

# **DIPONEGORO SCIENCE COMPETITION 2018**



**Subtema: Kesehatan**

## **PENERAPAN DATA MINING DAN ALGORITMA POHON KEPUTUSAN PADA ANALISIS FAKTOR PENYEBARAN HIV/AIDS DI INDONESIA**

*Diusulkan oleh*

Muhammad Iqbal Rahmadhan Putra; 10215054; 2015  
Dianisti Saraswati; 15315079; 2015

Institut Teknologi Bandung

Bandung

2018

### LEMBAR PENGESAHAN DSC 2018

1. Judul Karya Tulis : Penerapan Data Mining dan Algoritma Pohon Keputusan pada Analisis Faktor Penyebaran HIV/AIDS di Indonesia
2. Ketua Karya Tulis Ilmiah
  - a. Nama Lengkap : Muhammad Iqbal Rahmadhan Putra
  - b. NIM : 10215054
  - c. Perguruan Tinggi / Fakultas : Institut Teknologi Bandung / FMIPA
  - d. Alamat Rumah dan No. HP. : Jln. Pelesiran No 152/56, Bandung
  - e. Alamat email : miqbalrp@gmail.com
3. Anggota Penulis : Dianisti Saraswati
4. Dosen Pendamping
  - a. Nama Lengkap dan Gelar : Acep Purqon S.Si., M.Si., Ph.D
  - b. NIP : 19740915 199903 1 004
  - c. Alamat Rumah dan No. HP. : Jl. Ganesha 10 Bandung, 081221636620

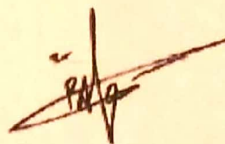
Bandung, 26 Juli 2018

Mengetahui,  
Guru Pembimbing



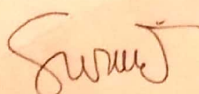
Acep Purqon S.Si., M.Si., Ph.D.  
NIP 19740915 199903 1 004

Ketua Pelaksana



Muhammad Iqbal Rahmadhan Putra  
NIM 10215054

Menyetujui,  
Wakil Dekan Kemahasiswaan



Dr. rer. nat. Sparisoma Viridi, M.Si

NIP 19731201 199903 1 002

## **LEMBAR PERNYATAAN ORISINALITAS**

Judul Karya Tulis : Penerapan Data Mining dan Algoritma Pohon  
Keputusan pada Analisis Faktor Penyebaran HIV/AIDS  
di Indonesia

Ketua Karya Tulis Ilmiah : Muhammad Iqbal Rahmadhan Putra

Anggota Karya Tulis Ilmiah : Dianisti Saraswati

Saya yang bertanda tangan di bawah ini menyatakan bahwa memang benar karya dengan judul tersebut merupakan karya orisinal dan belum pernah dipublikasikan/dilombakan diluar kegiatan Diponegoro Science Competition 2018.

Demikian pernyataan ini dibuat dengan sebenarnya dan apabila terbukti terdapat pelanggaran di dalamnya maka saya siap untuk didiskualifikasi dari kompetisi ini sebagai bentuk pertanggung jawaban saya.

Bandung, 28 Juli 2018

Ketua Karya Tulis



**Muhammad Iqbal Rahmadhan Putra**

NIM. 10215054

## Kata Pengantar

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa karena atas izinnya karya tulis ilmiah yang berjudul "Penerapan Data Mining dan Algoritma Pohon Keputusan pada Analisis Faktor Penyebaran HIV/AIDS di Indonesia" telah selesai dilakukan. Ucapan terimakasih kami sampaikan kepada Dosen Pembimbing serta pihak-pihak terkait penyelesaian karya tulis ini.

Karya tulis ini disusun dalam rangka mengikuti lomba karya tulis ilmiah (LKTI) Diponegoro Science Competition yang diselenggarakan oleh *Research Incubator Center* Fakultas Sains dan Matematika Universitas Diponegoro dengan tema "Inovasi Sains dan Teknologi demi tercapainya *Sustainable Development Goal* (SDGs) di Indonesia". Dari sekian sub tema yang disediakan oleh panitia, penulis memilih sub tema kesehatan dikarenakan pokok permasalahan kesehatan di Indonesia memiliki tantangan tersendiri, terutama terkait penyebaran HIV/AIDS.

Harapannya dengan makalah ini dapat menjadi solusi alternatif dalam pembuatan kebijakan terhadap pencegahan penyebaran HIV/AIDS dan membuka peluang untuk penelitian lebih lanjut.

Terakhir, kami mengucapkan terimakasih kepada panitia penyelenggara LKTI dan kiranya dapat memberikan penilaian yang baik pada makalah penelitian ini. Semoga kegiatan ini dapat berlanjut dan semakin baik kedepannya.

# Daftar Isi

<b>1</b>	<b>PENDAHULUAN</b>	<b>1</b>
1.1	Latar Belakang . . . . .	1
1.2	Rumusan Masalah . . . . .	2
1.3	Tujuan . . . . .	2
1.4	Manfaat . . . . .	2
<b>2</b>	<b>TINJAUAN PUSTAKA</b>	<b>4</b>
2.1	Faktor Penyebaran Epidemi HIV/AIDS . . . . .	4
2.2	Data Mining . . . . .	4
2.3	Algoritma Pohon Keputusan . . . . .	5
<b>3</b>	<b>METODOLOGI PENELITIAN</b>	<b>8</b>
<b>4</b>	<b>HASIL DAN PEMBAHASAN</b>	<b>11</b>
4.1	Deskripsi Statistik Data dan Prapemrosesan Data . . . . .	11
4.2	Hasil Implementasi . . . . .	11
4.3	Analisis Hasil . . . . .	13
<b>5</b>	<b>PENUTUP</b>	<b>16</b>
5.1	Kesimpulan . . . . .	16
5.2	Saran . . . . .	16

## Daftar Gambar

1	Tahapan dalam <i>knowledge discovery in database</i> . . . . .	5
2	Contoh penerapan pohon keputusan dalam klasifikasi resiko kredit .	6
3	Diagram alir metodologi penelitian . . . . .	8
4	Hasil pembentukan pohon keputusan . . . . .	12
5	Pohon Keputusan yang diperoleh berdasarkan kriteria <i>information gain</i> . . . . .	22
6	Pohon Keputusan yang diperoleh berdasarkan kriteria <i>information gain</i> . . . . .	22

## Daftar Tabel

1	Contoh struktur suatu himpunan data . . . . .	6
2	Deskripsi dan sumber data . . . . .	9
3	Deskripsi statistik variabel . . . . .	11
4	Contoh struktur suatu himpunan data . . . . .	12
5	Aturan yang dibentuk oleh Pohon Keputusan . . . . .	13
6	Himpunan data 34 provinsi di Indonesia . . . . .	21

## Abstrak

Mengakhiri epidemi AIDS pada tahun 2030 merupakan salah satu target dari tujuan ketiga (kesehatan yang baik) pada *Sustainable Development Goals* (SDGs). Namun target ini harus dihadapkan dengan data dari Kementerian Kesehatan RI yang menunjukkan terjadinya kenaikan laporan infeksi HIV tiap tahunnya dari tahun 2010 hingga tahun 2016. Penanganan penyebaran HIV/AIDS merupakan persoalan kompleks yang dipengaruhi berbagai faktor baik langsung maupun tidak langsung yang menjadi penentu tinggi atau rendahnya penyebaran. Beberapa dari faktor penentu tersebut adalah tingkat kemiskinan, penggunaan narkoba, seks komersil, tingkat buta huruf dan ketersediaan dokter. Melalui penelitian ini dilakukan analisis kuantitatif terhadap faktor-faktor tersebut pada 34 provinsi di Indonesia menggunakan konsep *data mining* dan algoritma pohon keputusan. *Data mining* merupakan proses penggalian pengetahuan dari suatu basis data yang dapat berupa keteraturan, pola, atau hubungan sejumlah kumpulan data. Algoritma pohon keputusan merupakan teknik klasifikasi data yang memiliki atribut numerik dan kategorikal yang menghasilkan aturan-aturan berupa diagram berstruktur pohon untuk memprediksi suatu data baru. Pada penelitian ini, tiap provinsi yang diteliti akan dikelompokkan menjadi provinsi dengan prevalensi HIV tinggi dan rendah. Hasil dari penelitian ini adalah terbentuknya aturan yang menjelaskan hubungan antar faktor sehingga menentukan tinggi atau rendahnya prevalensi pada tiap provinsi serta dapat diketahui signifikansi dari masing-masing faktor. Diharapkan hasil dari penelitian ini dapat dijadikan rekomendasi terhadap pembentukan kebijakan yang efektif dalam penanganan kasus penyebaran HIV/AIDS agar tercapainya Indonesia bebas epidemi AIDS pada tahun 2030.

**Kata kunci** - HIV/AIDS, data mining, algoritma pohon keputusan, prevalensi

# PENDAHULUAN

## 1.1 Latar Belakang

Berdasarkan laporan perkembangan HIV/AIDS dan PIMS di Indonesia Triwulan IV Tahun 2016 oleh Kementerian Kesehatan RI diketahui bahwa jumlah HIV yang dilaporkan terjadi peningkatan di tiap tahunnya dari jumlah kasus sebelum tahun 2005 sebanyak 859 kasus dan pada tahun 2016 tercatat 41.250 kasus sehingga total kasus yang telah dilaporkan adalah 232.323 kasus HIV. Sedangkan untuk kasus AIDS yang dilaporkan pada tahun 2005 adalah sebanyak 5.239 kasus dan pada tahun 2016 adalah sebanyak 7.491 kasus.

Melakukan upaya penanggulangan penyebaran epidemi HIV/AIDS bukan hanya menjadi permasalahan Indonesia. Secara khusus, mengakhiri epidemi AIDS tercantum dalam Tujuan Pembangunan Berkelanjutan (*Sustainable Development Goals*, SDGs) yang disahkan oleh Perserikatan Bangsa-bangsa. Keinginan untuk mengakhiri epidemi AIDS bergabung bersama beberapa isu terkait penyakit menular lain dalam tujuan ketiga, yakni kesehatan dan kesejahteraan yang baik.

Namun dengan melihat kondisi epidemi HIV/AIDS di Indonesia yang telah disebutkan di atas, memunculkan pertanyaan tentang kesanggupan Indonesia untuk mencapai SDGs pada tahun 2030. Upaya pemerintah dalam melakukan penanggulangan telah dilakukan dengan berbagai cara. Sejak pertama kali ditemukannya HIV/AIDS pada tahun 1987, telah muncul lebih dari 400 kebijakan pemerintah dari skala internasional hingga skala kabupaten/kota. Kebijakan pengendalian HIV/AIDS di Indonesia mengacu pada kebijakan global *Getting to Zeros*, yakni:

1. Menurunkan hingga meniadakan infeksi baru HIV;
2. Menurunkan hingga meniadakan kematian yang disebabkan oleh keadaan yang berkaitan dengan AIDS;
3. Meniadakan diskriminasi terhadap ODHA;

Akan tetapi, patut diingat bahwa epidemi HIV/AIDS merupakan persoalan kompleks yang dipengaruhi oleh banyak faktor, baik faktor langsung maupun faktor tidak langsung seperti kondisi ekologi objek. Oleh karena itu, pada penelitian ini kami melakukan analisis terhadap beberapa faktor yang mempengaruhi besar dan kecilnya penyebaran epidemi HIV/AIDS di Indonesia.

Analisis ini dilakukan menggunakan konsep *data mining* dengan memanfaatkan data statistik di 34 provinsi yang berkaitan dengan faktor-faktor tersebut, kemudian diolah menggunakan algoritma Pohon Keputusan untuk mendapatkan suatu aturan



(*rule*) berbentuk pohon yang dapat menjelaskan bagaimana pengaruh faktor-faktor tersebut dalam menentukan tinggi atau rendahnya prevalensi epidemi HIV/AIDS di tiap provinsi di Indonesia. Algoritma Pohon Keputusan dipilih karena memiliki keunggulan yakni mudah untuk diinterpretasikan oleh manusia karena outputnya yang berupa diagram aturan, berbeda dengan algoritma lain yang dijelaskan dengan proses kotak hitam (*black box*).

Dengan demikian, diharapkan hasil dari penelitian ini dapat memberikan alternatif solusi untuk menentukan kebijakan yang efektif demi tercapainya SDGs pada tahun 2030.

## **1.2 Rumusan Masalah**

Permasalahan yang akan diselesaikan pada penelitian ini adalah sebagai berikut:

1. Bagaimana faktor penyebaran HIV/AIDS di Indonesia?
2. Bagaimana implementasi Algoritma Pohon Keputusan pada analisis faktor penyebaran HIV/AIDS di Indonesia?
3. Bagaimana interpretasi dari aturan yang dibentuk oleh Algoritma Pohon Keputusan?

## **1.3 Tujuan**

Tujuan dari penelitian ini adalah sebagai berikut:

1. Melakukan pengambilan data terhadap faktor-faktor penyebaran HIV/AIDS di tiap provinsi di Indonesia.
2. Melakukan pembentukan pohon keputusan berdasarkan data yang telah diperoleh menggunakan algoritma C4.5
3. Melakukan interpretasi terhadap aturan pohon keputusan yang telah dibentuk.

## **1.4 Manfaat**

Setelah dilakukan penelitian ini, maka diharapkan mendapatkan manfaat sebagai berikut:

- Diketuinya faktor-faktor yang mempengaruhi penyebaran HIV/AIDS di Indonesia secara umum

- Hasil berupa aturan pohon keputusan dapat dijadikan acuan dalam menentukan kebijakan pemerintah yang efektif dalam penanggulangan HIV/AIDS
- Membuka peluang untuk penelitian yang lebih lanjut baik dengan metode yang sama atau dengan metode yang berbeda.

# TINJAUAN PUSTAKA

## 2.1 Faktor Penyebaran Epidemi HIV/AIDS

Telah dikumpulkan beberapa riset yang memberikan faktor-faktor yang mempengaruhi penyebaran HIV/AIDS. Mondal [1] dalam risetnya menganalisis empat faktor dengan analisis korelasi statistik terhadap 163 negara. Empat faktor tersebut diantaranya adalah prevalensi pengguna kontrasepsi, jumlah dokter terhadap populasi, proporsi dari populasi Muslim, kehamilan saat remaja, dan rata-rata waktu belajar penduduk.

Sedangkan Januaris melakukan penelitian sepuluh faktor mayor yang berkontribusi terhadap penyebaran HIV/AIDS pada negara berkembang. Sepuluh faktor tersebut adalah pergaulan bebas [2], ketidakpedulian [3], buta huruf [3], kemiskinan [4], penggunaan narkoba dan minuman beralkohol [5], stigma terhadap HIV/AIDS [6], faktor budaya, kekurangan terhadap layanan kehamilan, konflik suku dan perang saudara, serta faktor penduduk yang bermigrasi [7].

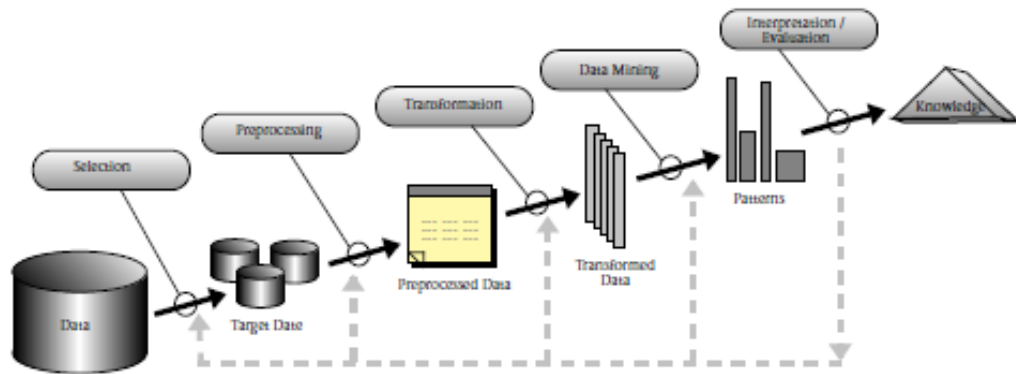
## 2.2 Data Mining

*Data Mining* (DM) merupakan suatu langkah analisis untuk mendapatkan pengetahuan (*knowledge*) di dalam suatu basisdata yang biasa disebut sebagai *Knowledge Discovery in Database* [8]. Pengetahuan yang dimaksud dapat berupa pola data atau relasi antar data yang valid (yang tidak diketahui sebelumnya). Tahapan-tahapan dalam *Knowledge Discovery in Database* terdiri dari *data cleaning*, *data integration*, *data selection*, *data transformation*, *data mining*, *pattern evaluation*, dan *knowledge presentation*. Urutan dari tahap ini dapat dilihat pada gambar dibawah.

Terlihat bahwa *data mining* merupakan satu bagian dari proses *knowledge discovery in database*. *Data Mining* melibatkan integrasi dari banyak disiplin keilmuan seperti teknologi basisdata dan *warehousing*, statistika, *machine learning*, komputasi tingkat tinggi, pengenalan pola, jaringan syaraf, visualisasi data, pemrosesan citra dan sinyal, serta analisis data spasial dan temporal [9]. Aplikasi dari *data mining* dapat ditemukan di berbagai bidang, seperti pemasaran, bisnis, sains dan teknologi, ekonomi, *games* dan bioinformatika.

Berdasarkan fungsinya, *data mining* dikelompokkan menjadi enam kelompok [8] [10] sebagai berikut:

- Klasifikasi, yakni melakukan generalisasi struktur yang diketahui untuk diaplikasikan pada data-data baru.



Gambar 1: Tahapan dalam *knowledge discovery in database*

- Klasterisasi, yakni mengelompokkan data ke dalam sejumlah kelompok tertentu sesuai dengan kemiripannya.
- Regresi, menemukan suatu fungsi yang memodelkan data dengan galat semaksimal mungkin.
- Deteksi anomali, yakni mengidentifikasi data yang tidak umum, misalnya berupa pencilan.
- Pembelajaran aturan asosiasi atau pemodelan ketergantungan
- Perangkuman, menyediakan representasi data yang lebih sederhana, meliputi visualisasi dan pembuatan laporan.

Kumpulan data yang diolah pada *data mining* disebut sebagai himpunan data (*data set*). Himpunan data dibangun dari objek-objek data, dimana objek data menyatakan sebuah entitas [10]. Objek data digambarkan menggunakan atribut (dalam istilah lain disebut sebagai fitur, dimensi, atau variabel). Atribut suatu objek atribut nominal, biner, numerik, dan ordinal. Objek-objek data yang disimpan dalam suatu basis data disebut sebagai *tuple*, dimana baris menyatakan objek-objek data dan kolom adalah atribut. Objek-objek data tersebut dapat dikelompokkan menjadi beberapa kelas yang dapat dipengaruhi oleh nilai dari atribut. Contoh struktur suatu himpunan data dapat dilihat pada tabel 1.

### 2.3 Algoritma Pohon Keputusan

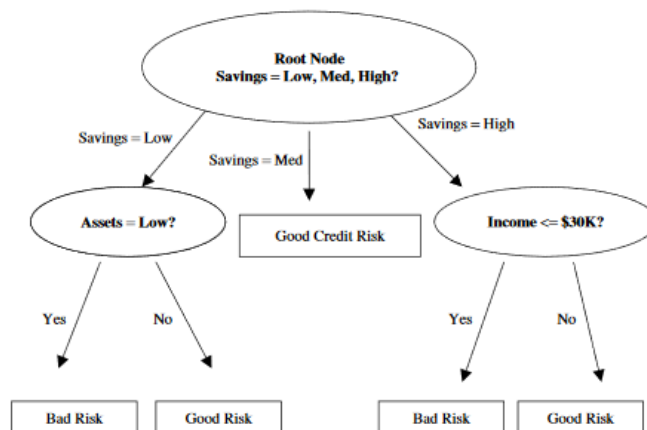
Salah satu algoritma dalam *data mining* yang digunakan untuk tujuan klasifikasi adalah Pohon Keputusan (*Decision Tree*). Pohon keputusan merupakan diagram

Tabel 1: Contoh struktur suatu himpunan data

Objek	Atribut 1	Atribut 2	Atribut 3	Kelas
Objek A	...	...	...	Kelas I
Objek B	...	...	...	Kelas I
Objek C	...	...	...	Kelas II
Objek D	...	...	...	Kelas II

alir yang membentuk struktur hirarki menyerupai struktur pohon, yang mana setiap *node* merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas [9]. *Node* paling atas pada pohon keputusan disebut sebagai *root*.

Pada gambar 2 ditampilkan contoh hasil dari penerapan algoritma pohon keputusan dalam mengklasifikasi resiko kredit dari sejumlah data dengan tiga atribut berupa *savings*, *assets*, dan *income* serta target berupa *good risk* dan *bad risk* [11].



Gambar 2: Contoh penerapan pohon keputusan dalam klasifikasi resiko kredit

Terdapat beberapa jenis algoritma pohon keputusan, yaitu CART (*Classification and Regression Trees*, diusulkan oleh Breiman dkk pada 1984), ID3 (*Iterative Dichotomiser 3*), dan C4.5.

Untuk menentukan ukuran dari suatu atribut dapat digunakan dua macam ukuran/kriteria, yakni *information gain* dan *gini index*.

### Information Gain

Dengan menggunakan ukuran *information gain*, kita mencoba melakukan perkiraan besarnya informasi yang terkandung oleh suatu atribut, atau dengan kata lain *information gain* didefinisikan sebagai ukuran efektifitas suatu atribut dalam meng-

klasifikasikan data [10]. Misalkan  $S$  adalah suatu himpunan data dan  $A$  adalah atribut dengan nilai  $v$  Secara matematis *information gain* suatu atribut dituliskan

$$InformationGain(S, A) = Entropy(S) - \sum_v \frac{|S_v|}{|S|} [Entropy((S_v))] \quad (1)$$

dengan entropi didefinisikan sebagai suatu parameter untuk mengukur keberagaman dalam suatu himpunan data, semakain heterogen suatu himpunan data, maka semakin besar nilai entropinya secara matematis dirumuskan

$$Entropi(S) = - \sum_{j=1}^k p_j \log_2 p_j \quad (2)$$

di mana:

$S$  = himpunan kasus

$k$  = banyaknya partisi  $S$

$p_j$  = rasio jumlah sampel di kelas  $j$  dengan jumlah semua sampel pada himpunan data

## Gini Index

*Gini index* adalah ukuran yang menyatakan seberapa sering suatu elemen yang dipilih acak akan teridentifikasi dengan tidak tepat. Sehingga untuk *gini index*, atribut yang memiliki nilai lebih kecil akan lebih baik. Secara matematis, *gini index* dinyatakan sebagai berikut.

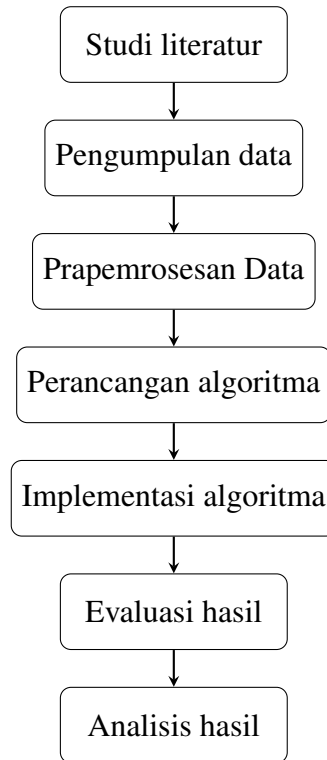
$$GiniIndex(S) = 1 - \sum_{j=1}^k p_j^2 \quad (3)$$

Untuk menentukan nilai *threshold* masing-masing atribut yang menempati *node* ditentukan dengan *threshold* yang memberikan nilai kriteria terbesar. Secara umum, algoritma untuk membangun Pohon Keputusan adalah sebagai berikut:

1. Tentukan atribut yang memiliki nilai kriteria terbesar sebagai *root node*
2. Buat cabang untuk tiap nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk setiap cabang hingga semua kasus memiliki kelas yang sama atau syarat kriteria terpenuhi.

## METODOLOGI PENELITIAN

Urutan pengerjaan dari penelitian ini ditampilkan sebagai berikut:



Gambar 3: Diagram alir metodologi penelitian

Hal pertama yang peneliti lakukan pada penelitian ini adalah melakukan studi terhadap literatur yang terkait untuk menambah pemahaman penulis akan topik dari penelitian. Studi literatur difokus mengenai faktor-faktor penyebaran HIV/AIDS di Indonesia, *data mining*, dan algoritma pohon keputusan. Hasil dari studi literatur dapat dilihat pada bagian Tinjauan Pustaka.

Berdasarkan hasil dari studi literatur terhadap faktor-faktor penyebaran HIV/AIDS, dilakukan pengumpulan data statistik dari masing-masing faktor di tiap provinsi di Indonesia. Data statistik yang berhasil dikumpulkan dan mencukupi untuk diolah ditampilkan pada tabel 2.

Dari data tersebut kemudian dicari karakteristik statistiknya (mean, median, standar deviasi, dll) sebelum dilakukan prapemrosesan data. Prapemrosesan data untuk menjadikan data siap untuk diolah menggunakan algoritma. Salah satu kelebihan algoritma pohon keputusan adalah tidak diperlukannya normalisasi data, sehingga pada prapemrosesan data tidak perlu dilakukan normalisasi seperti pada umumnya. Hal yang perlu dilakukan adalah menjadi data terbagi menjadi dua kelas

Tabel 2: Deskripsi dan sumber data

Variabel	Deskripsi	Sumber
Prevalensi HIV	Jumlah pengidap HIV tiap 100.000 penduduk pada tahun 2014	Kemenkes[12]
Persentase penduduk miskin	Jumlah penduduk miskin terhadap jumlah total penduduk pada tahun 2014	BPS[13]
Prevalensi pengguna narkoba	Jumlah pengguna narkoba dan obat-obatan terlarang tiap 100.000 penduduk yang berusia 10-59 tahun pada tahun 2014	BNN[14]
Jumlah lokalisasi	Banyaknya lokalisasi prostitusi yang belum ditutup dalam provinsi tahun 2014	Kemensos
Persentase penduduk buta huruf	Jumlah penduduk yang buta huruf terhadap total jumlah penduduk yang berusia 14-44 tahun pada tahun 2014	BPS[13]
Ketersediaan dokter	Jumlah dokter di Pusat Kesehatan Masyarakat yang tersedia tiap 100 orang penduduk	BPS[13]

berdasarkan nilai prevalensi HIV. Kelas ini terbagi menjadi kelas prevalensi tinggi dan prevalensi rendah yang dibagi mengacu pada median variabel target.

Selanjutnya adalah melakukan perancangan algoritma. Algoritma yang digunakan sesuai dengan hasil studi literatur. Digunakan dua macam kriteria penentuan *node*, yakni *information gain* dan *gini ratio*. Algoritma ini diimplementasikan menggunakan bahasa pemrograman Python 3.0 dengan menggunakan pustaka *scikit-learn*. Pustaka *scikit-learn* merupakan pustaka *machine learning* yang menyediakan beberapa algoritma klasifikasi, regresi, dan klusterisasi. Implementasi dari algoritma pohon keputusan pada Python 3.0 dapat dilihat di bawah (kode yang lengkap dapat dilihat di bagian lampiran).

```
import sklearn.datasets as datasets
from sklearn import tree
from sklearn.cross_validation import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
import graphviz

#Menggunakan algoritma dengan kriteria Gini Index
clf_gini = DecisionTreeClassifier(criterion = "gini",
random_state = 100, splitter = 'best',
min_impurity_split = 0.2, max_depth = 5
)
clf_gini.fit(X_train, Y_train)
```



```

dot_data_gini = tree.export_graphviz(clf_gini ,
out_file=None, feature_names=X.columns ,
filled=True , rounded=True ,
special_characters=True)
graph_gini = graphviz.Source(dot_data_gini)

#Melakukan pengujian dan mengukur akurasi
Y_pred = clf_gini.predict(X_train)
print("Accuracy is ", accuracy_score(Y_train , Y_pred)*100)

```

```

#Menggunakan algoritma dengan kriteria Gini Index
clf_entropy = DecisionTreeClassifier(criterion = "entropy" ,
random_state = 100, min_impurity_split = 0.5 ,
max_depth = 5
)
clf_entropy.fit(X_train , Y_train)
dot_data_entropy = tree.export_graphviz(clf_entropy ,
out_file=None, feature_names=X.columns ,
filled=True , rounded=True ,
special_characters=True)
graph_entropy = graphviz.Source(dot_data_entropy)

#Melakukan pengujian dan mengukur akurasi
Y_pred_entropy = clf_entropy.predict(X_train)
print("Accuracy is ", accuracy_score(Y_train ,
Y_pred_entropy)*100)

```

Setelah algoritma diimplementasikan, didapatkan hasil berdasarkan dua kriteria (*information gain* dan *gini ratio*). Dari dua macam hasil ini kemudian dievaluasi berdasarkan akurasi yang didapatkan. Hasil yang memiliki akurasi lebih tinggi digunakan untuk proses selanjutnya.

Proses terakhir yang dilakukan adalah analisis dari hasil yang didapatkan. Masing-masing atribut yang menempati *node* akan diteliti lebih jauh beserta kaitannya dengan kondisi aktual tiap provinsi maupun Indonesia secara umum.

## HASIL DAN PEMBAHASAN

### 4.1 Deskripsi Statistik Data dan Prapemrosesan Data

Himpunan data yang telah dikumpulkan terdiri dapat dilihat pada bagian lampiran. Himpunan data tersebut terdiri dari 34 provinsi yang memiliki 6 atribut (dimana Y, X1, X2, dst berturut-turut sesuai dengan tabel 3). Masing-masing atribut dicari deskripsi statistiknya seperti pada tabel 3. Dengan *N*, *min*, *maks*, *mean*, *med*, dan *SD* berturut-turut menyatakan jumlah data, data minimal, data maksimal, rata-rata, median, dan standar deviasi. Penjelasan masing-masing atribut dapat dilihat kembali pada tabel 2.

Tabel 3: Deskripsi statistik variabel

Variabel	N	Min	Maks	Mean	Med	SD
Prevalensi HIV	34	1.222	106.049	16.969	8.266	23.305
Persentase penduduk miskin	34	4.09	27.8	11.532	9.695	5.989
Prevalensi pengguna narkoba	34	1.21	5.01	2.032	1.865	0.704
Jumlah lokalisasi	34	0	53	5	1	10.57
Persentase penduduk buta huruf	34	0.08	28.5	1.873	0.61	4.814
Ketersediaan dokter	34	3.06	20.38	9.284	9.015	4.017

Dari tabel dapat diketahui bahwa atribut Prevalensi HIV memiliki rata-rata 16.969 dan median 8.266. Untuk membagi objek-objek data menjadi dua kelas, maka digunakan nilai median sebagai nilai ambang batas (*threshold*) sehingga untuk objek data dengan atribut Prevalensi HIV di bawah median dikelompokkan menjadi kelas "Prevalensi Rendah" dan selebihnya dikelompokkan menjadi kelas "Prevalensi Tinggi". Selanjutnya untuk dapat diolah menggunakan algoritma, maka untuk kelas "Prevalensi Rendah" dinyatakan dengan nilai 0 dan untuk kelas "Prevalensi Tinggi" dinyatakan dengan nilai 1. Pada tabel 4 ditampilkan cuplikan dari pengkategorian provinsi berdasarkan nilai Prevalensi HIV.

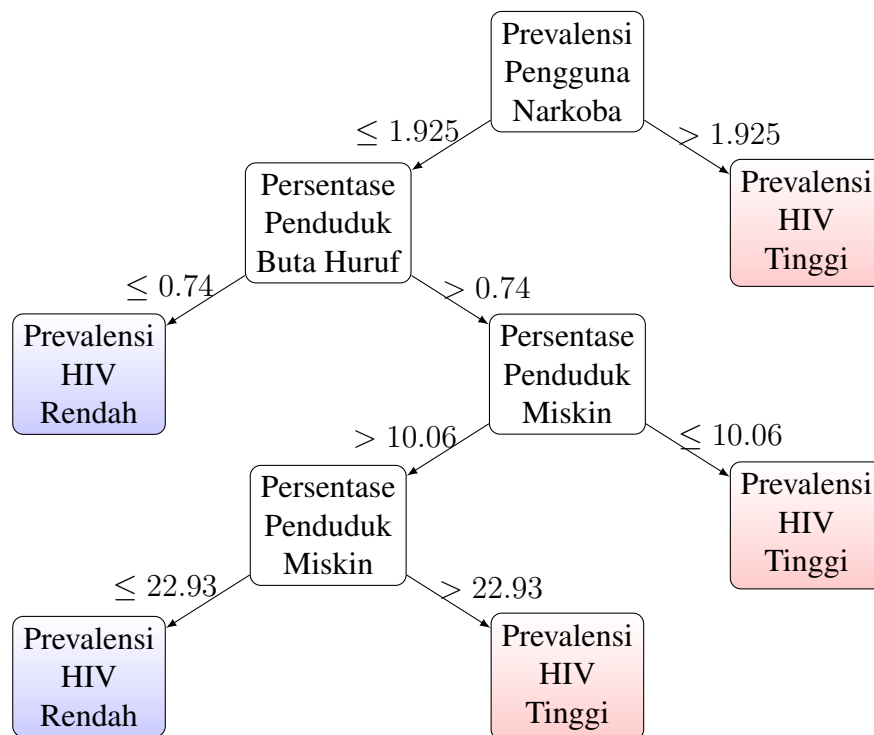
### 4.2 Hasil Implementasi

Pada bagian lampiran ditampilkan dua macam hasil Pohon Keputusan berdasarkan himpunan data yang telah dilakukan prapemrosesan. Pada masing-masing gambar juga ditampilkan nilai kriteria untuk masing-masing atribut. Untuk gambar 5 merepresentasikan Pohon Keputusan yang diperoleh menggunakan kriteria *information gain* sedangkan pada gambar 6 merepresentasikan Pohon Keputusan yang

Tabel 4: Contoh struktur suatu himpunan data

Nama Provinsi	Nilai Prevalensi HIV	Kategori Pravelensi HIV
Aceh	1.2228	Rendah
Sumatera Utara	11.8254	Tinggi
Sumatera Barat	6.25499	Rendah
...	...	...

dibentuk menggunakan kriteria *gini index*. Masing-masing hasil memperoleh nilai akurasi 97.06% untuk *information gain* dan 91.17% untuk *gini index*, sehingga bentuk pohon keputusan yang dipilih adalah kriteria *information gain*. Berdasarkan kriteria Pohon Keputusan tersebut, dibentuk Pohon Keputusan yang lebih sederhana seperti pada gambar 4.



Gambar 4: Hasil pembentukan pohon keputusan

Pada gambar 4 terlihat bahwa hanya tiga atribut dari total lima atribut yang berpengaruh terhadap tinggi atau rendahnya Prevalensi HIV di tiap provinsi. Tiga atribut itu adalah prevalensi pengguna narkoba, persentase penduduk buta huruf, dan persentase penduduk miskin. Sedangkan atribut yang tidak termasuk adalah jumlah lokalisasi prostitusi dan ketersediaan dokter. Aturan yang dibentuk berdasarkan

Pohon Keputusan dapat dilihat pada tabel 5.

Tabel 5: Aturan yang dibentuk oleh Pohon Keputusan

Kondisi	Hasil
Jika prevalensi pengguna narkoba $\leq 1.925$ , persentase penduduk buta huruf $\leq 0.74$	Maka prevalensi HIV rendah
Jika prevalensi pengguna narkoba $\leq 1.925$ , persentase penduduk buta huruf $> 0.74$ , persentase penduduk miskin diantara 10.06 dan 22.93	Maka prevalensi HIV rendah
Jika prevalensi pengguna narkoba $\leq 1.925$ , persentase penduduk buta huruf $> 0.74$ , persentase penduduk miskin $> 22.93$	Maka prevalensi HIV tinggi
Jika prevalensi pengguna narkoba $\leq 1.925$ , persentase penduduk buta huruf $> 0.74$ , persentase penduduk miskin $\leq 10.06$	Maka prevalensi HIV tinggi
Jika prevalensi pengguna narkoba $> 1.925$	Maka prevalensi HIV tinggi

### 4.3 Analisis Hasil

Pada bagian ini akan dilakukan analisis mendalam terhadap masing-masing atribut yang mempengaruhi tinggi atau rendahnya Prevalensi HIV.

#### Prevalensi pengguna narkoba

Atribut prevalensi pengguna narkoba menempati *root node* karena memiliki nilai kriteria paling tinggi untuk seluruh data baik kriteria *gini index* (0.5) maupun kriteria *information gain* (1.0). Nilai *threshold* dari atribut ini adalah 1.925. Melihat pada tabel 3 diketahui rata-rata atribut bernilai 2.032 yang berarti lebih besar 5.5% dari nilai *threshold*.

Atribut prevalensi pengguna narkoba diambil dari faktor yang diajukan oleh Liu bersama dengan penggunaan minuman beralkohol. Penggunaan narkoba menjadi faktor penting karena penggunaan jarum suntik bersama yang sangat rentan menjadi perantara penyebaran HIV. Selain itu, pengguna narkoba (dan minuman beralkohol) cenderung tidak mampu membuat keputusan yang bijak terhadap aktivitas seksualnya.

Prevalensi pengguna narkoba tertinggi berada di DKI Jakarta dengan nilai 5.01 dengan prevalensi HIV bernilai 58.072 (tiap 100.000 orang) yang merupakan tertinggi kedua setelah Papua. Untuk dapat mengurangi prevalensi HIV, berdasarkan aturan yang telah dibuat, maka DKI Jakarta dan provinsi lain yang memiliki prevalensi narkoba lebih besar dari 1.925 harus menurunkan hingga melewati nilai *tre-*

*shold*. Karena apabila tidak, maka dapat dipastikan tergolong pada prevalensi HIV tinggi.

### **Persentase penduduk buta huruf**

Atribut selanjutnya yang menempati *node* tingkat kedua adalah persentase penduduk buta huruf dengan nilai *threshold* sebesar 0.74. Untuk provinsi yang memiliki nilai persentase penduduk buta huruf yang lebih kecil dari 0.74 dapat dikatakan memiliki prevalensi HIV yang rendah, sedangkan untuk kondisi lebih besar dari 0.74 maka melewati syarat selanjutnya.

Kondisi buta huruf menurut Kelly & Bain memungkinkan masyarakat tidak tahu apa-apa mengenai cara penularan HIV serta tindakan pencegahannya. Selain itu, masyarakat yang buta huruf biasanya lebih mudah untuk dipengaruhi oleh keyakinan lama, mitos, dan kesalah-pahaman tentang penyakit-penyakit, terutama HIV/AIDS yang menyebabkan sulitnya penanganan apabila telah terkena penyakit tersebut.

Provinsi yang memiliki tingkat buta huruf tertinggi pada tahun 2014 berada di provinsi Papua dengan nilai 28.5, hal ini sesuai dengan nilai prevalensi HIV sebesar 106.04 yang menempati urutan pertama di antara provinsi lain.

### **Persentase penduduk miskin**

Atribut terakhir yang mempengaruhi adalah persentase penduduk miskin. Atribut ini muncul dua kali pada Pohon Keputusan yang dapat disederhakan pemahamannya dengan melihat pada tabel ???. Kondisi yang didapatkan menurut peneliti cukup berbeda dengan hipotesis umum dan menurut penelitian sebelumnya. Scott menyatakan bahwa kemiskinan merupakan penyebab dari munculnya pekerjaan seks komersial dan keterbatasan akses pada pendidikan. Namun jika ditelisik pada Pohon Keputusan memberikan kesimpulan yang cukup berbeda.

Ada tiga macam kondisi yang diperoleh yang mengacu pada persentase penduduk miskin. Kondisi pertama yakni nilai persentase penduduk miskin yang lebih kecil dari 10.06 menghasilkan Prevalensi HIV yang tinggi. Kondisi kedua apabila persentase penduduk miskin berada di antara 10.06 sampai 22.93 menghasilkan Prevalensi HIV yang rendah. Terakhir, kondisi ketiga adalah apabila persentase penduduk miskin lebih besar dari 22.93.

Dari ketiga kondisi tersebut yang sesuai dengan hipotesis dan literatur adalah kondisi ketiga (lebih besar dari 22.93). Seperti yang telah dijelaskan sebelumnya, hal ini mengakibatkan munculnya seks komersial serta kurangnya akses pada pendidikan.

Namun pada kondisi pertama yang menyatakan kondisi untuk provinsi dengan penduduk miskin yang cenderung lebih sedikit (dibawah 10.06) malah mengindikasikan Prevalensi HIV yang tinggi. Yang perlu diperhatikan adalah bahwa kondisi ini terpenuhi apabila telah melewati syarat sebelumnya yakni besarnya penduduk yang buta huruf. Berdasarkan fakta ini menunjukkan kesejahteraan yang tinggi namun tidak didukung dengan tingkat literasi cenderung memberikan ruang untuk penyebaran HIV.

Atribut ini hanya menyumbangkan satu kondisi untuk Prevalensi HIV rendah yakni untuk kondisi ekonomi di pertengahan, di antara 10.06 sampai 22.93. Secara statistik berarti menggambarkan provinsi yang memiliki ekonomi pertengahan cenderung lebih terbebas dari penyebaran HIV dibandingkan dua kondisi lainnya.

# PENUTUP

## 5.1 Kesimpulan

Berdasarkan uraian pada bagian-bagian sebelumnya, maka dapat ditarik kesimpulan sebagai berikut:

1. Secara umum terdapat faktor-faktor yang mempengaruhi penyebaran HIV seperti yang dijelaskan pada bagian Tinjauan Pustaka. Dari faktor-faktor tersebut, yang dianalisis pada penelitian ini adalah:
  - Persentase penduduk miskin
  - Prevelansi pengguna narkoba
  - Jumlah lokalisasi prostitusi
  - Persentase penduduk buta huruf
  - Ketersediaan dokter

Penjelasan dari data yang diperoleh berdasarkan faktor-faktor tersebut dapat dilihat pada tabel 2.

2. Telah dilakukan implementasi algoritma Pohon Keputusan terhadap himpunan data menggunakan dua macam kriteria dan hasil yang didapatkan menunjukkan kriteria *Information Gain* memiliki akurasi yang lebih tinggi. Pohon keputusan yang dibentuk dapat dilihat pada gambar 4 yang menunjukkan hanya tiga faktor yang berpengaruh yaitu sebagai berikut (diurutkan berdasarkan tingkat signifikannya):
  - (a) Prevalensi pengguna narkoba
  - (b) Persentase penduduk buta huruf
  - (c) Persentase penduduk miskin
3. Analisis dari hasil menunjukkan secara umum sesuai dengan hipotesis dari literatur, kecuali pada faktor Persentase Penduduk Miskin. Analisis secara lengkap dapat dilihat pada bagian 4.3.

## 5.2 Saran

Berikut merupakan beberapa saran yang diajukan setelah melakukan penelitian ini.

- Melalui penelitian dapat diketahui bahwa Algoritma Pohon Keputusan dapat diimplementasikan pada kasus analisis faktor-faktor penyebaran HIV/AIDS. Untuk penelitian lebih lanjut dapat diterapkan metode lain sebagai pembandingan.
- Penelitian ini menggunakan objek data berupa provinsi di Indonesia yang berjumlah 34 provinsi. Hal ini memberikan ketelitian yang terbatas. Oleh karena itu, untuk mendapatkan hasil yang lebih teliti pada penelitian selanjutnya dapat digunakan skala yang lebih kecil, misalnya skala kabupaten/kota sebagai objek data.



## Pustaka

- [1] M. Mondal and M. Shitan, "Factors affecting the hiv/aids epidemic: an ecological analysis of global data," *African health sciences*, vol. 13, no. 2, pp. 301–310, 2013.
- [2] K.-H. Choi, D. R. Gibson, L. Han, and Y. Guo, "High levels of unprotected sex with men and women among men who have sex with men: a potential bridge of hiv transmission in beijing, china," *AIDS education and Prevention*, vol. 16, no. 1: Special issue, pp. 19–30, 2004.
- [3] M. J. Kelly and B. Bain, *Education and HIV/AIDS in the Caribbean*. Ian Randle Publishers, 2005.
- [4] E. Scott, T. Simon, A. La Foucadeb, T. Karl, and K.-A. Gittens-Baynes, "Poverty, employment and hiv/aids in trinidad and tobago," *International Journal of Business and Social Science*, vol. 2, no. 15, 2011.
- [5] H. Liu, O. Grusky, X. Li, and E. Ma, "Drug users: a potentially important bridge population in the transmission of sexually transmitted diseases, including aids, in china," *Sexually transmitted diseases*, vol. 33, no. 2, pp. 111–117, 2006.
- [6] A. D. Grant and K. M. De Cock, "The growing challenge of hiv/aids in developing countries," *British Medical Bulletin*, vol. 54, no. 2, pp. 369–381, 1998.
- [7] G. C. Bond, J. Kreniske, I. Susser, J. Vincent, *et al.*, *AIDS in Africa and the Caribbean*. Westview Press, Inc., 1997.
- [8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.
- [9] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
- [10] Suyanto, *Data Mining: untuk klasifikasi dan klasterisasi data*. Penerbit Informatika, 2017.
- [11] D. Larose, "Discovering knowledge in data. new jersey: John willey & sons," *Inc. ISBN 0-471-66657-2*, 2005.
- [12] Kementerian Kesehatan RI, "Laporan perkembangan hiv-aids triwulan i tahun 2016," 2015.
- [13] Badan Pusat Statistik, "Statistik Indonesia 2015," 2015.
- [14] Pusat Penelitian Data dan Informasi Badan Narkotika Nasional, "Survei Nasional Penyalahgunaan Narkoba di 34 Provinsi Tahun 2017," 2017.

## Lampiran

### Kode Implementasi Algoritma Pohon Keputusan pada Python 3.0

Pada bagian ini dilakukan pengambilan himpunan data dari file spreadsheet bertipe .csv yang telah disiapkan terlebih dahulu. File disimpan dengan nama HIV\_python. Selanjutnya dilihat perhitungan statistik masing-masing atribut berupa nilai minimum, maksimum, rata-rata, median, dan standar deviasi. Hasil perhitungan dapat dilihat pada tabel 3.

```
import numpy as np
import pandas as pd
from sklearn.cross_validation import train_test_split

#Mengambil dan melihat penggalan data
df = pd.read_csv("D:/py_project/HIV_python.csv")
print(df.head())

#Melihat deskripsi statistik data
print(df.describe())
```

Setelah diketahui nilai yang dicari di atas, selanjutnya objek-objek data dibagi menjadi dua kelas berdasarkan atribut Prevalensi HIV. Untuk objek data yang di atas nilai median digolongkan menjadi Prevalensi HIV Tinggi dan yang di bawah nilai median menjadi Prevalensi HIV Rendah, secara berturut-turut disimpan dalam angka biner 1 dan 0. Variabel ini disimpan sebagai Prevalensi biner.

Pada bagian ini atribut-atribut disimpan dalam variabel X dan target disimpan dalam variabel Y. Selanjutnya diatur jumlah objek data sebagai data latih dan data uji. Pada penelitian ini, seluruh data menjadi data latih dan sekaligus data uji. Hal ini dikarenakan jumlah objek data yang tidak begitu besar sehingga dirasa tidak mencukupi untuk dilakukan *split*.

```
#Menyimpan data sebagai variabel
#dan mengubah variabel target menjadi biner
X=df.loc[:, "Persentase_Penduduk_Miskin": "Ketersediaan_Dokter"]
Ys df.loc[:, "Prevalensi_biner"]

#Melakukan pembagian data untuk data latih dan data uji
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,
test_size=0, random_state = 100)

#Menampilkan jumlah data latih dan data uji
print('Data_latih:', Y_train.size)
print('Data_uji:', Y_test.size)
```

Pada bagian ini dilakukan implementasi algoritma Pohon Keputusan menggunakan **scikit-learn**. Pertama menggunakan kriteria *Gini Index* lalu ditampilkan grafik. Kedua menggunakan kriteria *Information Gain/Entropy* lalu ditampilkan grafik.

```
import sklearn.datasets as datasets
from sklearn import tree
from sklearn.cross_validation import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
```

```

import matplotlib.pyplot as plt
import graphviz

#Menggunakan algoritma dengan kriteria Gini Index
clf_gini = DecisionTreeClassifier(criterion = "gini",
random_state = 100, splitter = 'best',
min_impurity_split = 0.2, max_depth = 5
)
clf_gini.fit(X_train , Y_train)
dot_data_gini = tree.export_graphviz(clf_gini ,
out_file=None, feature_names=X.columns ,
filled=True, rounded=True,
special_characters=True)
graph_gini = graphviz.Source(dot_data_gini)

#Melakukan pengujian dan mengukur akurasi
Y_pred = clf_gini.predict(X_train)
print("Accuracy is_", accuracy_score(Y_train , Y_pred)*100)

```

```
graph_gini
```

```

#Menggunakan algoritma dengan kriteria Gini Index
clf_entropy = DecisionTreeClassifier(criterion = "entropy",
random_state = 100, min_impurity_split = 0.5,
max_depth = 5
)
clf_entropy.fit(X_train , Y_train)
dot_data_entropy = tree.export_graphviz(clf_entropy ,
out_file=None, feature_names=X.columns ,
filled=True, rounded=True,
special_characters=True)
graph_entropy = graphviz.Source(dot_data_entropy)

#Melakukan pengujian dan mengukur akurasi
Y_pred_entropy = clf_entropy.predict(X_train)
print("Accuracy is_", accuracy_score(Y_train ,
Y_pred_entropy)*100)

```

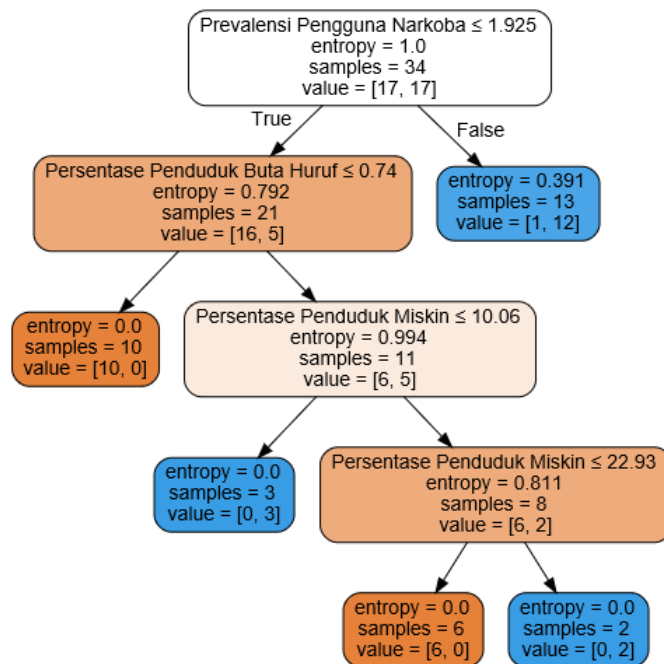
```
graph_entropy
```

## Tabel Himpunan Data yang Diteliti

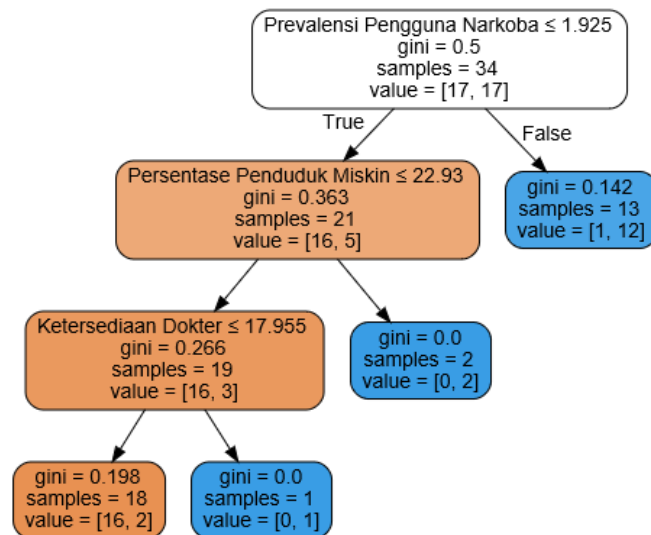
Tabel 6: Himpunan data 34 provinsi di Indonesia

Nama Provinsi	Y	X1	X2	X3	X4	X5
Aceh	1.2228	16.98	1.91	0	0.43	15.529
Sumatra Utara	11.825	9.85	3.2	1	0.66	10.307
Sumatra Barat	6.255	6.89	1.72	0	0.43	8.262
Riau	8.8876	7.99	1.97	9	0.48	7.7888
Jambi	5.0831	8.39	1.71	2	0.57	9.03
Sumatra Selatan	3.1732	13.62	1.74	1	0.52	5.5531
Bengkulu	4.987	17.09	1.62	1	0.54	10.408
Lampung	3.1896	14.21	1.24	3	0.42	6.3168
Kepulauan Bangka	8.4084	4.97	1.68	10	0.91	9.8222
Kepulauan Riau	50.746	6.4	2.77	10	0.38	16.533
DKI Jakarta	58.073	4.09	5.01	0	0.08	7.6127
Jawa Barat	8.1252	9.18	2.41	13	0.41	4.0843
Jawa Tengah	8.5524	13.58	1.94	3	0.65	5.5455
DI Yogyakarta	16.882	14.55	2.24	0	0.09	10.035
Jawa Timur	11.676	12.28	1.99	53	1.43	4.3175
Banten	5.8095	5.51	1.74	5	0.48	3.0585
Bali	51.865	4.76	2	3	1.06	7.7956
Nusa Tenggara Barat	3.1212	17.05	1.58	0	3.54	4.2524
Nusa Tenggara Timur	4.9435	19.6	1.43	0	3.48	6.2538
Kalimantan Barat	14.822	8.07	1.86	0	2.06	6.8701
Kalimantan Tengah	4.6313	6.07	1.86	12	0.32	11.599
Kalimantan Selatan	5.7867	4.81	1.89	0	0.28	8.9987
Kalimantan Timur	16.083	6.31	3.24	32	0.19	10.951
Kalimantan Utara	13.588	1.36	1.63	5	1.36	20.382
Sulawesi Utara	16.425	8.26	2.43	5	0.18	19.232
Sulawesi Tengah	4.6268	13.61	1.89	0	1.38	8.512
Sulawesi Selatan	9.95	9.54	2.25	0	2.58	7.4121
Sulawesi Tenggara	6.5357	12.77	1.5	0	1.62	9.9669
Gorontalo	2.1513	17.41	1.61	0	1.1	9.7705
Sulawesi Barat	2.3845	12.05	1.87	0	3.93	6.5178
Maluku	24.979	18.44	2.35	0	0.81	9.1107
Maluku Utara	5.5326	7.41	1.78	0	0.57	13.173
Papua Barat	70.605	26.26	1.52	0	2.27	9.2963
Papua	106.05	27.8	1.21	2	28.5	11.356

## Pohon Keputusan yang Dihasilkan melalui Dua Kriteria



Gambar 5: Pohon Keputusan yang diperoleh berdasarkan kriteria *information gain*



Gambar 6: Pohon Keputusan yang diperoleh berdasarkan kriteria *gini index*