**LAPORAN UAS DATA MINING KLASIFIKASI DECISION TREE MUSHROOM DATASET NEW**

****

Disusun Oleh :

Muhammad Syahdaffa - 21105133

Jefta Supraja - 2110511131

Heydar Emir Alvaro - 2110511150

Ahmad Surya Ramadhan - 2110511127

Muhammad Ghariza Pranaya 2110511126

**JURUSAN INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS PEMBANGUNAN VETERAN**

**JAKARTA**

**29 NOVEMBER**

**PENDAHULUAN**

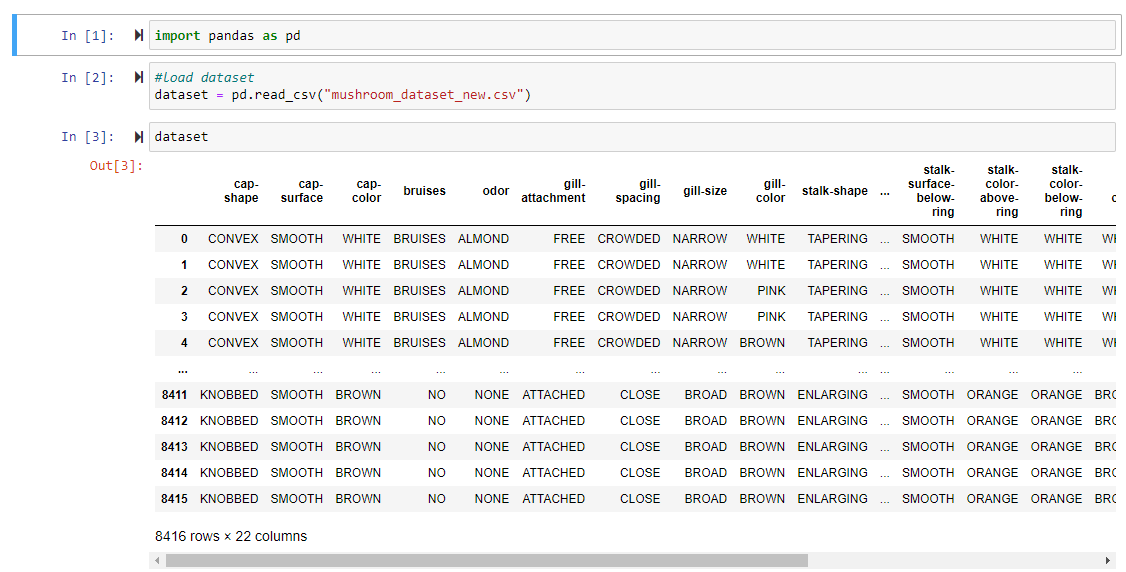
Data mining merupakan salah satu cabang ilmu komputer yang cukup banyak digunakan dan dipelajari oleh para ahli komputer dan para programmer. Data mining merupakan sebuah konsep yang diperuntukkan untuk menemukan pengetahuan atau informasi berharga yang tersembunyi di dalam database. Data mining adalah proses semi otomatik yang menerapkan matematika, teknik statistika, machine learning, dan kecerdasan buatan untuk menguraikan dan mengidentifikasi informasi pengetahuan dan berguna yang terkandung di dalam database besar.

Klasifikasi merupakan penempatan objek-objek ke salah satu dari beberapa kategori yang telah ditetapkan sebelumnya. Klasifikasi telah banyak ditemui dalam berbagai aplikasi. Sebagai contoh, pendeteksian pesan email spam berdasarkan header dan isi atau mengklasifikasikan galaksi berdasarkan bentuk bentuknya. Dalam bab ini akan dibahas mengenai konsep klasifikasi, beberapa isi penting dalam klasifikasi dan menyatakan metode untuk mengevaluasi dan membandingkan kinerja teknik klasifikasi. Klasifikasi adalah tipe analisis data yang dapat membantu orang menentukan kelas label dari sampel yang  ingin di klasifikasi. Klasifikasi merupakan Metode  supervised learning, metode yang mencoba menemukan hubungan antara atribut masukan dan atribut target. Tujuan klasifikasi untuk meningkatkan keandalan hasil yang diperoleh dari data.

Salah satu metode supervised machine learning yang digunakan untuk klasifikasi atau regresi adalah metode *decision tree.* Algoritma ini menghasilkan suatu model yang dapat memprediksi kategori data dengan cara mempelajari aturan penentuan kategori berdasarkan fitur-fitur yang dimiliki oleh data. Berdasarkan tipe kategori datanya decision tree dibedakan menjadi 2 jenis yaitu *classification tree* dan *regression tree.* Classification tree memiliki kategori berupa data diskrit  terhingga, sedangkan  regression tree memiliki kategori  berupa data diskrit berhingga atau data kontinu.

**PEMBAHASAN**

1. membaca data menjadi file csv



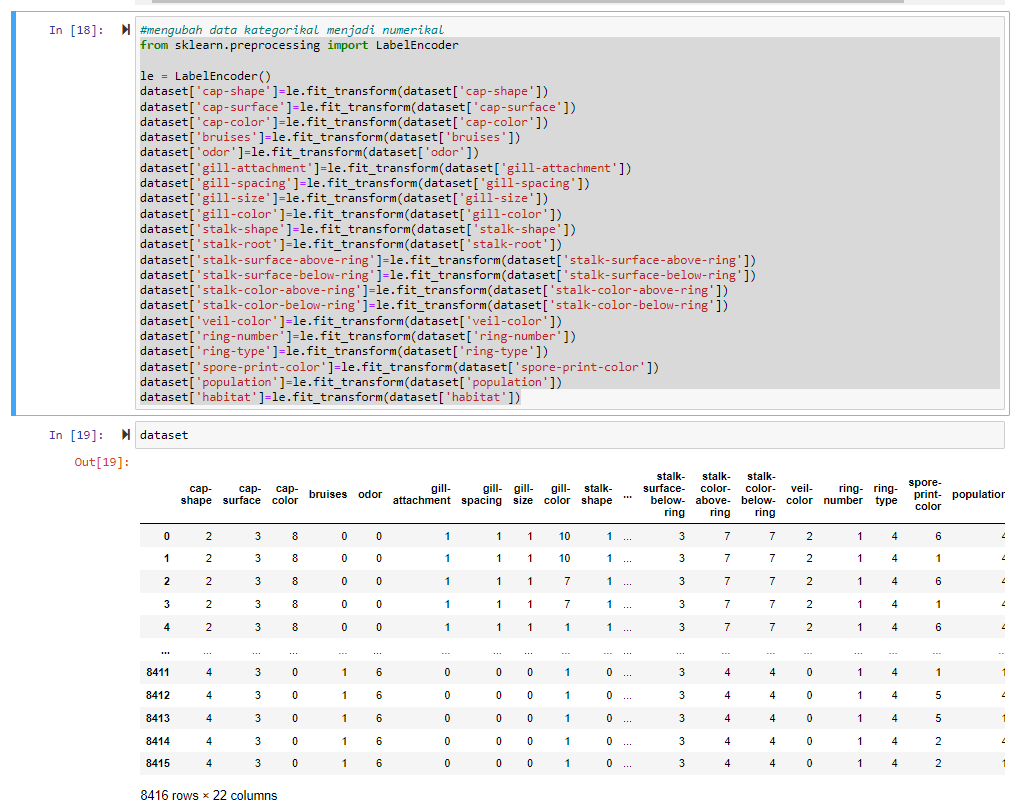
1. mengecek apakah ada data yang null



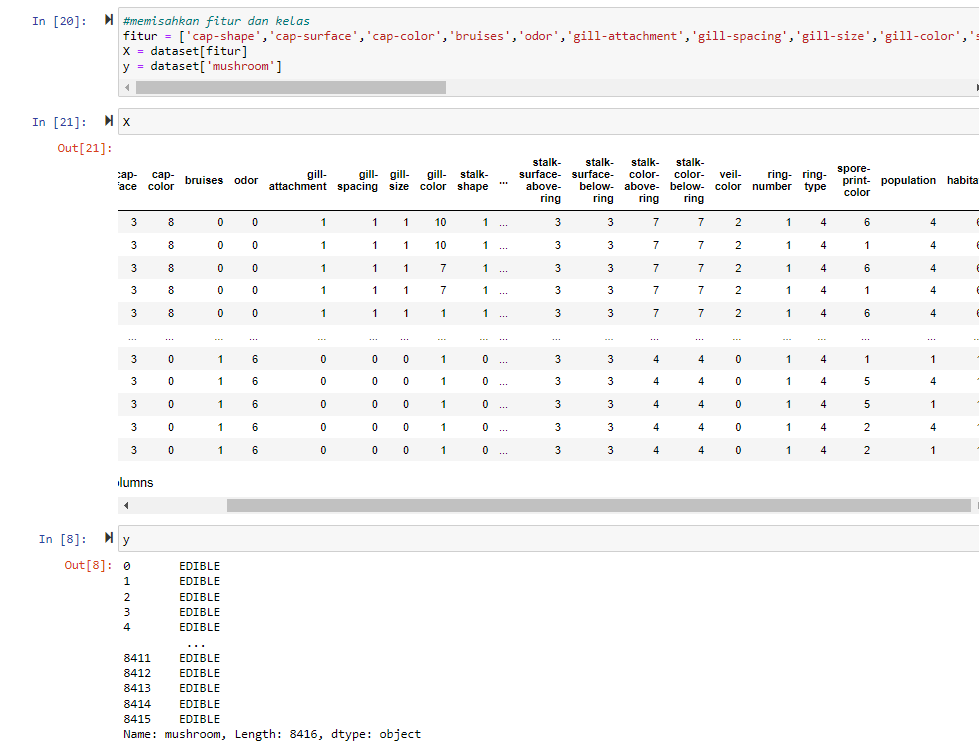
1. Menghandle missing values



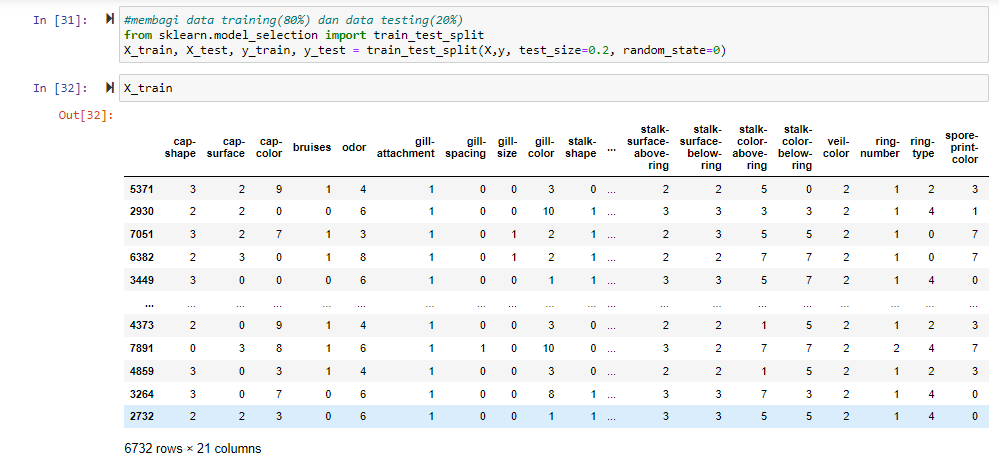
1. mengubah data menjadi numerik

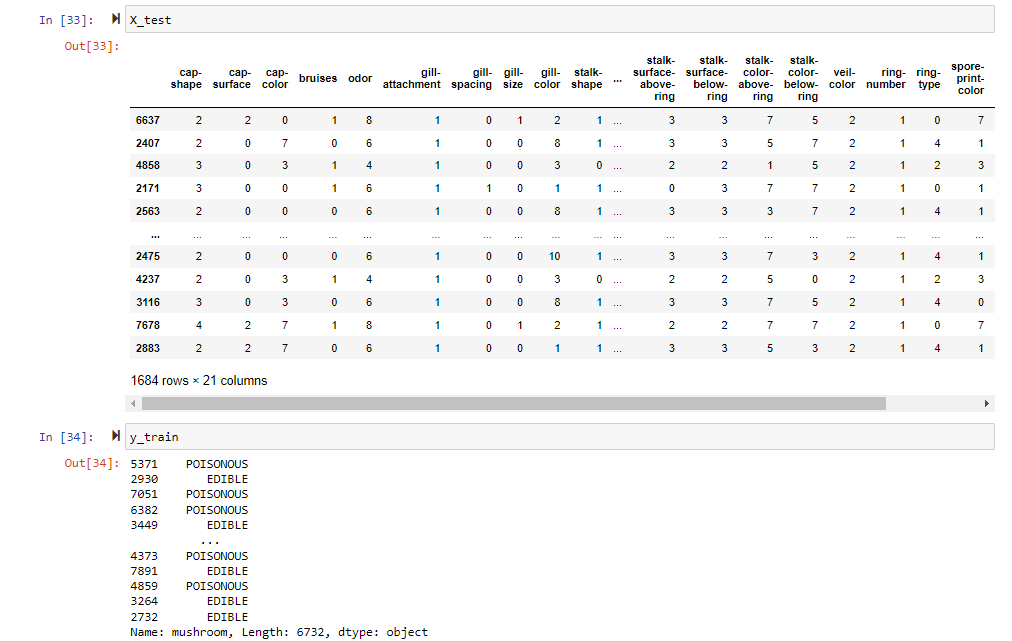


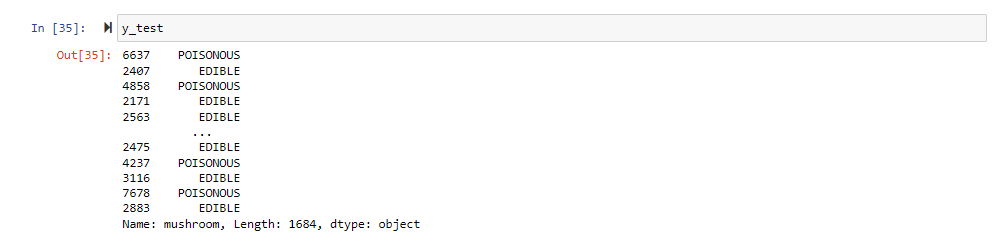
1. memisahkan fitur('cap-shape','cap-surface','cap-color','bruises','odor','gill-attachment','gill-spacing','gill-size','gill-color','stalk-shape','stalk-root','stalk-surface-above-ring','stalk-surface-below-ring','stalk-color-above-ring','stalk-color-below-ring','veil-color','ring-number','ring-type','spore-print-color','population','habitat') dan class (mushroom)



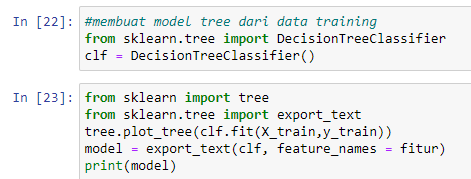
1. Memisahkan data training dan test 80 dan 20

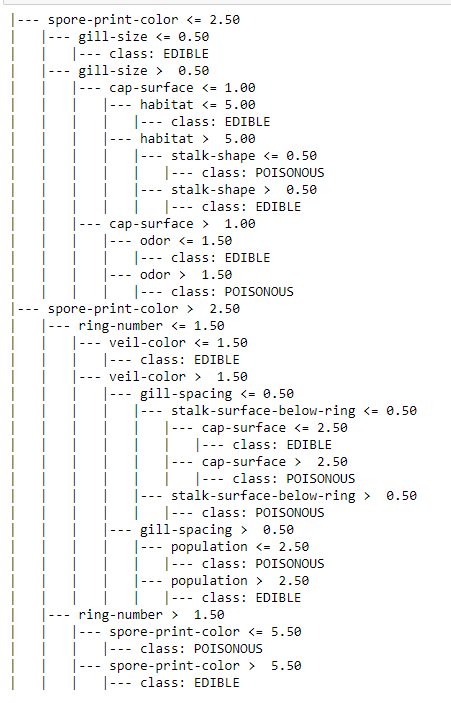


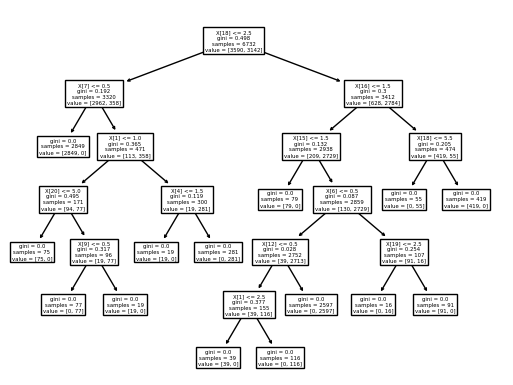




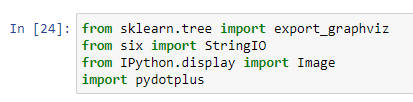
1. Membuat model tree dari data training

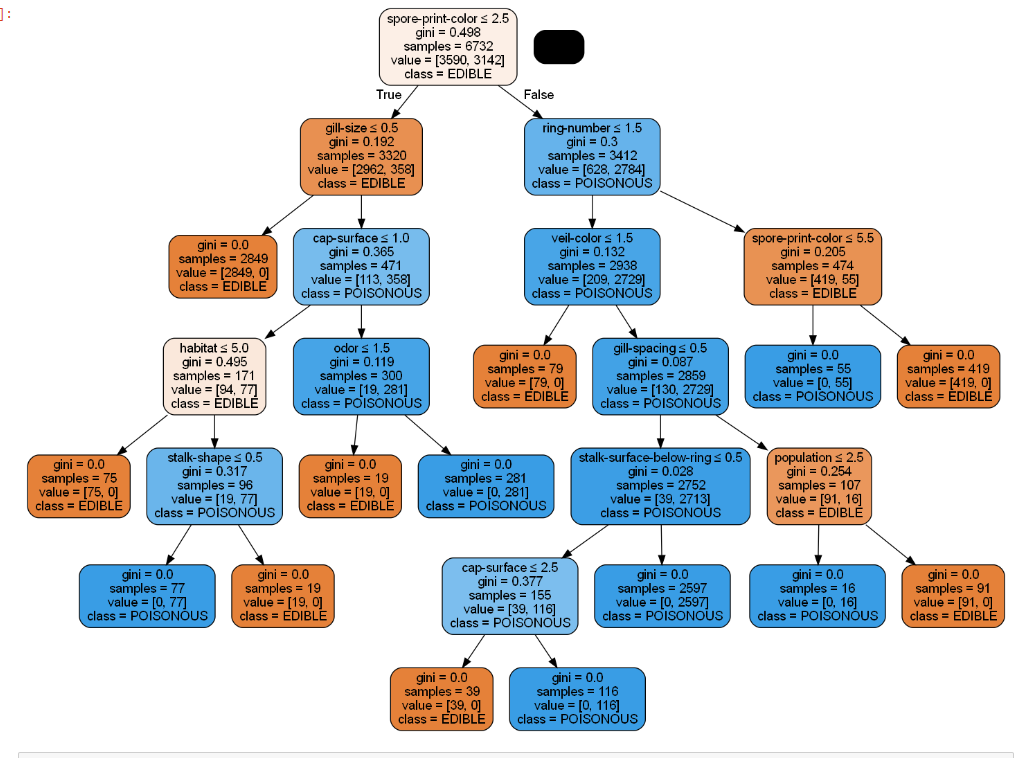
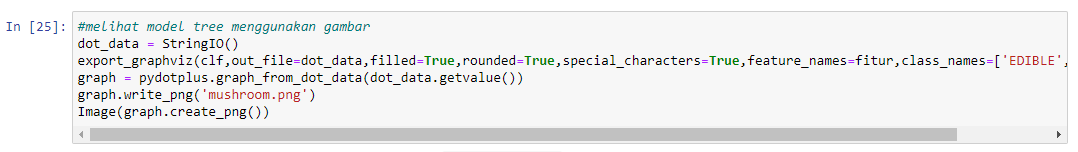




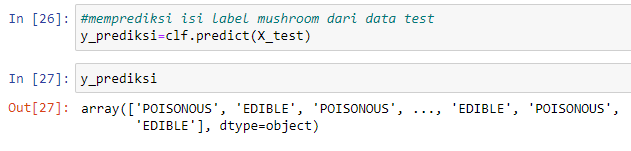


1. Melihat model tree menggunakan gambar (graphviz)

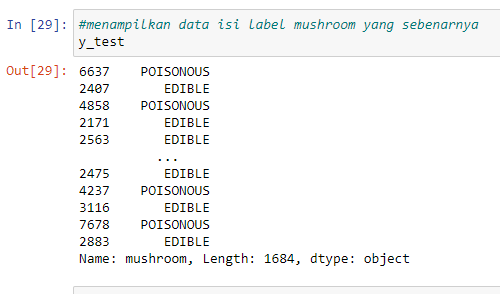




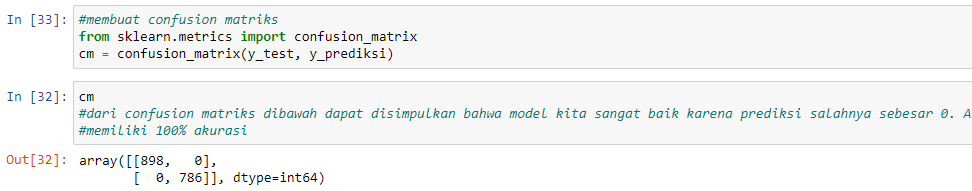
1. Memprediksi isi class mushroom dari data test (X\_test) menggunakan model yang tree yang sudah dibuat



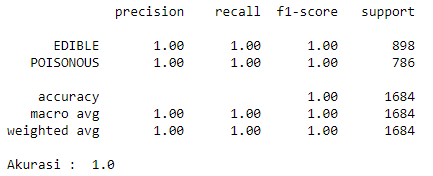
1. Menampilkan isi data class mushroom yang sebenarnya (y\_test)



1. Membuat confusion matriks



1. Menampilkan akurasi dari model decision tree yang sudah dibuat



Cara Perhitungan

* Mushroom (Y-test)

|  |  |  |  |
| --- | --- | --- | --- |
| Total Kasus | Sum (Edible) | Sum (Poisonous) | Entropi Total |
| 5 | 3 | 2 | 0.970951 |

S = (-( 3 / 5 ) . log2 3/5 ) + (-(2 / 5) . log2 2/5 = 0.970951

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Atribut | Nilai | SUM (Nilai) | SUM (Edible) | SUM (Poisonous) | Entropi | Gain |
| Cap – Shape | Convex | 3 | 2 | 1 | 0.918296 |  |
| Flat | 2 | 1 | 1 | 1 |
|  | | | | | 0.0199734 |
| Cap – Surface | Scaly | 1 | 0 | 1 | 0 |  |
| Fibrous | 4 | 3 | 1 | 0.811278 |
|  | | | | | 0.321929 |
| Cap – Color | Brown | 3 | 2 | 1 | 0.918296 |  |
| Red | 1 | 1 | 0 | 0 |
| Gray | 1 | 0 | 1 | 0 |
|  | | | | | 0.419973 |
| Bruises | No | 3 | 1 | 2 | 0.918296 |  |
| Bruises | 2 | 2 | 0 | 0 |
|  | | | | | 0.419973 |
| Odor | Spicy | 1 | 0 | 1 | 0 |  |
| None | 3 | 3 | 0 | 0 |
| Foul | 1 | 0 | 1 | 0 |
|  | | | | | 0.970951 |
| Gill – Attachment | Free | 5 | 3 | 2 | 0.970951 |  |
|  | | | | | 0 |
| Gill - Spacing | Close | 4 | 2 | 2 | 1 |  |
| Crowded | 1 | 1 | 0 | 0 |
|  | | | | | 0.170951 |
| Gill – Size | Narrow | 1 | 0 | 1 | 0 |  |
| Broad | 4 | 3 | 1 | 0.811278 |
|  | | | | | 0.321929 |
| Gill – Color | Buff | 1 | 0 | 1 | 0 |  |
| Purple | 2 | 2 | 0 | 0 |
| Chocolate | 1 | 0 | 1 | 0 |
| Brown | 1 | 1 | 0 | 0 |
|  | | | | | 0.970951 |
| Stalk – Shape | Tapering | 4 | 3 | 1 | 0.811278 |  |
| Enlarging | 1 | 0 | 1 | 0 |
|  | | | | | 0.321929 |
| Stalk – Root | Bulbous | 4 | 2 | 2 | 1 |  |
| Equal | 1 | 1 | 0 | 0 |
|  | | | | | 0.170951 |

* Analisis Atribut (X-test)

Contoh rumus Perhitungan entropy dan gain atribut

Cap – shape= -( 2 / 3 ) . log2 2 / 3 ) + (-(1 / 3) . log2 1/3 = 0.918296

Gain = 0.970951 – ((3/5) x 0.918296 + (2/5) x 1 = 0.0199734

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Stalk-surface-above-ring | Smooth | 3 | 2 | 1 | 0.918296 |  |
| Silky | 1 | 0 | 1 | 0 |  |
| Fibrous | 1 | 1 | 0 | 0 |  |
|  | | | | | 0.419973 |
| Stalk-surface-below-ring | Smooth | 4 | 3 | 1 | 0.811278 |  |
| Silky | 1 | 0 | 1 | 0 |  |
|  | | | | | 0.321929 |
| stalk-color-above-ring | White | 2 | 1 | 1 | 1 |  |
| Pink | 1 | 1 | 0 | 0 |  |
| Buff | 1 | 0 | 1 | 0 |  |
| Gray | 1 | 1 | 0 | 0 |  |
|  | | | | | 0.170951 |
| stalk-color-below-ring | White | 3 | 3 | 0 | 0 |  |
| Pink | 2 | 0 | 2 | 0 |  |
|  | | | | | 0.970951 |
| veil-color | White | 5 | 3 | 2 | 0.970951 |  |
|  | | | | | 0 |
| ring-number | One | 5 | 3 | 2 | 0.970951 |  |
|  | | | | | 0 |
| Ring-Type | Evanescent | 2 | 1 | 1 | 1 |  |
| Pendant | 2 | 2 | 0 | 0 |  |
| Large | 1 | 0 | 1 | 0 |  |
|  | | | | | 0.170951 |
| spore-print-color | White | 1 | 0 | 1 | 0 |  |
| Brown | 3 | 3 | 0 | 0 |  |
| Chocolate | 1 | 0 | 1 | 0 |  |
|  | | | | | 0.970951 |
| population | Several | 3 | 2 | 1 | 0.918296 |  |
| Solitary | 1 | 0 | 1 | 0 |  |
| Abundant | 1 | 1 | 0 | 0 |  |
|  | | | | | 0.419973 |
| habitat | Woods | 3 | 2 | 1 | 0.918296 |  |
| Grasses | 2 | 1 | 1 | 1 |  |
|  | | | | | 0.0199734 |

Kesimpulan:

Dapat dilihat data diatas nilai gain paling tinggi yaitu 0.970951 yaitu pada atribut spore-print-color, jadi atribut tersebut di taruh pada root paling teratas, sama seperti pada grafik

**PENUTUPAN**

Data yang dianalisa dengan menerapkan algortima C4.5 dalam pembentukan Decision Tree digunakan untuk menampilkan informasi mengenai klasifikasi jamur dengan teknik data mining. Berdasarkan proses data mining yang telah dilakukan, kami mengambil kesimpulan terhadap penerapan data mining dalam dataset mushroom yang kami analisa menggunakan metode Decision Tree sebagai berikut:

1. Pengolahan data mushroom yang kami dapatkan menggunakan metode Decision Tree menghasilkan sebuah kesimpulan data berupa faktor perbandingan yang memperngaruhi apakah jamur tersebut dapat dimakan (EDIBLE) atau beracun (POISONOUS).
2. Hasil dari data mining menggunakan Decision Tree dibuat dengan susunan kegiatan dalam bentuk klasifikasi yang saling berurutan sehingga mudah dipahami dengan melihat gambar decision tree.
3. Metode decision tree yang digunakan memberikan akurasi 100% dalam analisa klasifikasi dataset mushroom yang telah dibuat.
4. Perhitungan dengan algortima C4.5 memberikan nilai gain tertinggi yaitu 0.970951 sehingga dapat ditempatkan sebagai root paling atas

Dari fitur yang didapatkan pada dataset yaitu cap-shape, cap-surface, cap-color,  bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, vail-color, ring-number, ring-type, spore-print-color, population, habitat didapatkan rule yang tepat dalam menentukan apakah jamur tersebut layak dimakan atau beracun. Analisa perhitungan menghasilkan nilai yang cocok digunakan untuk penentuan klasifikasi data jamur sesuai dengan algoritma C4.5