



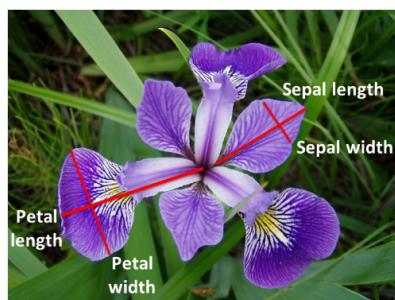
## Hands-On

Hands-On ini digunakan pada kegiatan Microcredential Associate Data Scientist 2021

## Tugas Mandiri Pertemuan 11

Pertemuan 11 (sebelas) pada Microcredential Associate Data Scientist 2021 menyampaikan materi mengenai Membangun Model 2 (Regresi Non Linier, Support Vector Machine, dll). silakan Anda kerjakan Latihan 1 s/d 20. Output yang anda lihat merupakan panduan yang dapat Anda ikuti dalam penulisan code :)

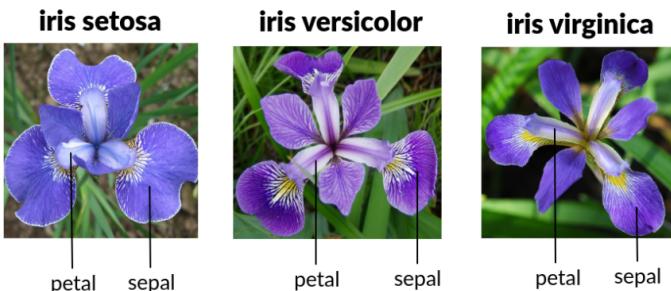
### About Iris dataset



The iris dataset contains the following data (**Before Cleansing**)

- 50 samples of 3 different species of iris (150 samples total)
- Measurements: sepal length, sepal width, petal length, petal width
- The format for the data: (sepal length, sepal width, petal length, petal width)

The variables are:



- `sepal_length`: Sepal length, in centimeters, used as input.
- `sepal_width`: Sepal width, in centimeters, used as input.
- `petal_length`: Petal length, in centimeters, used as input.
- `petal_width`: Petal width, in centimeters, used as input.
- `class`: Iris Setosa, Versicolor, or Virginica, used as the target.

## Contents

### Data Preprocessing

- Include Libraries
- Import DataSet
- Handle Missing Value (sudah dilakukan pada pert 8)

### Data Visualization

- Scatterplot
- Pairplot
- Barplot
- Violin
- Areaplot
- Correlation

### Feature Engineering

#### Machine learning Model (Regresi Non Linier, Support Vector Machine, dll)

- Logistic Regression
- Naive Bayes
- KNN
- Support Vector Machine

## 1. Data Preprocessing

## Latihan (1)

Melakukan import library yang dibutuhkan

```
In [1]: 1 # import Library pandas
2 import pandas as pd
3
4 # Import Library numpy
5 import numpy as np
6
7 # Import Library matplotlib dan seaborn untuk visualisasi
8 import matplotlib.pyplot as plt
9 import seaborn as sns
10
11 # me-non aktifkan peringatan pada python
12 import warnings
13 warnings.filterwarnings('ignore')
```

Load Dataset

```
In [2]: 1 #Panggil file (Load file bernama CarPrice.Assignment.csv) dan simpan dalam dataframe
2 dataset = pd.read_csv('Iris.AfterClean.csv')
3 iris = pd.DataFrame(dataset)
```

## Latihan (2)

Review Dataset

```
In [3]: 1 # tampilkan 5 baris awal dataset dengan function head()
2 iris.head(5)
```

```
Out[3]:
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	5.0	3.5	1.4	0.2	Iris-setosa
1	4.4	3.0	1.3	0.2	Iris-setosa
2	4.9	3.4	1.5	0.2	Iris-setosa
3	4.9	3.4	1.4	0.2	Iris-setosa
4	4.6	3.6	1.0	0.2	Iris-setosa

```
In [4]: 1 # tampilkan unique value dari species
2 iris['Species'].unique()
```

```
Out[4]: array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

dari output diatas, dataset ini memiliki tiga varietas tanaman Iris.

```
In [5]: 1 # melihat statistik data untuk data numeric dan non numeric
2 iris.describe(include='all')
```

```
Out[5]:
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
count	140.000000	140.000000	140.000000	140.000000	140
unique	NaN	NaN	NaN	NaN	3
top	NaN	NaN	NaN	NaN	Iris-virginica
freq	NaN	NaN	NaN	NaN	50
mean	5.002857	3.752857	1.587300	0.375000	NaN
std	0.800000	0.800000	0.800000	0.800000	NaN
min	4.300000	2.000000	1.000000	0.100000	NaN
25%	4.900000	3.000000	1.300000	0.300000	NaN
50%	5.000000	3.400000	1.500000	0.500000	NaN
75%	5.100000	3.700000	1.700000	0.800000	NaN
max	7.900000	4.400000	2.500000	1.900000	NaN

```
In [6]: 1 # Melihat Informasi lebih detail mengenai struktur DataFrame dapat dilihat menggunakan fungsi info()
2 iris.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 140 entries, 0 to 139
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype  
 ---  -- 
 0   SepalLengthCm    140 non-null    float64
 1   SepalWidthCm     140 non-null    float64
 2   PetalLengthCm    140 non-null    float64
 3   PetalWidthCm     140 non-null    float64
 4   Species          140 non-null    object  
dtypes: float64(4), object(1)
memory usage: 5.6+ K
```

Seperi yang kita lihat di atas distribusi titik data di setiap kelas adalah sama sehingga Iris adalah dataset seimbang

## Latihan (3)

Checking if there are any missing values



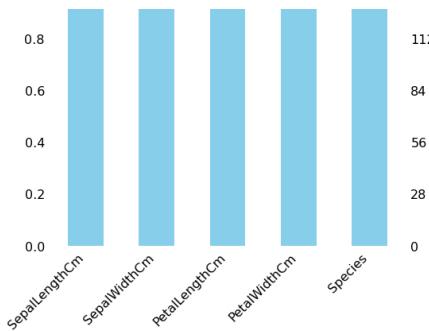
```
In [7]: 1 # cek jumlah nilai yang hilang / missing values dari setiap kolom dengan function isnull() dan sum()
2 iris.isnull().sum()
```

```
Out[7]:
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
count	140	140	140	140	140
unique	NaN	NaN	NaN	NaN	3
top	NaN	NaN	NaN	NaN	Iris-virginica
freq	NaN	NaN	NaN	NaN	50

```
In [8]: 1 # cek missing values dengan visualisasi menggunakan Library: Missingno adalah pustaka khusus untuk menampilkan nilai yang hilang
2 # jenis: barchart
3 import missingno as msno
4 msno.bar(iris,figsize=(8,6),color='skyblue')
5 plt.show()
```





Dataset IrisAfterclean.csv ini adalah dataset yang telah melewati proses cleansing pada pertemuan 8 kemarin sehingga dataset ini sudah bersih

## 2. Data Visualization

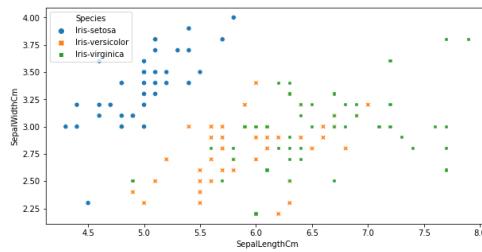
### 2.1 Scatter Plot

Scatter plot adalah visualisasi data dua dimensi yang menggunakan titik untuk mewakili nilai yang diperoleh untuk dua variabel berbeda, satu diplot sepanjang sumbu x dan yang lainnya diplot sepanjang sumbu y. Kita dapat memplot scatter plot di antara dua fitur.

#### Latihan (4)

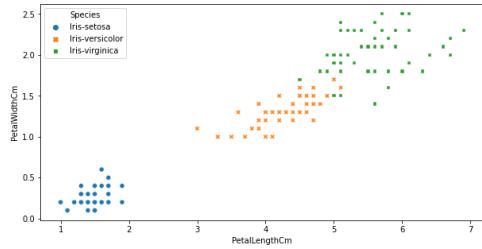
buat visualisasi scatter plot 'Sepal Length' dan 'Sepal Width'

```
In [9]: 1 # visualisasi scatter plot dengan seaborn antara Sepal Length dan Sepal Width dengan parameter hue dan style='species'
2 g=sns.scatterplot(x="SepallengthCm",y="SepalWidthCm",data=iris,hue="Species",style="Species")
3 g.figure.set_size_inches(10,5)
4 plt.show()
```



buat visualisasi scatter plot 'Petal Length' dan 'Petal Width'

```
In [10]: 1 # visualisasi scatter plot dengan seaborn antara Petal Length dan Petal Width dengan parameter hue dan style='species'
2 g=sns.scatterplot(x="PetallengthCm",y="PetawidthCm",data=iris,hue="Species",style="Species")
3 g.figure.set_size_inches(10,5)
4 plt.show()
```



Seperti yang dapat kita lihat bahwa Fitur Petal memberikan pembagian cluster yang lebih baik dibandingkan dengan fitur Sepal. Ini merupakan indikasi bahwa Petal dapat membantu dalam Prediksi yang lebih baik dan akurat dari pada Sepal.

### 2.2 Pairplot

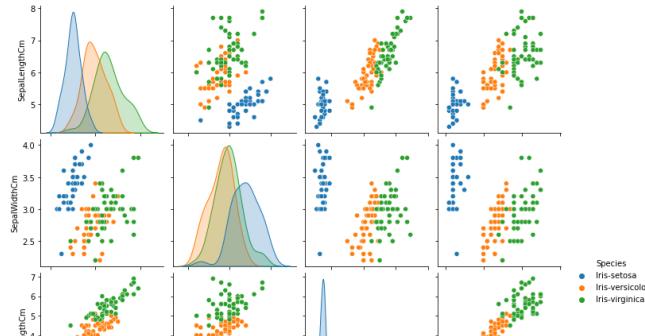
Pair Plots adalah cara yang sangat sederhana (satu baris kode sederhananya) untuk memvisualisasikan hubungan antara setiap variabel. Ini menghasilkan matriks hubungan antara setiap variabel dalam data kita untuk pemeriksaan data instan. Pair Plots memberikan scatter plot dari fitur yang berbeda. Pair Plots untuk kumpulan data iris.

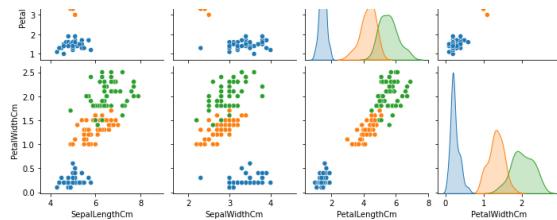
#### Latihan (5)

buat visualisasi Pair Plots dari data iris

```
In [11]: 1 # buat visualisasi Pair Plots dari data iris dengan parameter hue='species'
2 sns.pairplot(iris,hue="Species")
3 plt.show()
```

Out[11]: <function matplotlib.pyplot.show(close=None, block=None)>

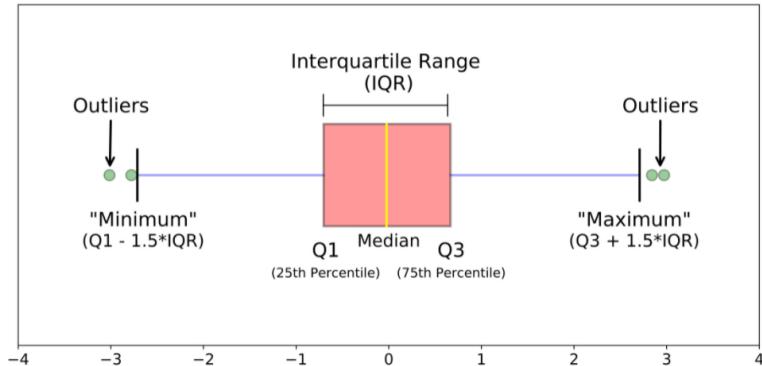




dari grafik kita dapat melihat scatter plot antara dua fitur dan distribusinya, dari sebaran di atas petal length memisahkan iris setosa dari yang tersisa, dari plot antara petal length dan petal width kita dapat memisahkan bunga

### 2.3 BoxPlot

boxplot adalah cara standar untuk menampilkan distribusi data berdasarkan ringkasan lima angka ("minimum", kuartil pertama (Q1), median, kuartil ketiga (Q3), dan "maksimum"). Ini dapat memberi tahu kita tentang outlier dan apa lainnya. Ini juga dapat memberi tahu kita apakah data kita simetris, seberapa ketat data kita dikelompokkan, dan bagaimana jika data kita miring.

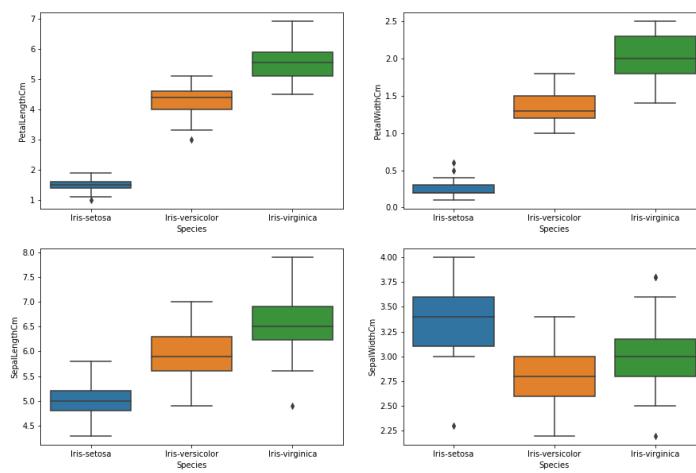


### Latihan (6)

buat visualisasi box plot dari setiap kolom feature terhadap species

```
In [12]: 1 plt.figure(figsize=(15,10))
2 plt.subplots(2,2,1)
3 sns.boxplot(x='Species',y='PetalLengthCm',data=iris)
4 plt.subplot(2,2,2)
5 sns.boxplot(x='Species',y='PetalWidthCm',data=iris)
6 plt.subplot(2,2,3)
7 sns.boxplot(x='Species',y='SepalLengthCm',data=iris)
8 plt.subplot(2,2,4)
9 sns.boxplot(x='Species',y='SepalWidthCm',data=iris)
10 plt.show()
```

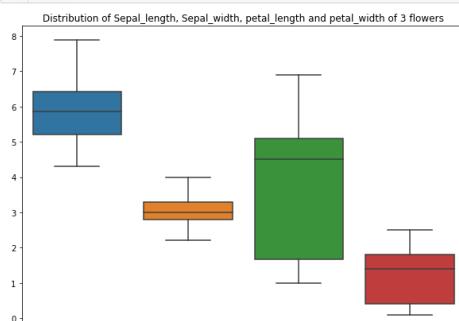
```
Out[12]: <function matplotlib.pyplot.show(close=None, block=None)>
```



### Latihan (7)

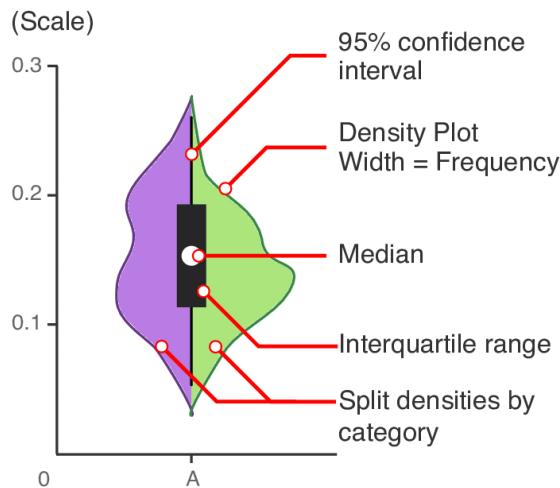
buat visualisasi box plot distribusi setiap kolom feature

```
In [13]: 1 plt.subplots(figsize=(10,7))
2 sns.boxplot(data=iris).set_title("Distribution of Sepal_length, Sepal_width, petal_length and petal_width of 3 flowers")
3 plt.show()
```



## 2.4 Violin

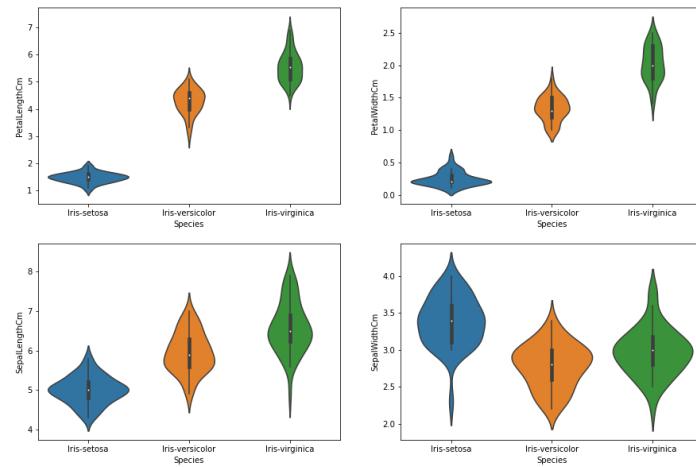
Violin Plot adalah metode untuk memvisualisasikan distribusi data numerik dari variabel yang berbeda. Ini mirip dengan Box Plot tetapi dengan plot yang diputar di setiap sisi, memberikan lebih banyak informasi tentang perkiraan kepadatan pada sumbu y. Kepadatan dicerminkan dan dibalik dan bentuk yang dihasilkan diciptakan gambar yang menyerupai biola. Kelebihan dari Violin Plot adalah dapat menampilkan nuansa dalam distribusi yang tidak terlihat dalam boxplot. Di sisi lain, boxplot lebih jelas menunjukkan outlier dalam data.



## Latihan (8)

buat visualisasi violin plot setiap kolom feature

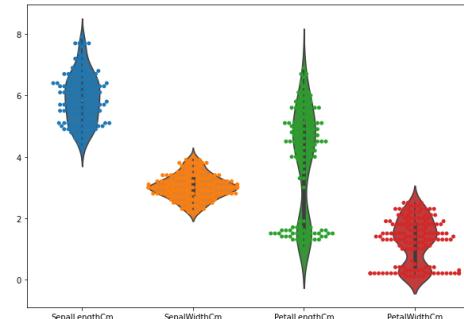
```
In [14]: 1 plt.figure(figsize=(15,10))
2 plt.subplots(2,2,1)
3 sns.violinplot(x='Species',y='PetalLengthCm',data=iris)
4 plt.subplots(2,2,2)
5 sns.violinplot(x='Species',y='PetalWidthCm',data=iris)
6 plt.subplots(2,2,3)
7 sns.violinplot(x='Species',y='SepallengthCm',data=iris)
8 plt.subplots(2,2,4)
9 sns.violinplot(x='Species',y='SepalWidthCm',data=iris)
10 plt.show()
```



## Latihan (9)

buat visualisasi violin plot dengan swarm plot

```
In [15]: 1 plt.subplots(figsize=(10,7))
2 sns.violinplot(data=iris)
3 sns.swarmplot(data=iris)
4 plt.show()
```



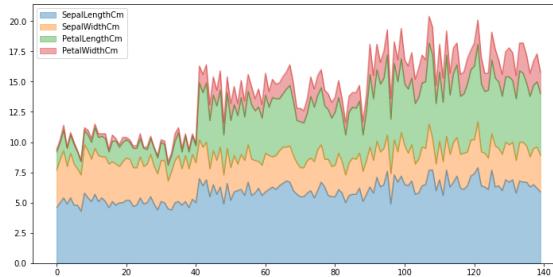
## 2.5 Area Plot

Area Plot memberi kita representasi visual dari Berbagai dimensi bunga Iris dan jangkaunya dalam dataset.

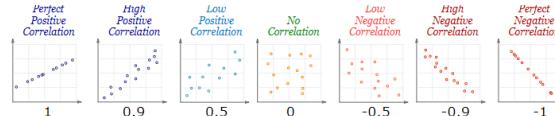
## Latihan (10)

buat visualisasi area plot pada setiap feature kolom

```
In [16]: 1 iris.plot.area(y=["SepalLengthCm","SepalWidthCm","PetalLengthCm","PetalWidthCm"],alpha=0.4,figsize=(12, 6));
```



## 2.6 Correlation



Sekarang, ketika kami melatih algoritma apa pun, jumlah fitur dan korelasinya memainkan peran penting. Jika ada fitur dan banyak fitur yang sangat berkorelasi, maka melatih suatu algoritma dengan semua fitur akan mengurangi akurasi. Dengan demikian pemilihan fitur harus dilakukan dengan hati-hati. Dataset ini memiliki fitur yang lebih sedikit tetapi kita masih akan melihat korelasinya.

## Latihan (11)

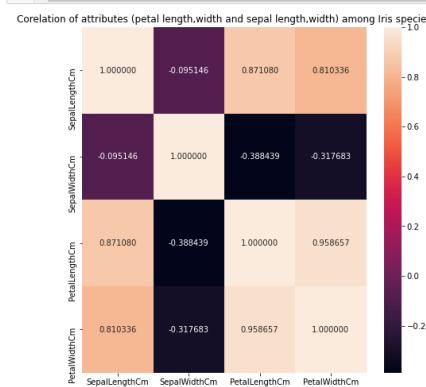
lihat korelasi dataset dan visualisasi dengan heatmap pada feature kolom

```
In [17]: 1 # lihat korelasi dengan function corr()  
2 iris.corr()
```

Out[17]:

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
SepalLengthCm	1.000000	-0.095146	0.871080	0.810336
SepalWidthCm	-0.095146	1.000000	-0.388439	-0.317683
PetalLengthCm	0.871080	-0.388439	1.000000	0.958657
PetalWidthCm	0.810336	-0.317683	0.958657	1.000000

```
In [18]: 1 # lihat korelasi dengan visualisasi heatmap  
2 plt.figure(figsize = (8,8))  
3 sns.heatmap(iris.corr(),annot=True,fmt=".2f").set_title("Corelation of attributes (petal length,width and sepal length,width)  
4 plt.show()
```



### Observasi :

Sepal Width dan Sepal Length tidak berkorelasi || Petal Width and Petal Length sangat berkorelasi

Kami akan menggunakan semua fitur untuk melatih algoritme dan memeriksa keakuratannya.

## Dividing data into features and labels

### Feature Selection

Full Feature Set



Identify Useful Features



Selected Feature Set



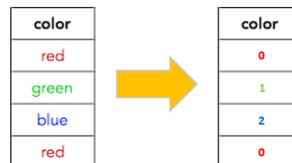
Seperti yang kita lihat, dataset berisi lima kolom: SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm dan Species. Fitur yang sebenarnya dijelaskan oleh kolom 1-4. Kolom terakhir berisi label sampel. Pertama kita perlu membagi data menjadi dua array: X (fitur) dan y (label).

## Latihan (12)

definisi variabel X(feature kolom) dan y(species/label):

```
In [19]: 1 X=iris.iloc[:, 0:4].values  
2 y=iris.iloc[:, 4].values
```

### Label encoding



Seperti yang kita lihat, label bersifat kategoris. KNeighborsClassifier tidak menerima label string. Kita perlu menggunakan LabelEncoder untuk mengubahnya menjadi angka. Iris-setosa sesuai dengan 0, Iris-versicolor sesuai dengan 1 dan Iris-virginica sesuai dengan 2.

### Latihan (13)

transform label data species dengan menggunakan library LabelEncoder

```
In [20]: 1 from sklearn import preprocessing
2 le = preprocessing.LabelEncoder()
3 y = le.fit_transform(y)
```

## 3. Building Machine Learning Models

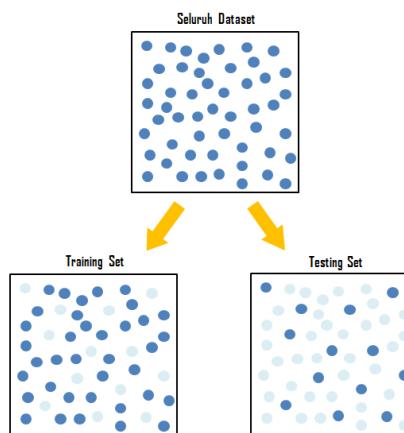
### Latihan (14)

import library dalam kebutuhan membangun model

```
In [21]: 1 #Metrics
2 from sklearn.metrics import make_scorer, accuracy_score, precision_score
3 from sklearn.metrics import classification_report
4
5
6 # Import library confusion matrix
7 from sklearn.metrics import confusion_matrix
8
9 from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
10
11 #Model Select
12 from sklearn.model_selection import KFold, train_test_split, cross_val_score
13 from sklearn.model_selection import train_test_split
14
15 # Import Library Logistic Regression
16 from sklearn.linear_model import LogisticRegression
17
18 from sklearn import linear_model
19 from sklearn.linear_model import SGDClassifier
20
21 # Import Library KNN
22 from sklearn.neighbors import KNeighborsClassifier
23
24 # Import Library Support Vector Machines dan Linier Support Vector Machines
25 from sklearn.svm import SVC
26 from sklearn.svm import LinearSVC
27
28 # Import Library Gaussian Naive Bayes
29 from sklearn.naive_bayes import GaussianNB
```

### Splitting The Data into Training And Testing Dataset

Train/test split adalah salah satu metode yang dapat digunakan untuk mengevaluasi performa model machine learning. Metode evaluasi model ini membagi dataset menjadi dua bagian yakni bagian yang digunakan untuk training data dan untuk testing data dengan proporsi tertentu. Train data digunakan untuk fit model machine learning, sedangkan test data digunakan untuk mengevaluasi hasil fit model tersebut.



Python memiliki library yang dapat mengimplementasikan train/test split dengan mudah yaitu Scikit-Learn. Untuk menggunakannya, kita perlu mengimport Scikit-Learn terlebih dahulu, kemudian setelah itu kita dapat menggunakan fungsi train\_test\_split().

### Latihan (15)

split data train dan test dengan function train\_test\_split() dengan train\_size=0.7, test\_size=0.3 dan random\_state=0

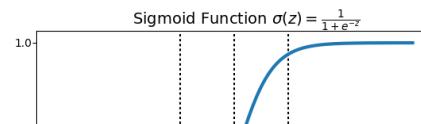
```
In [22]: 1 #Train and Test split
2 X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=0)
```

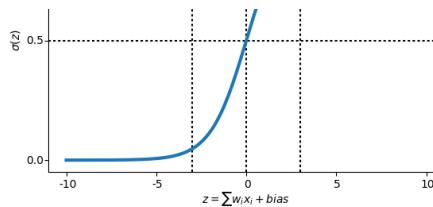
Sekarang kita akan melatih beberapa model Machine Learning dan membandingkan hasilnya. Perhatikan bahwa karena set data tidak memberikan label untuk set pengujiannya, kita perlu menggunakan prediksi pada set pelatihan untuk membandingkan algoritme satu sama lain.

### 3.2 Logistic Regression:

Logistic Regression adalah algoritma Machine Learning yang digunakan untuk masalah klasifikasi, ini adalah algoritma analisis prediktif dan berdasarkan konsep probabilitas.

Kita dapat menyebut Logistic Regression sebagai model Regresi Linier tetapi Regresi Logistik menggunakan fungsi biaya yang lebih kompleks, fungsi biaya ini dapat didefinisikan sebagai 'fungsi Sigmoid' atau juga dikenal sebagai 'fungsi logistik' daripada fungsi linier.





### Bangun model LogisticRegression dan akurasi nya

```
In [23]: 1 logreg = LogisticRegression(solver='lbfgs',max_iter=400)
2 logreg.fit(X_train,y_train)
3 y_pred = logreg.predict(X_test)
4 accuracy_lr=round(accuracy_score(y_test,y_pred)* 100, 2)
5 acc_log = round(logreg.score(X_train, y_train) * 100, 2)
6
7
8 cm = confusion_matrix(y_test, y_pred)
9 accuracy = accuracy_score(y_test,Y_pred)
10 precision = precision_score(y_test, Y_pred,average='micro')
11 recall = recall_score(y_test, Y_pred,average='micro')
12 f1 = f1_score(y_test, Y_pred,average='micro')
13 print('Confusion matrix for Logistic Regression\n',cm)
14 print('accuracy Logistic Regression : %.3f' %accuracy)
15 print('precision Logistic Regression : %.3f' %precision)
16 print('recall Logistic Regression: %.3f' %recall)
17 print('f1-score Logistic Regression : %.3f' %f1)
```

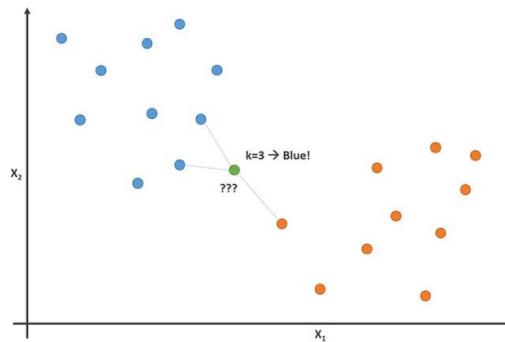
Confusion matrix for Logistic Regression  
[[11 0 ]  
[ 0 14 ]  
[ 0 14]]  
accuracy\_Logistic Regression : 0.952  
precision\_Logistic Regression : 0.952  
recall\_Logistic Regression: 0.952  
f1-score\_Logistic Regression : 0.952

### 3.3 K Nearest Neighbor:

K-Nearest Neighbor adalah salah satu algoritma Machine Learning yang paling sederhana berdasarkan teknik Supervised Learning.

Algoritma K-NN mengasumsikan kesamaan antara kasus/data baru dengan kasus yang tersedia dan memasukkan kasus baru ke dalam kategori yang paling mirip dengan kategori yang tersedia.

Algoritma K-NN menyimpan semua data yang tersedia dan mengklasifikasikan titik data baru berdasarkan kesamaan. Artinya ketika data baru muncul maka dapat dengan mudah diklasifikasikan ke dalam kategori well suited dengan menggunakan algoritma K-NN.



### Latihan (16)

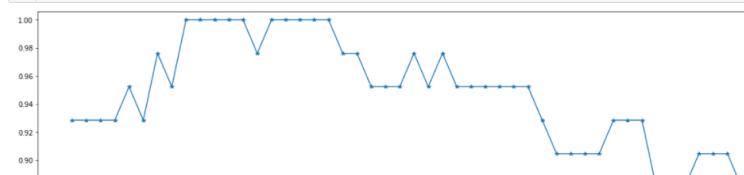
#### Bangun model KNN dan akurasi nya

```
In [24]: 1 knn = KNeighborsClassifier(n_neighbors = 3)
2 knn.fit(X_train,y_train)
3 y_pred = knn.predict(X_test)
4 accuracy_knn=round(accuracy_score(y_test,Y_pred)* 100, 2)
5 acc_knn = round(knn.score(X_train, y_train) * 100, 2)
6
7 cm = confusion_matrix(y_test, y_pred)
8 accuracy = accuracy_score(y_test,Y_pred)
9 precision = precision_score(y_test, Y_pred,average='micro')
10 recall = recall_score(y_test, Y_pred,average='micro')
11 f1 = f1_score(y_test, Y_pred,average='micro')
12 print('Confusion matrix for KNN\n',cm)
13 print('accuracy_KNN : %.3f' %accuracy)
14 print('precision_KNN : %.3f' %precision)
15 print('recall_KNN: %.3f' %recall)
16 print('f1-score_KNN : %.3f' %f1)
```

Confusion matrix for KNN  
[[11 0 ]  
[ 0 14 ]  
[ 0 13]]  
accuracy\_KNN : 0.929  
precision\_KNN : 0.929  
recall\_KNN: 0.929  
f1-score\_KNN : 0.929

#### Mari kita periksa akurasi untuk berbagai nilai n untuk Model KNN

```
In [25]: 1 plt.subplots(figsize=(20,5))
2 a_index=list(range(1,50))
3 a=pd.Series()
4 x=range(1,50)
5 #x=[1,2,3,4,5,6,7,8,9,10]
6 for i in list(range(1,50)):
7     model=KNeighborsClassifier(n_neighbors=i)
8     model.fit(X_train,y_train)
9     prediction=model.predict(X_test)
10    a=a.append(pd.Series(accuracy_score(y_test,prediction)))
11 plt.plot(a_index, a,marker="*")
12 plt.xticks(x)
13 plt.show()
```





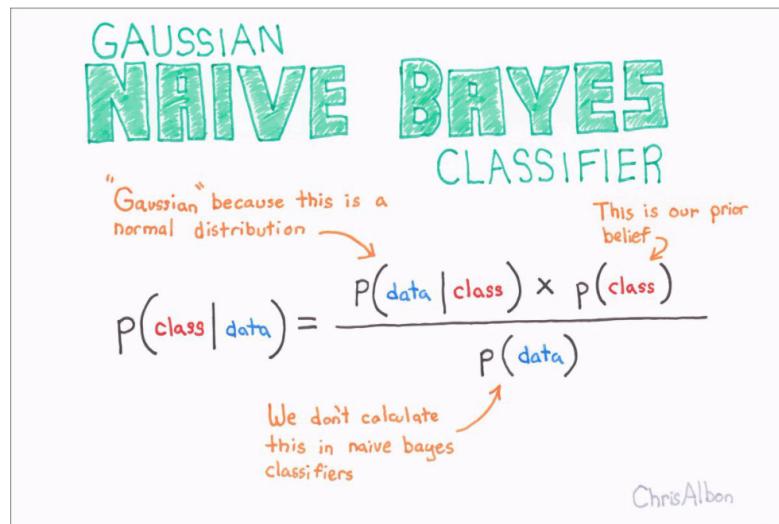
Di atas adalah grafik yang menunjukkan akurasi untuk model KNN menggunakan nilai n yang berbeda.

### 3.4 Gaussian Naive Bayes:

Naive Bayes adalah algoritma klasifikasi untuk masalah klasifikasi biner (dua kelas) dan multi kelas. Teknik ini paling mudah dipahami ketika dijelaskan menggunakan nilai input biner atau kategoris.

Disebut naive bayes atau idiot bayes karena perhitungan probabilitas untuk setiap hipotesis disederhanakan untuk membuat perhitungannya dapat dilakukan. Daripada mencoba menghitung nilai dari setiap nilai atribut  $P(d_1, d_2, d_3|h)$ , mereka dianggap independen bersyarat dengan nilai target dan dihitung sebagai  $P(d_1|h) \cdot P(d_2|h)$  dan seterusnya.

Ini adalah asumsi yang sangat kuat yang paling tidak mungkin dalam data nyata, yaitu bahwa atribut tidak berinteraksi. Namun demikian, pendekatan ini bekerja dengan sangat baik pada data di mana asumsi ini tidak berlaku.



### Latihan (17)

Bangun model gaussian Naive Bayes dan akurasi nya

```
In [26]: 1 gaussian = GaussianNB()
2 gaussian.fit(X_train, y_train)
3 y_pred = gaussian.predict(X_test)
4 accuracy_nb=round(accuracy_score(y_test,Y_pred)* 100, 2)
5 acc_gaussian = round(gaussian.score(X_train, y_train) * 100, 2)

6 cm = confusion_matrix(y_test, Y_pred)
7 accuracy = accuracy_score(y_test,Y_pred)
8 precision = precision_score(y_test, Y_pred,average='micro')
9 recall = recall_score(y_test, Y_pred,average='micro')
10 f1 = f1_score(y_test,Y_pred,average='micro')
11 print('Confusion matrix for Naive Bayes\n',cm)
12 print('accuracy_Naive Bayes: %.3f' %accuracy)
13 print('precision_Naive Bayes: %.3f' %precision)
14 print('recall_Naive Bayes: %.3f' %recall)
15 print('f1-score_Naive Bayes : %.3f' %f1)

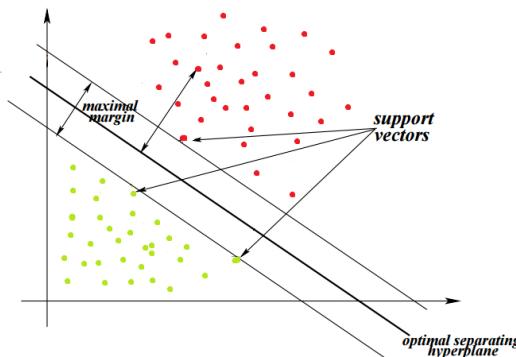
Confusion matrix for Naive Bayes
[[12  0  0]
 [ 0 14  1]
 [ 0  2 13]]
accuracy_Naive Bayes: 0.929
precision_Naive Bayes: 0.929
recall_Naive Bayes: 0.929
f1-score_Naive Bayes : 0.929
```

### Latihan (18)

Bangun model gaussian Naive Bayes dan akurasi nya

### 3.5 Linear Support Vector Machine:

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin terawali yang dapat digunakan untuk klasifikasi atau regresi. Namun, sebagian besar digunakan dalam masalah klasifikasi. Dalam algoritma SVM, kami memplot setiap item data sebagai titik dalam ruang n-dimensi (di mana n adalah jumlah fitur yang Anda miliki) dengan nilai setiap fitur menjadi nilai koordinat tertentu. Kemudian, kami melakukan klasifikasi dengan menemukan hyper-plane yang membedakan kedua kelas dengan sangat baik



### Latihan (19)

Bangun model Linear Support Vector Machines dan akurasi nya

```
In [27]: 1 linear_svc = LinearSVC(max_iter=4000)
2 linear_svc.fit(X_train, y_train)
3 y_pred = linear_svc.predict(X_test)
4 accuracy_svc= round(accuracy_score(y_test,Y_pred)* 100, 2)
```

```

5 acc_linear_svc = round(linear_svc.score(X_train, y_train) * 100, 2)
6
7 cm = confusion_matrix(y_test, Y_pred)
8 accuracy = accuracy_score(y_test,Y_pred)
9 precision = precision_score(y_test, Y_pred,average='micro')
10 recall = recall_score(y_test, Y_pred,average='micro')
11 f1 = f1_score(y_test,Y_pred,average='micro')
12 print('Confusion matrix for SVC\n',cm)
13 print('accuracy_SVC: %3f' %accuracy)
14 print('precision_SVC: %3f' %precision)
15 print('recall_SVC: %3f' %recall)
16 print('f1-score_SVC : % .3f' %f1)

```

Confusion matrix for SVC  
[[2 0 0]  
[0 13 2]  
[0 2 12]]  
accuracy\_SVC: 0.905  
precision\_SVC: 0.905  
recall\_SVC: 0.905  
f1-score\_SVC : 0.905

## Latihan (20)

Model mana yang terbaik ?

```

In [28]: 1 results = pd.DataFrame({
2     'Model': [ 'KNN',
3                'Logistic Regression',
4                'Naive Bayes',
5                'Support Vector Machine'],
6     'Score': [acc_knn,
7               acc_log,
8               acc_gaussian,
9               acc_linear_svc],
10    "Accuracy_score": [accuracy_knn,
11                      accuracy_lr,
12                      accuracy_nb,
13                      accuracy_svc
14                     ])
15 result_df = results.sort_values(by="Accuracy_score", ascending=False)
16 result_df = result_df.reset_index(drop=True)
17 result_df.head(9)

```

	Model	Score	Accuracy_score
0	Logistic Regression	96.94	95.24
1	KNN	96.94	92.86
2	Naive Bayes	96.94	92.86
3	Support Vector Machine	95.92	90.48

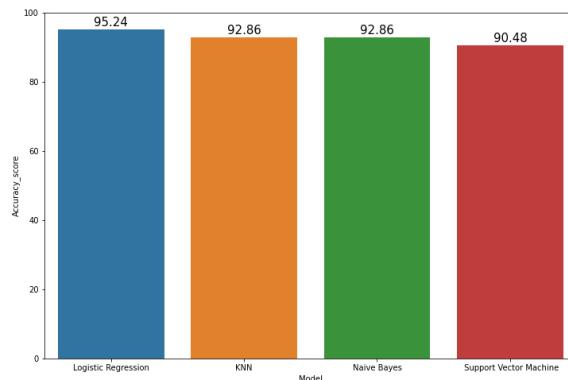
Seperti yang kita lihat Model terbaik diberikan oleh Logistic Regression (Akurasi 95%).

### Visualisasikan Hasil Akurasi

```

In [29]: 1 plt.figure(figsize=(12,8))
2 ax=sns.barplot(x="Model",y="Accuracy_score",data=result_df)
3 labels = (result_df["Accuracy_score"])
4 # add result numbers on barchart
5 for i, v in enumerate(result_df['Accuracy_score']):
6     ax.text(i, v+1, str(v), horizontalalignment = 'center', size = 15, color = 'black')

```



### Hasil Observasi:

Hal ini seperti yang diharapkan dapat terlihat pada heatmap di atas bahwa korelasi antara Sepal Width dan Sepal Length sangat rendah sedangkan korelasi antara Petal Width dan Petal Length sangat tinggi. Dengan model terbaik diberikan oleh Logistic Regression dengan akurasi 95%

Thank you!!