      0.979    0.999     0.999
Weighted Avg.    0.990     0.014     0.990       0.990   0.990      0.979    0.999     0.999

=== Confusion Matrix ===

   a   b   <-- classified as
 249   1 |   a = ckd
   3 147 |   b = notckd
```

Gambar 4.4 Hasil akurasi algoritma DT pada dataset *Chronic Kidney Disease*

Dari Gambar di atas kita dapat melihat bahwa dari 400 *instances*, sebanyak 396 *instances* termasuk Correctly Classified, dan 4 instances termasuk Incorrectly Classified. *Confusion matrix* yang diperoleh jika dibuatkan tabel bisa terlihat pada tabel berikut:

**Tabel 4.4 Tabel confusion matrix Decision Tree**

| | Prediksi (yes) | Prediksi (no) |
|---|---|---|
| Aktual (yes) | 249 | 1 |
| Aktual (no) | 3 | 147 |

Dari tabel di atas bisa diketahui bahwa:
TP = 249
FN = 1
FP = 3
TN = 147
Nilai akurasi = $\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{249+147}{400} = \frac{396}{400} = 0.99$
Jadi nilai akurasi yang diperoleh sebesar 99 %.

### *4.1.3* Hasil Akurasi Algoritma *Naive Bayes*

Pada algoritma kalsifikasi *Naive Bayes* yang dilakukan pengukuran akurasi menggunakan aplikasi Weka dengan metode *cross validation* dan diperoleh nilai akurasi sebagai berikut:

*a.* *Naive Bayes* dengan dataset *Breast Cancer*

Setelah dilakukan pengolahan data mining pada aplikasi Weka dengan algoritma *Naive Bayes* diperoleh nilai akurasi sebesar 74.83 % seperti terlihat pada gambar dibawah ini.

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

```
=== Summary ===

Correctly Classified Instances         214               74.8252 %
Incorrectly Classified Instances        72               25.1748 %
Kappa statistic                          0.2684
Mean absolute error                      0.2777
Root mean squared error                  0.4195
Relative absolute error                 76.3847 %
Root relative squared error             98.5193 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC A
                 0.858     0.603     0.820       0.858   0.839      0.270    0.760     0.904
                 0.397     0.142     0.466       0.397   0.429      0.270    0.760     0.442
Weighted Avg.    0.748     0.493     0.736       0.748   0.741      0.270    0.760     0.794

=== Confusion Matrix ===

   a   b   <-- classified as
 187  31 |   a = no
  41  27 |   b = yes
```

Gambar 4.5 Hasil akurasi algoritma NB pada dataset *Breast Cancer*

Dari Gambar di atas kita dapat melihat bahwa dari 286 *instances*, sebanyak 214 *instances* termasuk Correctly Classified, dan 72 instances termasuk Incorrectly Classified. *Confusion matrix* yang diperoleh jika dibuatkan tabel bisa terlihat pada tabel berikut:

**Tabel 4.5 Tabel confusion matrix *Naive Bayes***

| | Prediksi (yes) | Prediksi (no) |
|---|---|---|
| Aktual (yes) | 187 | 31 |
| Aktual (no) | 41 | 27 |

Dari tabel di atas bisa diketahui bahwa:
TP = 187
FN = 31
FP = 41
TN = 27
Nilai akurasi = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{187+27}{286} = \frac{214}{286} = 0.7483$
Jadi nilai akurasi yang diperoleh sebesar 74.83 %.

*b.* *Naive Bayes* dengan dataset *Chronic Kidney Disease*
Setelah dilakukan pengolahan data mining pada aplikasi Weka dengan algoritma *Naive Bayes* diperoleh nilai akurasi sebesar 95 % seperti terlihat pada gambar dibawah ini

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

```
=== Summary ===

Correctly Classified Instances         380               95       %
Incorrectly Classified Instances        20                5       %
Kappa statistic                          0.8961
Mean absolute error                      0.0479
Root mean squared error                  0.2046
Relative absolute error                 10.2125 %
Root relative squared error             42.2526 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC A:
                 0.920     0.000     1.000       0.920   0.958      0.901    1.000     1.000
                 1.000     0.080     0.882       1.000   0.938      0.901    1.000     1.000
Weighted Avg.    0.950     0.030     0.956       0.950   0.951      0.901    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 230  20 |   a = ckd
   0 150 |   b = notckd
```

Gambar 4.6 Hasil akurasi algoritma NB pada dataset *Chronic Kidney Disease*

Dari Gambar di atas kita dapat melihat bahwa dari 400 *instances*, sebanyak 380 *instances* termasuk Correctly Classified, dan 20 instances termasuk Incorrectly Classified. *Confusion matrix* yang diperoleh jika dibuatkan tabel bisa terlihat pada tabel berikut:

**Tabel 4.6 Tabel confusion matrix *Naive Bayes***

| | Prediksi (yes) | Prediksi (no) |
|---|---|---|
| Aktual (yes) | 230 | 20 |
| Aktual (no) | 0 | 150 |

Dari tabel di atas bisa diketahui bahwa:
TP = 241
FN = 9
FP = 0
TN = 150
Nilai akurasi = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{230+150}{400} = \frac{380}{400} = 0.95$
Jadi nilai akurasi yang diperoleh sebesar 95 %.

## 4.2 Pembahasan

Dari hasil penelitian di atas dapat dibuat tabel perbandingan hasil nilai akurasinya seperti berikut:

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

**Tabel 4.7 Tabel perbandingan hasil akurasi**

| ALGORITMA | DATA SET | | RATA-RATA |
|---|---|---|---|
| | ***BREAST CANCER*** | ***CHRONIC KIDNEY DISEASE*** | |
| *Support Vector Machine* | 76.57 % | 97.75% | 87.16 % |
| *Decision Tree* | 75.87 % | 99 % | 87.43 % |
| *Naive Bayes* | 74.83 % | 95 % | 84.92 % |

Pada penelitian ini penulis memfokuskan pada perbandingan nilai akurasi yang dilakukan algoritma klasifikasi *support verctor machine, decision tree,* dan *Naive Bayes*. Dalam penelitian ini ternyata pada algoritma *support verctor machine* dan *decesion tree* memiliki nilai rata-rata akurasi yang hampir sama. Pada algoritma *support verctor machine* memiliki nilai akurasi rata-rata sebesar 87.16 %, pada algoritma *decesion tree* memiliki nilai akurasi rata-rata sebesar 87.43 %, dan pada algoritma *Naive Bayes* memiliki nilai akurasi rata-rata sebesar 84.92 %,

## 5. KESIMPULAN DAN SARAN

### 5.1 Kesimpulan

Dari hasil penelitian yang membandingkan algoritma *Support Vector Machine, decision trees,* dan *Naive Bayes* pada dataset *Chronic Kidney Disease* dan *Breast Cancer* dapat disimpulkan beberapa hal sesuai tujuan dari penelitian ini, diantaranya:

1. Nilai akurasi rata-rata menggunakan algoritma decision tree adalah sebesar 87.43 % dan merupakan nilai akurasi tertinggi dari kedua algoritma lainnya.
2. Nilai akurasi rata-rata menggunakan algoritma *Support Vector Machine* adalah sebesar 87.16 %.
3. Nilai akurasi rata-rata menggunakan algoritma *Naive Bayes* adalah sebesar 84.92 %.

### 5.2 Saran

Dari hasil yang telah penulis bahas diatas, maka penulis mencoba Pada penelitian ini dapat diketahui kinerja algoritma klasifikasi *Support Vector Machine*, decision trees, dan *Naive Bayes*. Hasil penelitian ini mungkin akan mendorong penelitian lanjutan dalam pengolahan data mining. Beberapa saran terkait hasil penelitian ini dan penelitian selanjutnya diantaranya:

1. Penelitian ini bisa menjadi referensi untuk melakukan klasifikasi dengan melihat hasil akurasi tertinggi dari penelitian ini.
2. Penelitian ini berfokus pada perbandingan hasil akurasi sebagai pembanding kinerja dari optimasinya, dimungkinkan penelitian selanjutnya bisa membandingkan waktu proses atau kinerja lainnya dari hasil optimasi yang telah dilakukan.

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

## DAFTAR PUSTAKA

[1] A. B. Wibisono dan A. Fahrurozi, "Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner," *Jurnal Ilmiah Teknologi dan Rekayasa,* pp. 161-170, 2019.

[2] A. P. A. Rahayuningsih dan R. Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner," *Jurnal Khatulistiwa Informatika,* pp. 20-28, 2018.

[3] D. Prajarini, "Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit," *Informatic Journal,* pp. 137-141, 2016.

[4] D. Nofriansyah, Konsep Data Mining vs Sistem Pendukung Keputusan, Yogyakarta: Deepublish, 2012.

[5] Kusrini dan E. T. Lutfi, Algoritma Data Mining, Yogyakarta: Andi, 2009.

[6] D. Nurdiyah dan I. A. Muwakhid , "Perbandingan Support Vector Machine Dan K-Nearest Neighbor Untuk Klasifikasi Telur Fertil Dan Infertil Berdasarkan Analisis Texture GLCM," *Jurnal Transformatika,* pp. 29-34, 2016.

[7] D. R. D. Lestari, "Perbandingan Klasifikasi Beasiswa Toyota Astra Menggunakan K-Nearest Neighbor Classifier Dan Naïve Bayes Sebagai Penentu Metode Klasifikasi Pada Spk Penerimaan Beasiswa Toyota Astra," 2017.

[8] X. Li, L. Wang dan E. Sung, "AdaBoost with SVM-based component classifiers," *Engineering Applications of Artificial Intelligence,* pp. 785-795, 2008.

[9] J. Han, M. Kamber dan J. Pei, Data mining: Concepts and Techniques (3th ed.), Waltham: Elsevier, 2012.

[10] Kusnawi, "Pengantar Solusi Data Mining," *Seminar Nasional Teknologi 2007,* pp. 1-9, 2007.

[11] P. Rahmadi, N. K. Tachjar dan L. S. Istiyowati, "Extracting Features On Indonesian Rupiah Notes Using 2DPCA Algorithm For Forged Detection," *International Conference for Emerging Markets,* pp. 1-6, 2013.

**ISSN 2549-4805**
**Volume 5, Nomor 2, Jul - Okt 2020**

[12] E. Prasetyo, Data Mining: Konsep dan Aplikasi menggunakan Matlab, Yogyakarta: Andi, 2012.

[13] P. Flach, "Machine Learning: The Art and Science of Algorithms that Make Sense of Data".

[14] S. Fitri, "Perbandingan Kinerja Algoritma Klasifikasi Naïve Bayesian, Lazy-Ibk, Zero-R, Dan Decision Tree- J48," *DASI*, pp. 33-37, 2014.

[15] N. Ratama, "IMPLEMENTASI METODE FUZZY TSUKAMOTO UNTUK DETEKSI DINI AUTISME PADA BALITA BERBASIS ANDROID," vol. 3, no. 2, pp. 129–139, 2020, [Online]. Available: https://e-journal.stmiklombok.ac.id/index.php/jire/article/view/269.

[16] Munawaroh, "Penerapan Metode Fuzzy Inference System Dengan Algoritma Tsukamoto," *J. Inform. J. Pengemb. IT Poltek Tegal*, vol. 03, no. 02, pp. 184–189, 2018.