

MODUL

STMIK WIDYA PRATAMA



DATA MINING

Decision Tree



DESKRIPSI MATAKULIAH

CAPAIAN PEMBELAJARAN

DOSEN PENGAMPU

SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN KOMPUTER

(STMIK) WIDYA PRATAMA

CAPAIAN PEMBELAJARAN

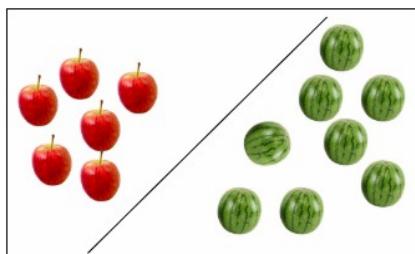
MATERI PEMBELAJARAN

Decision Tree

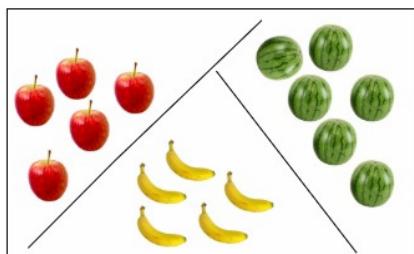
CLASSIFICATION

BINARY CLASSIFICATION

Seperti yang telah Anda ketahui, klasifikasi adalah teknik untuk menentukan kelas atau kategori berdasarkan atribut yang diberikan. Klasifikasi yang menghasilkan dua kategori disebut klasifikasi biner, sedangkan klasifikasi yang menghasilkan 3 kategori atau lebih disebut multiclass classification atau klasifikasi banyak kelas.



Klasifikasi biner



Klasifikasi banyak kelas

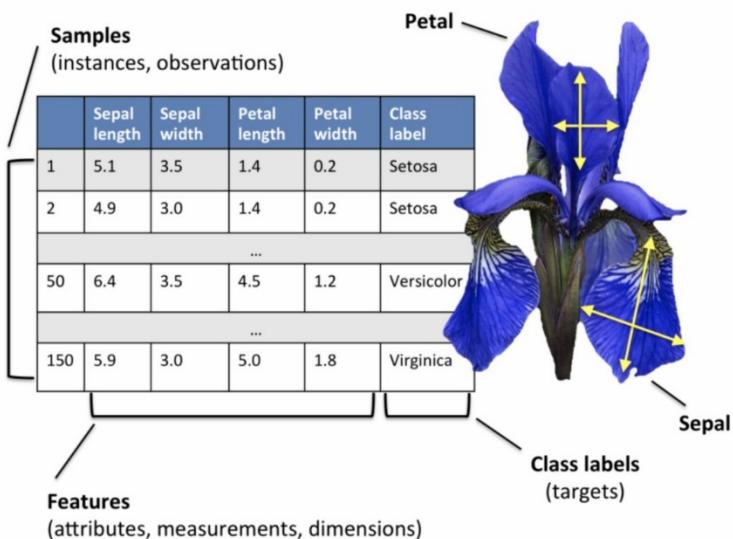
Bayangkan sebuah pertanyaan yang jawabannya adalah iya atau tidak, atau pertanyaan yang membuat Anda memilih antara satu pilihan atau pilihan yang lain. Klasifikasi biner bertujuan untuk membedakan dua kelas yang berbeda. Contohnya, klasifikasi buah semangka atau apel, klasifikasi laki-laki atau perempuan, klasifikasi email spam, dan lain-lain. Pada kasus klasifikasi email sebagai spam atau bukan, pertanyaan yang diajukan adalah: "Apakah email ini adalah spam?"

Sampai sini sudah paham ya. Selanjutnya, pada modul selanjutnya, kita akan belajar tentang klasifikasi banyak kelas.

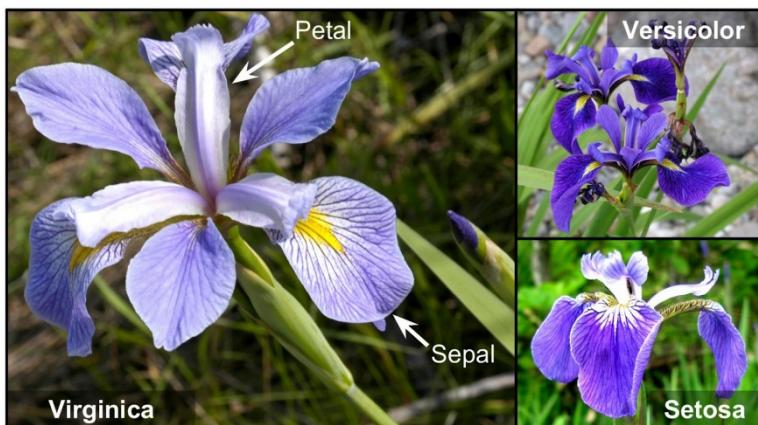
Multiclass Classification

Untuk lebih mudah dalam memahami klasifikasi banyak kelas, kita akan menggunakan contoh dataset Iris. Dataset iris merupakan salah satu dataset populer untuk belajar bagaimana ML dipakai dalam klasifikasi. Dataset ini berisi 150 sampel dari 3 spesies bunga iris.

Pada dataset Iris terdapat 4 kolom atribut yaitu panjang sepal, lebar sepal, panjang petal, dan lebar petal. Untuk label terdapat 3 kelas yaitu Setosa, Versicolor dan Virginica. Kelas adalah kategori atau jenis yang terdapat pada dataset. Dalam hal ini pada dataset terdapat 3 kelas yaitu Setosa, Versicolor, dan Virginica.



Sebuah model classification bertujuan untuk menentukan kelas berdasarkan atribut tertentu. Pada kasus klasifikasi Iris sebuah model bertugas untuk memprediksi spesies sebuah bunga iris berdasarkan atributnya yaitu panjang sepal, lebar sepal, panjang petal, dan lebar petalnya.



Contohnya panjang petal dari Iris Setosa lebih pendek dari spesies versicolor dan virginica. Jika panjang petal pendek maka kemungkinan spesies Iris tersebut adalah Setosa.

Seperti yang sudah dijelaskan pada modul sebelumnya, klasifikasi biner terdapat hanya 2 kelas pada dataset. Sedangkan pada klasifikasi banyak kelas terdapat lebih dari 2 kelas. Pada contoh dataset Iris seperti di atas terdapat 3 kelas maka dataset tersebut adalah kasus klasifikasi banyak kelas.

Decision Tree

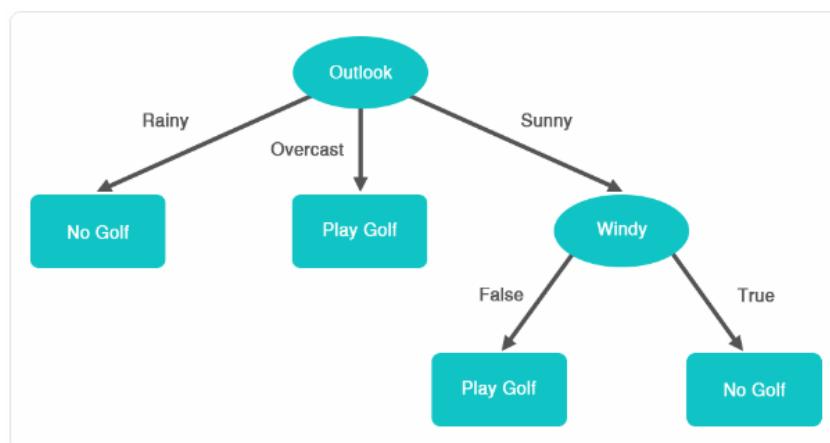
Decision tree atau pohon keputusan adalah salah satu algoritma supervised learning yang dapat dipakai untuk masalah klasifikasi dan regresi. Decision tree merupakan algoritma yang powerful alias mampu dipakai dalam masalah yang kompleks. Decision tree juga merupakan komponen pembangun utama algoritma Random Forest, yang merupakan salah satu algoritma paling powerful saat ini.

Decision tree memprediksi sebuah kelas (klasifikasi) atau nilai (regresi) berdasarkan aturan-aturan yang dibentuk setelah mempelajari data.

Misalnya kita memiliki data seperti di bawah. Data berisi informasi mengenai kondisi cuaca pada hari tertentu dan apakah cocok untuk bermain golf di kondisi cuaca tersebut.

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cold	Normal	False	Yes
Sunny	Cold	Normal	True	No
Overcast	Cold	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cold	Normal	False	Yes
Rainy	Mild	Normal	False	Yes

Sebuah pohon keputusan dapat dibuat dari data sebelumnya. Perhatikan contoh pohon keputusan di bawah. Pohon ini menggunakan hanya 2 atribut yaitu kondisi langit dan kecepatan angin untuk menentukan bermain golf atau tidak.



Latihan Praktikum SKLearn Decision Tree

Tujuan

Pada latihan ini, kita akan melakukan klasifikasi data yang kita miliki dengan teknik Decision Tree menggunakan **dataset iris**, salah satu dataset paling populer yang sering digunakan untuk belajar machine learning.

Tahapan Latihan

Dataset iris terdiri dari 4 atribut yaitu panjang sepal, lebar sepal, panjang petal, dan lebar petal. Terdapat 3 kelas target pada dataset ini. Data ini digunakan untuk masalah klasifikasi, di mana kita memprediksi jenis spesies sebuah bunga berdasarkan atribut-atribut yang diberikan.

Tahapan yang ada pada latihan ini antara lain:

1. Ubah dataset ke dalam dataframe.
2. Hapus kolom 'Id' pada dataframe serta pisahkan antara atribut dan label.
3. Bagi dataset menjadi data latih dan data uji.
4. Buat dan latih model Decision Tree.
5. Lakukan pengujian model dengan menggunakan data uji.
6. Lakukan prediksi dengan model yang telah dilatih.
7. Visualisasi model Decision Tree yang telah dilatih.

Codelab

Pertama kita akan mengimpor library yang dibutuhkan dan mempersiapkan dataset. Dataset dapat anda unduh di [tautan](#) berikut. Setelah data diunduh, masukkan berkas **Iris.csv** ke dalam Google Colab. Lalu jangan lupa konversi dataset menjadi Pandas dataframe.

```
1. import pandas as pd  
2.  
3. # Membaca file iris.csv  
4. iris = pd.read_csv('Iris.csv')
```

Untuk melihat informasi mengenai data, gunakan fungsi `info()`. Selain itu, Anda juga bisa melihat lima data teratas pada dataset menggunakan fungsi `head()`.

```
1. # Melihat informasi dataset  
2. iris.info()  
3.  
4. # melihat informasi dataset pada 5 baris pertama  
5. iris.head()
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Id          150 non-null    int64  
 1   SepalLengthCm 150 non-null   float64 
 2   SepalWidthCm  150 non-null   float64 
 3   PetalLengthCm 150 non-null   float64 
 4   PetalWidthCm  150 non-null   float64 
 5   Species      150 non-null   object  
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB
```

Dari output di atas, kita dapat mengidentifikasi kolom yang tidak penting pada dataset yaitu kolom "Id". Untuk menghilangkan kolom tersebut, gunakan fungsi `drop()`.

1. `# menghilangkan kolom yang tidak penting`
2. `iris.drop('Id',axis=1,inplace=True)`

Sebelum melatih model kita perlu memisahkan atribut dengan label. Selain itu, kita juga perlu membagi dataset menjadi data latih dan data uji. Jalankan kode berikut untuk menerapkan tahapan di atas.

```
1. # memisahkan atribut dan label
2. X = iris[['SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm' ]]
3. y = iris['Species']
4.
5. # Membagi dataset menjadi data latih & data uji
6. from sklearn.model_selection import train_test_split
7. X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=123)
```

Selanjutnya, definisikan model decision tree yang akan kita gunakan. Kemudian, latih model menggunakan data latih menggunakan fungsi `fit()`.

```
1. from sklearn.tree import DecisionTreeClassifier  
2.  
3. # membuat model Decision Tree  
4. tree_model = DecisionTreeClassifier()  
5.  
6. # Melatih model dengan menggunakan data latih  
7. tree_model = tree_model.fit(X_train, y_train)
```

Setelah model dilatih, uji model menggunakan data uji untuk melihat seberapa baik model yang telah kita buat. Pengujian model ini bisa dilakukan dengan menggunakan fungsi `predict()`.

Berikutnya, gunakan metrik akurasi untuk melihat seberapa baik model yang telah kita latih. Penjelasan terkait metrik akurasi ini akan dibahas pada modul selanjutnya.

```
1. # Evaluasi Model  
2. from sklearn.metrics import accuracy_score  
3.  
4. y_pred = tree_model.predict(X_test)  
5.  
6. acc_secore = round(accuracy_score(y_pred, y_test), 3)  
7.  
8. print('Accuracy: ', acc_secore)
```

Accuracy: 0.933

Nah, kita bisa mencoba model yang telah kita buat untuk memprediksi spesies dari sebuah bunga Iris. Masih ingat bukan, atribut yang menjadi masukan dari model adalah panjang sepal, lebar sepal, panjang petal, dan lebar petal? Kita akan memasukkan nilai yang sesuai dengan format tersebut secara berurutan dalam satuan centimeter.

Pada contoh berikut, kita ingin memprediksi spesies dari sebuah bunga iris yang memiliki panjang sepal 6,2 centimeter, lebar sepal 3,4 centimeter, panjang petal 5,4 centimeter, dan lebar petal 2,3 centimeter.

```
1. # prediksi model dengan tree_model.predict([[SepalLength, SepalWidth,  
PetalLength, PetalWidth]])  
  
2. print(tree_model.predict([[6.2, 3.4, 5.4, 2.3]])[0])
```

→ Iris-virginica

Selain melakukan prediksi, kita juga bisa melihat visualisasi dari decision tree yang kita buat terhadap data menggunakan library **Graphviz**. Hasil dari graphviz adalah **dot file** yang akan muncul pada folder file di panel sebelah kiri Google Colab (jika Anda menggunakan Google Colab).

```
1. from sklearn.tree import export_graphviz  
2. export_graphviz(  
3.     tree_model,  
4.     out_file = "iris_tree.dot",  
5.     feature_names = ['SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm'],  
6.     class_names = ['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'],  
7.     rounded= True,  
8.     filled =True)
```

Setelah kode di atas berhasil dijalankan, Anda akan mendapatkan output berupa berkas **iris_tree.dot**, sebagai berikut:

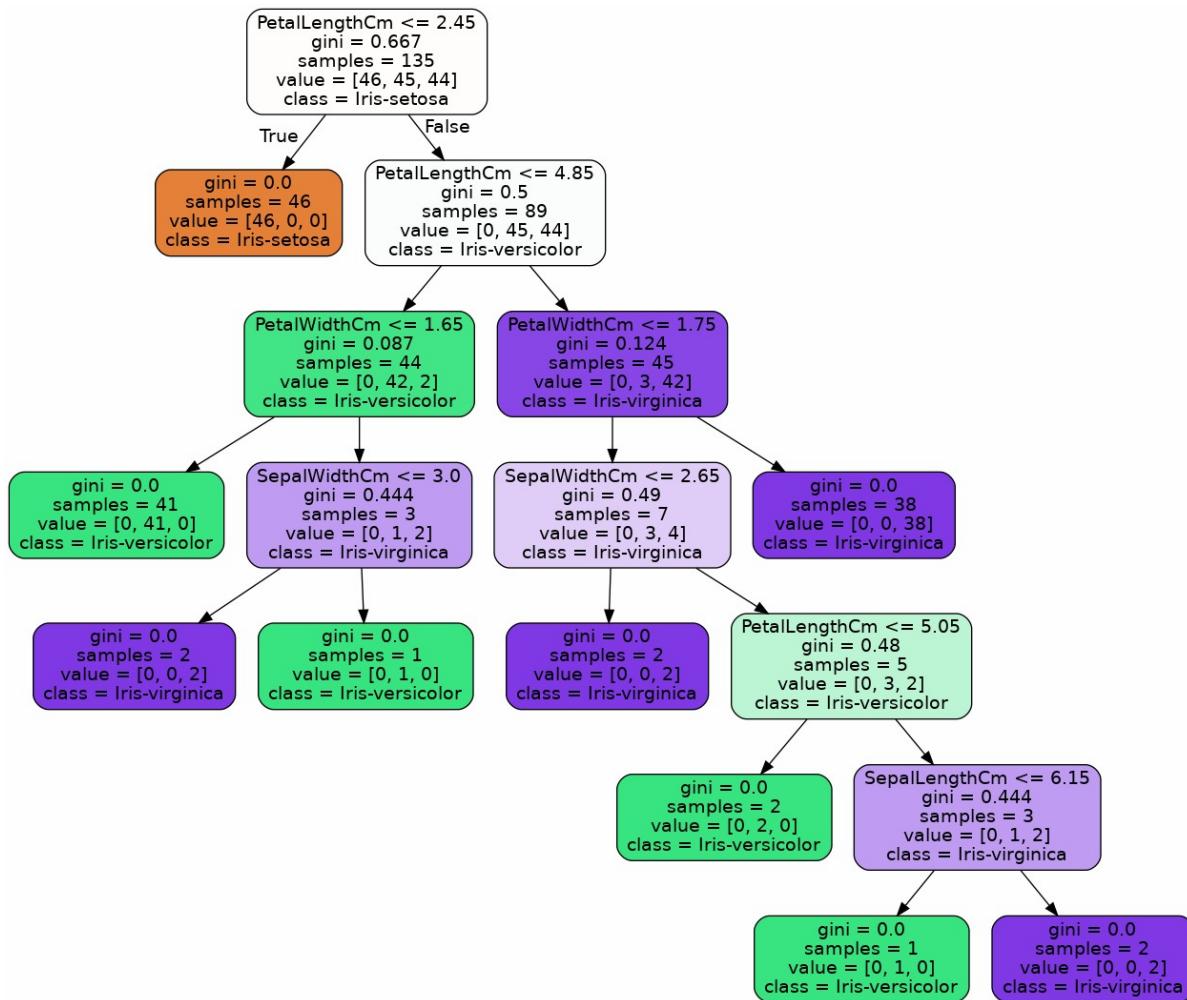


Untuk mengunduh berkas **iris_tree.dot** pada gambar di atas, kita dapat melakukan klik kanan pada berkas tersebut kemudian mengunduhnya.

Jika kita ingin melihat visualisasi decision tree, lakukan konversi dot file ke dalam file **png** menggunakan situs konversi berkas berikut ini <https://onlineconvertfree.com/converter/images/>.

Catatan : Jangan lupa ganti opsi ke images sebelum menekan tombol convert

Berikut merupakan hasil visualisasi dari model decision tree yang telah kita gunakan:



Selamat! Anda telah berhasil membuat sebuah model decision tree untuk klasifikasi spesies bunga Iris. Anda juga telah berhasil menguji model anda untuk memprediksi spesies dari sebuah bunga iris. Untuk belajar lebih mendalam tentang decision tree, kunjungi [tautan](https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052) (<https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>)