Session 29

Introduction to Machine Learning

Table of Content

What will We Learn Today?

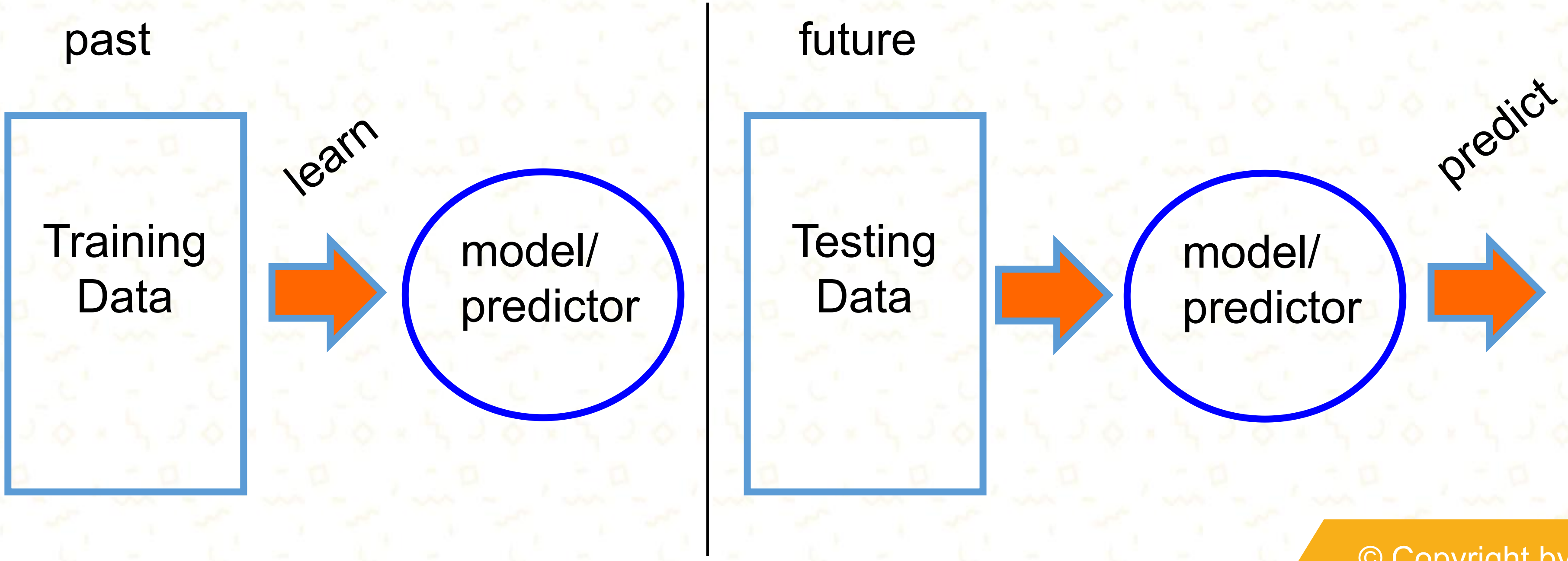
1. Machine Learning
2. ML approaches
3. Bias and Variance Tradeoff
4. Classification vs Regression
5. Logistic Regression

DigitalSkola
Uncover The World Of Digital Skills With Us

Machine Learning

What is Machine Learning

- Cabang kecerdasan buatan (Artificial Intelligence/ AI), yang berkaitan dengan desain dan pengembangan algoritma yang memungkinkan komputer mengembangkan perilaku berdasarkan data empiris.
- Karena kecerdasan membutuhkan pengetahuan, maka komputer perlu memperoleh pengetahuan.

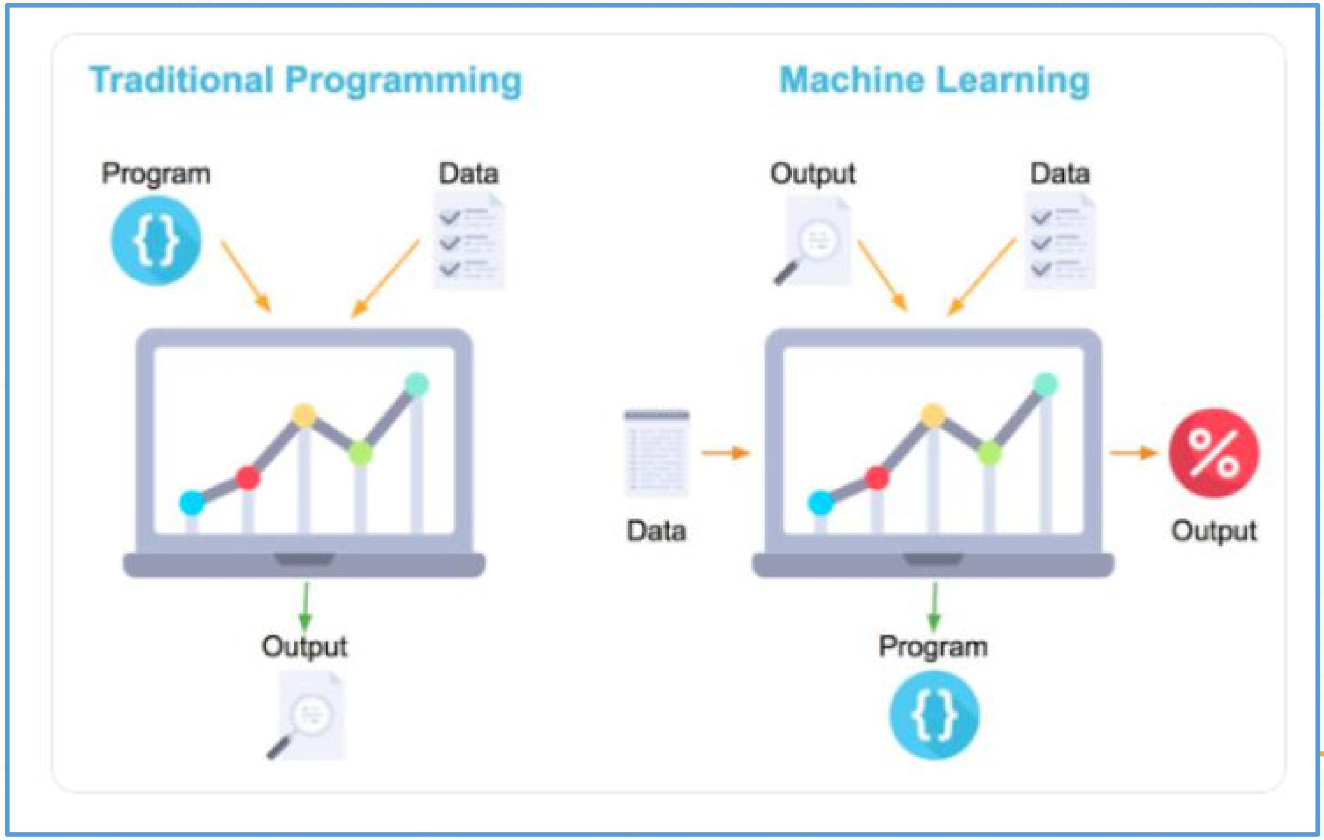


Why “Learn”?

- Machine learning adalah pemrograman komputer untuk mengoptimalkan kinerja menggunakan contoh data atau pengalaman masa lalu.
- Learning digunakan ketika:
  - Keahlian manusia tidak ada (navigating on Mars),
  - Manusia tidak mampu menjelaskan keahliannya (speech recognition)
  - Solusi yg perlu disesuaikan dengan kasus tertentu (user biometrics)

ML vs Traditional Programming

- Traditional programming adalah proses manual—artinya seseorang (programmer) membuat program.
- Sedangkan di machine learning, algoritma secara otomatis merumuskan aturan (rules) dari data.

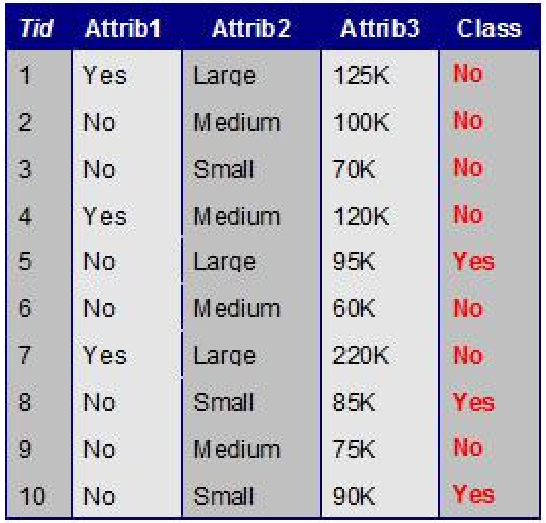


https://morioh.com/p/1063d43ef15e

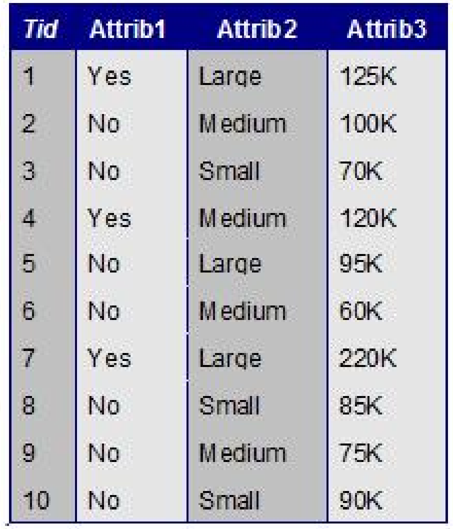
ML Approaches

Types of Learning

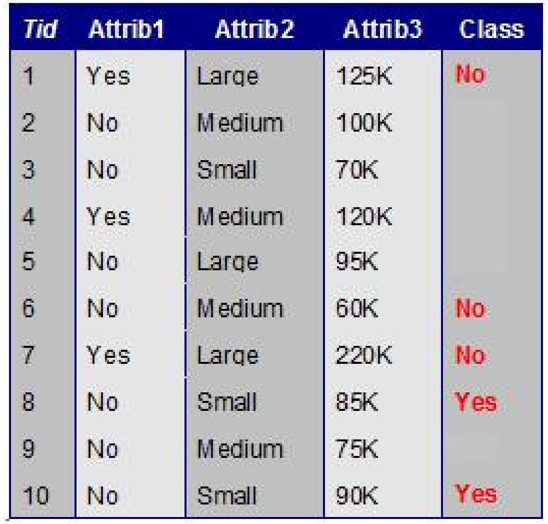
1. Supervised learning
   - Training data mempunyai target class
   - Classification, regression/ prediction
2. Unsupervised learning
   - Training data tidak mempunyai target class
   - Clustering
3. Semi-supervised learning
   - Sebagian training data memiliki outputs
4. Reinforcement learning
   - Rewards diberikan ketika agent mengerjakan tugas tertentu



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes



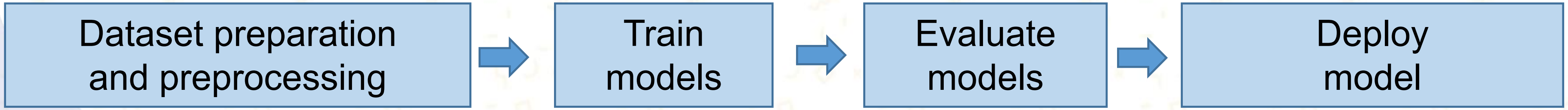
Tid	Attrib1	Attrib2	Attrib3
1	Yes	Large	125K
2	No	Medium	100K
3	No	Small	70K
4	Yes	Medium	120K
5	No	Large	95K
6	No	Medium	60K
7	Yes	Large	220K
8	No	Small	85K
9	No	Medium	75K
10	No	Small	90K



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	
3	No	Small	70K	
4	Yes	Medium	120K	
5	No	Large	95K	
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	
10	No	Small	90K	Yes

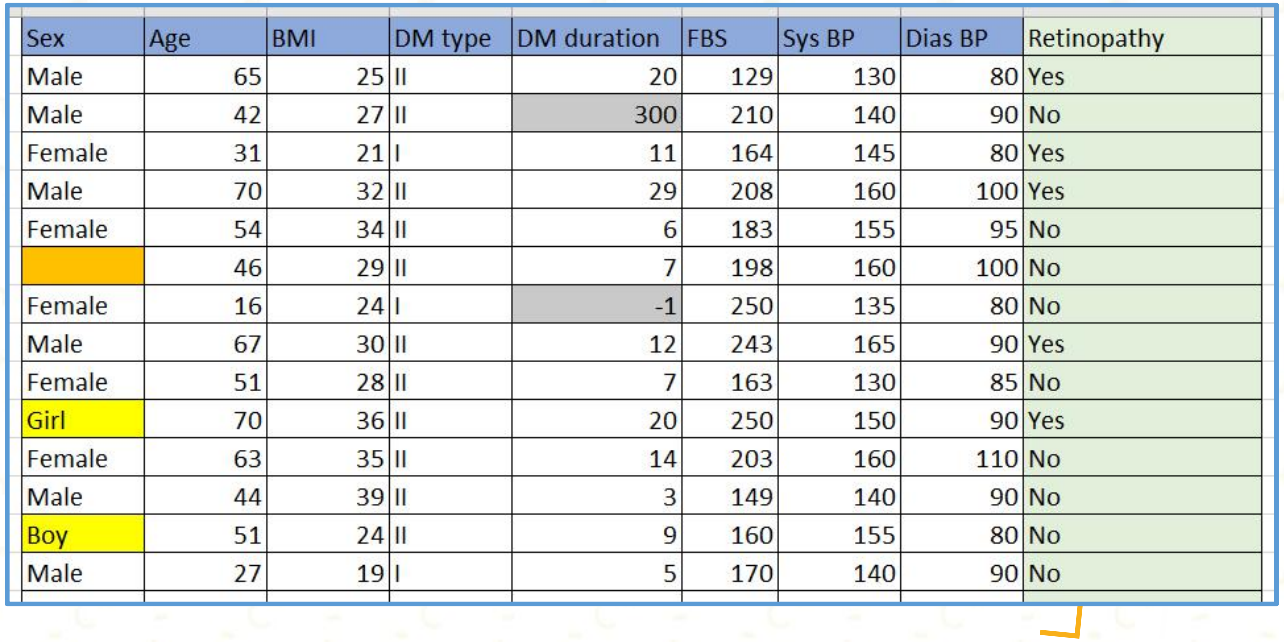
Stage in Machine Learning

- Data preprocessing
  - Data cleaning, filling missing value, remove outlier
- Train models
  - Select the algorithm
  - Feature selection and extraction
- Evaluate model
  - Assess performance
  - Model comparison
- Deploy model
  - Apply model to new data
  - Real-time demonstration



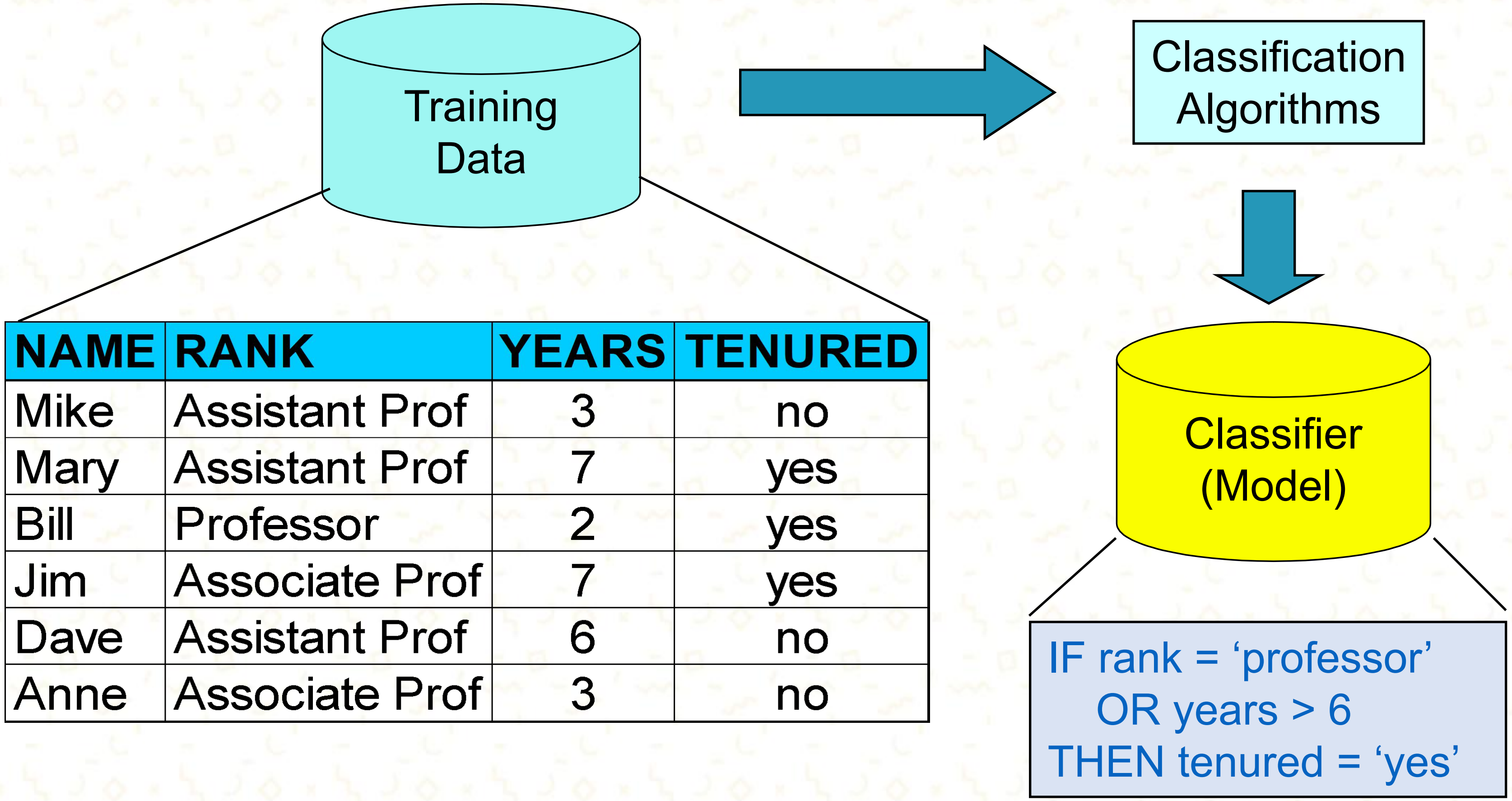
Why Data Preprocessing?

- Data in the real world is dirty
  - Missing or incomplete: lacking attribute values,
    - e.g., occupation=“ ”
  - Noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - Inconsistent: containing discrepancies in codes or names
    - e.g., sex=“Girl” vs. sex=“Female”
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data



Sex	Age	BMI	DM type	DM duration	FBS	Sys BP	Dias BP	Retinopathy
Male	65	25	II	20	129	130	80	Yes
Male	42	27	II	300	210	140	90	No
Female	31	21	I	11	164	145	80	Yes
Male	70	32	II	29	208	160	100	Yes
Female	54	34	II	6	183	155	95	No
	46	29	II	7	198	160	100	No
Female	16	24	I	-1	250	135	80	No
Male	67	30	II	12	243	165	90	Yes
Female	51	28	II	7	163	130	85	No
Girl	70	36	II	20	250	150	90	Yes
Female	63	35	II	14	203	160	110	No
Male	44	39	II	3	149	140	90	No
Boy	51	24	II	9	160	155	80	No
Male	27	19	I	5	170	140	90	No

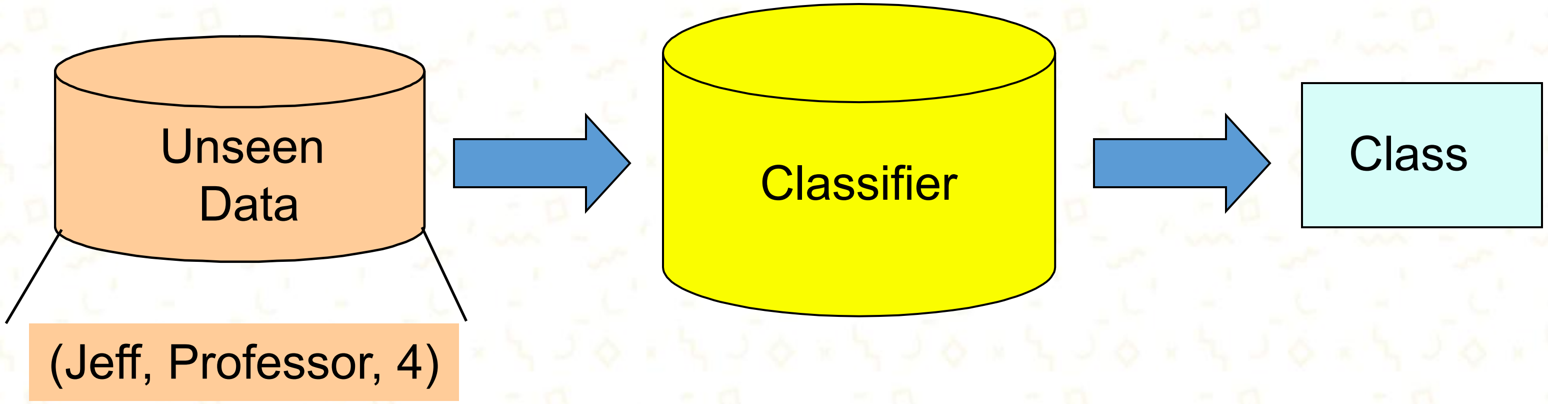
Model construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

```
IF rank = ‘professor’
   OR years > 6
THEN tenured = ‘yes’
```

Use the Model in Prediction



Tenured?

Yes

Bias and variance tradeoff

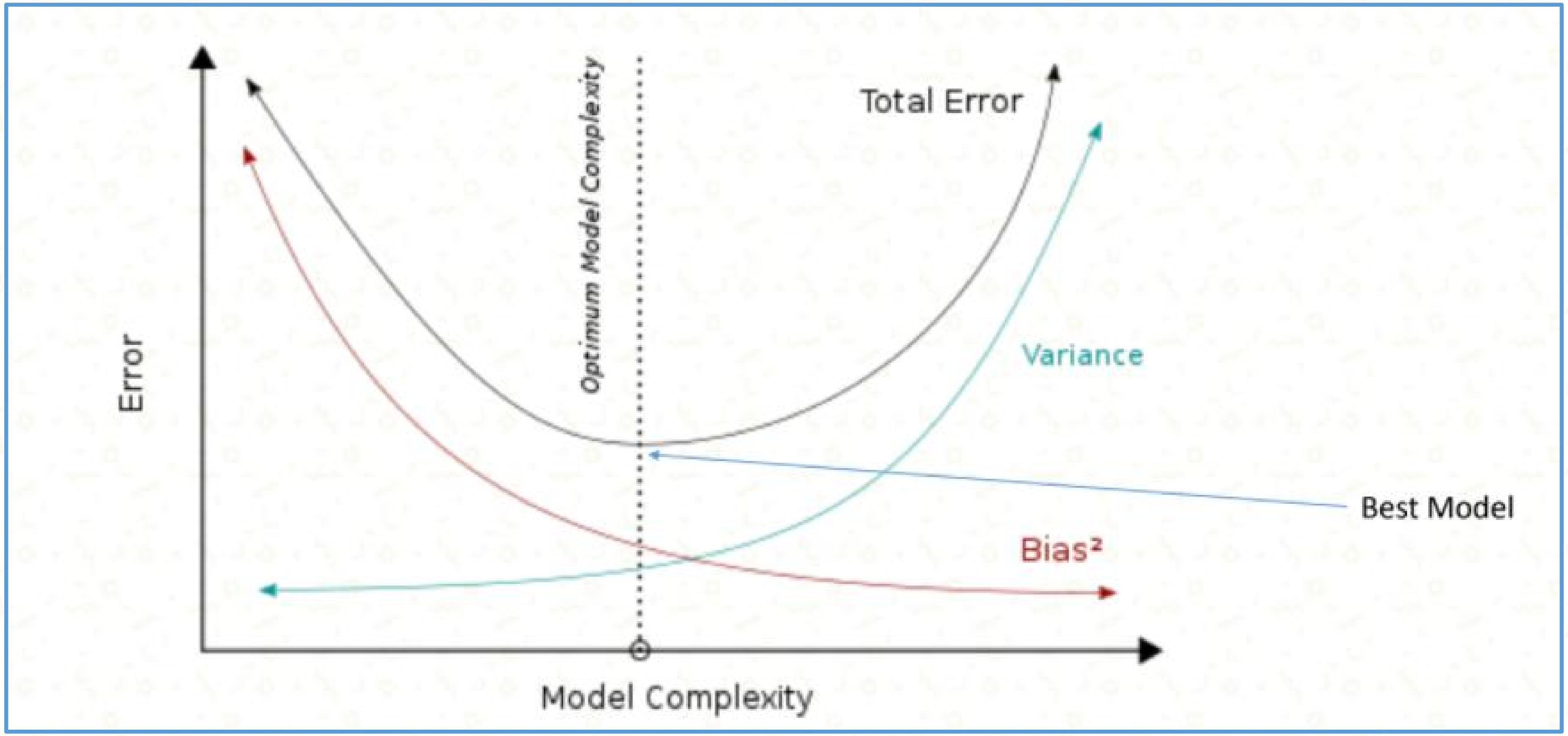
Bias and variance

- **Bias**
- Bias adalah perbedaan antara rata rata hasil prediksi dari model ML yang kita develop dengan data nilai yang sebenarnya.
- Bias yang tinggi dikarenakan dalam pembangunan model ML, dilakukan terlalu sederhana (oversimplified).

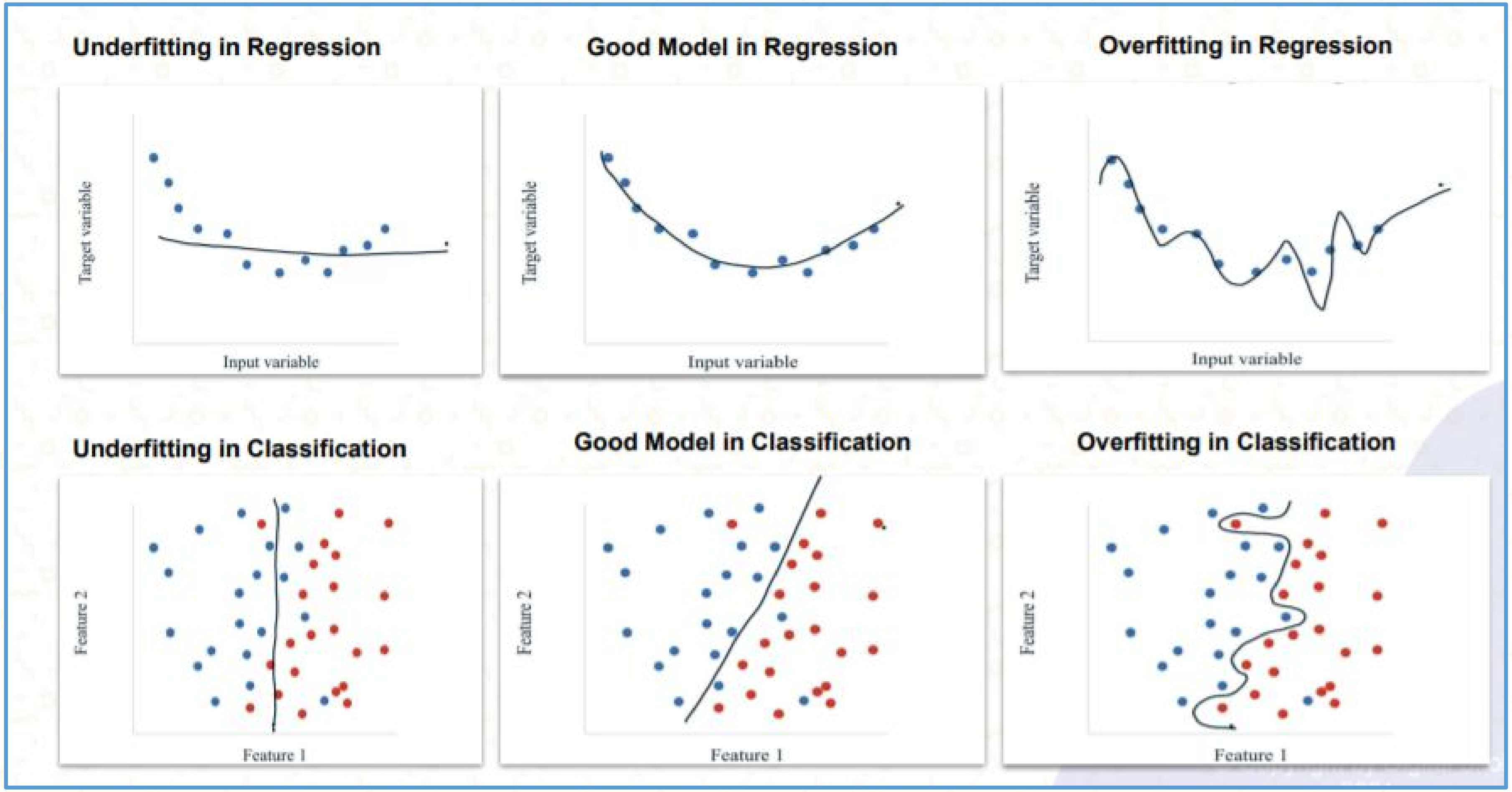
- **Variance**
- Variance adalah variabel dari prediksi yang memberikan kita informasi perserbaran data hasil prediksi.
- Model yang memiliki variance tinggi memiliki korelasi kuat hanya pada training set, sehingga akan berkinerja baik pada training data saja.

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Bias variance tradeoff



Underfitting and overfitting



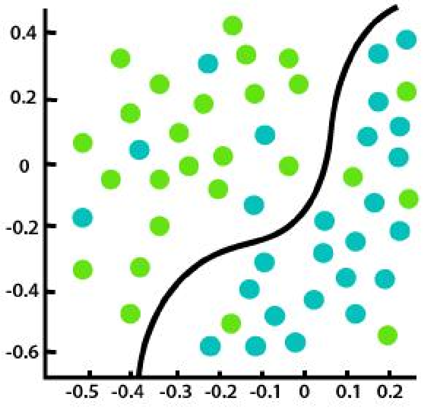
Classification vs Regression

Regression

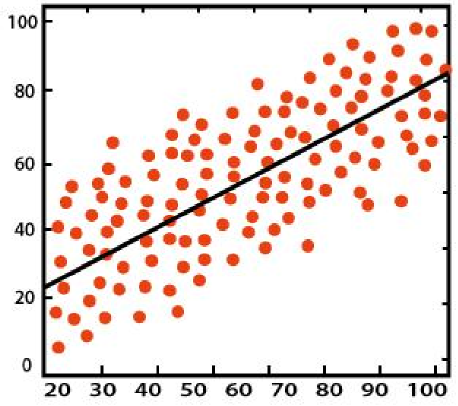
- *Regression* (regresi) = metode yang mencoba untuk menentukan kekuatan dan karakter hubungan antara satu variabel dependen dan serangkaian variabel lainnya (variabel independen).

- Algoritma regresi = nilai kontinu (seperti harga, gaji, usia, dll).
- Algoritma klasifikasi = nilai diskrit (seperti stroke atau normal, spam atau bukan spam, dll)

- Both are supervised learning

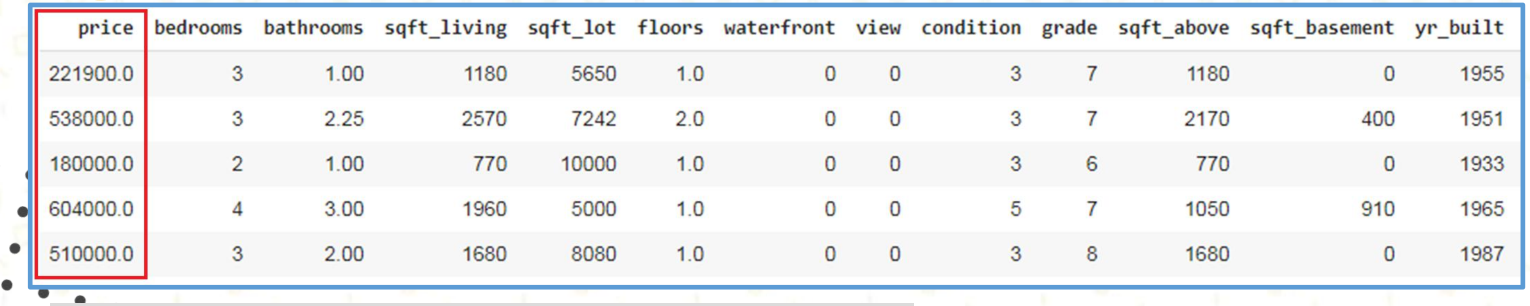


Classification



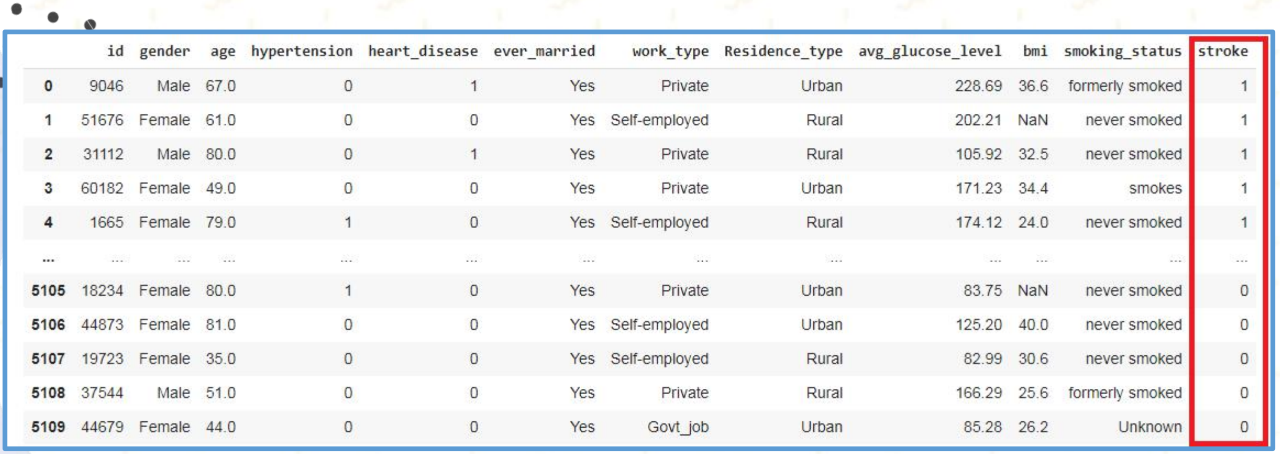
Regression

Classification, regression, clustering



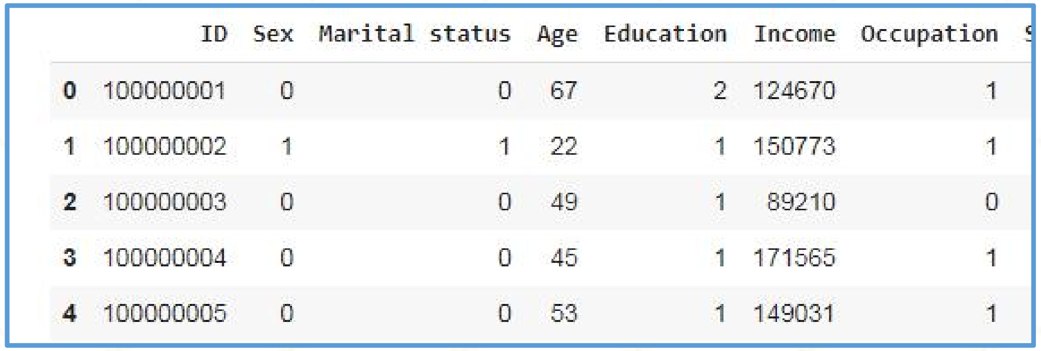
	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built
	221900.0	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955
	538000.0	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951
	180000.0	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933
	604000.0	4	3.00	1960	5000	1.0	0	0	5	7	1050	910	1965
	510000.0	3	2.00	1680	8080	1.0	0	0	3	8	1680	0	1987

Regression (house price dataset)



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

Classification (stroke dataset)

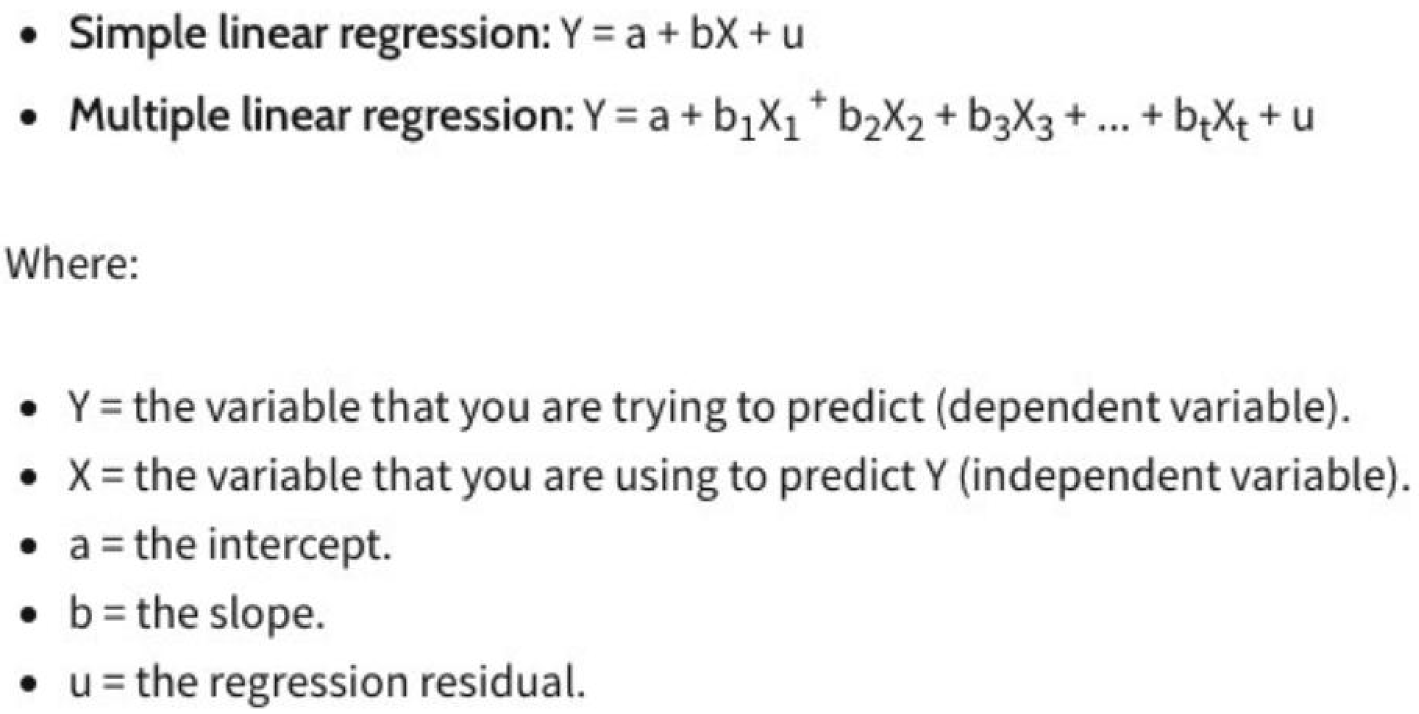


	ID	Sex	Marital status	Age	Education	Income	Occupation
0	100000001	0	0	67	2	124670	1
1	100000002	1	1	22	1	150773	1
2	100000003	0	0	49	1	89210	0
3	100000004	0	0	45	1	171565	1
4	100000005	0	0	53	1	149031	1

Clustering (customer dataset)

Linear Regression

- Membentuk hubungan antara dua variabel menggunakan garis lurus.



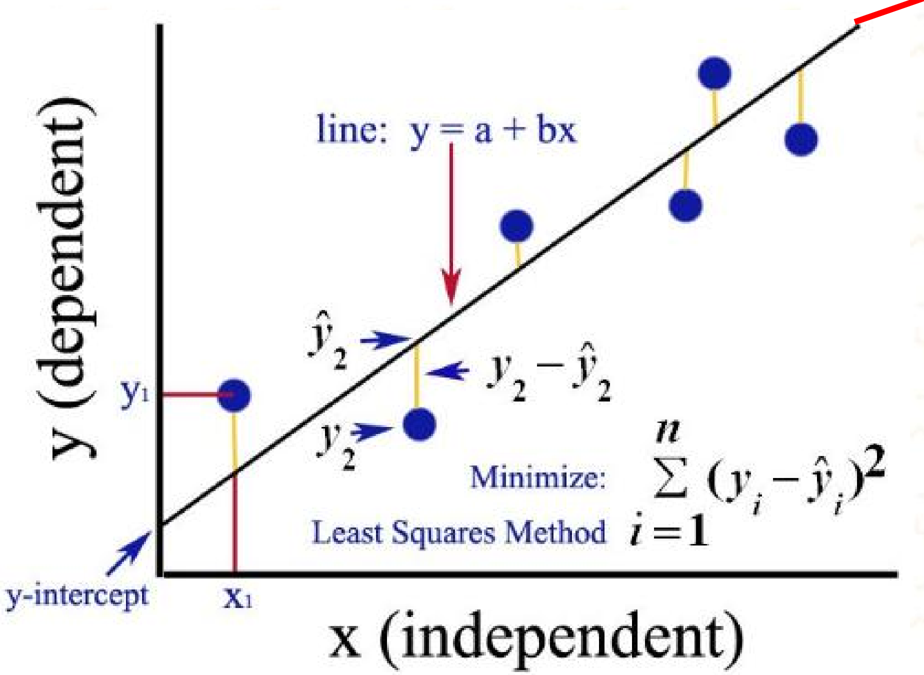
- **Simple linear regression:** $Y = a + bX + u$
- **Multiple linear regression:** $Y = a + b_1X_1 {}^{+} b_2X_2 + b_3X_3 + \ldots + b_tX_t + u$

Where:

- Y = the variable that you are trying to predict (dependent variable).
- X = the variable that you are using to predict Y (independent variable).
- a = the intercept.
- b = the slope.
- u = the regression residual.

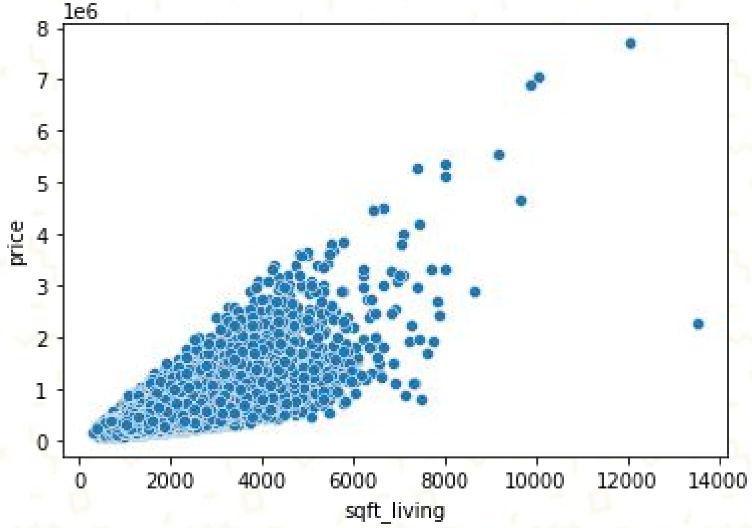
Linear Regression

- Regresi linier mencoba menggambar garis yang paling dekat dengan data dengan menemukan *slope* dan *intercept* dan meminimalkan *regression errors*.
- Ordinary Least Squares (OLS) adalah metode estimasi yang paling umum untuk model linier



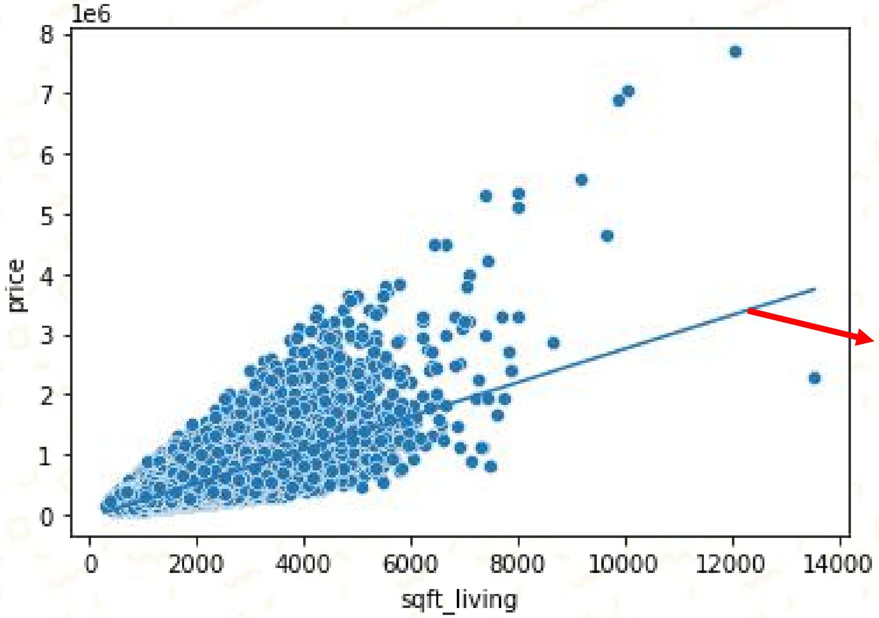
the best line would have the lowest sum of squared errors (SSE)

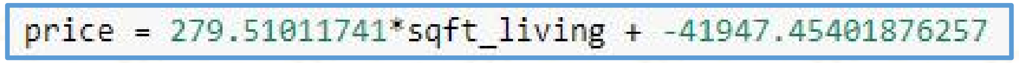
$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2$$



Linear Regression

- Example
  - y (dependent variable) = *price* (house price)
  - x (independent variable) = *sqft_living* (square feet)





```
price = 279.51011741*sqft_living + -41947.45401876257
```

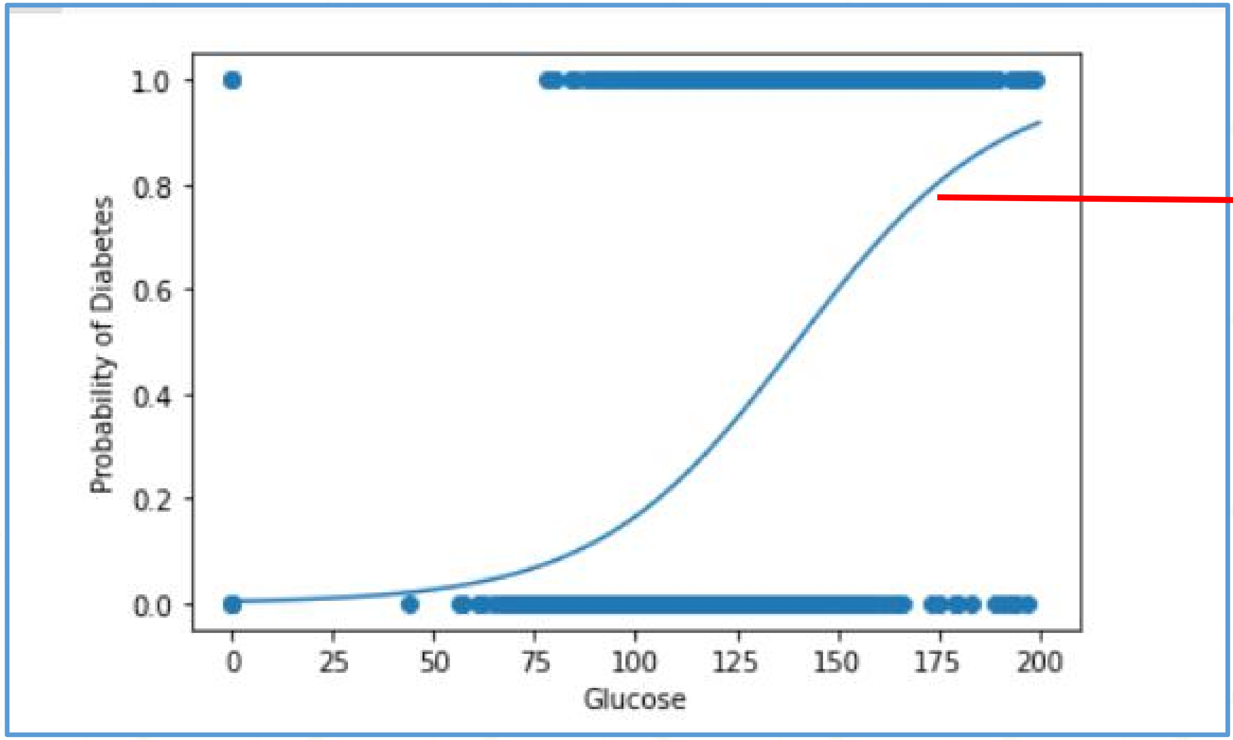
Q = House with 1000 square feet, approximate price?
A = USD 237562.663

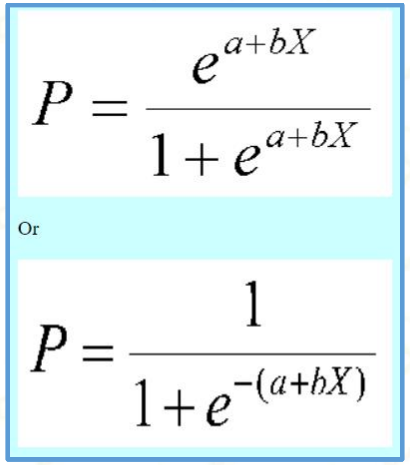
DigitalSkola
Uncover The World Of Digital Skills With Us

Logistic Regression

Logistic Regression

- Logistic Regression adalah algoritma klasifikasi Machine Learning yang digunakan untuk memprediksi ketika variabel dependen (target) adalah kategoris.
- Target adalah variabel biner yang berisi kelas 1 (untuk kasus benar/ya) atau 0 (untuk kasus salah/tidak).





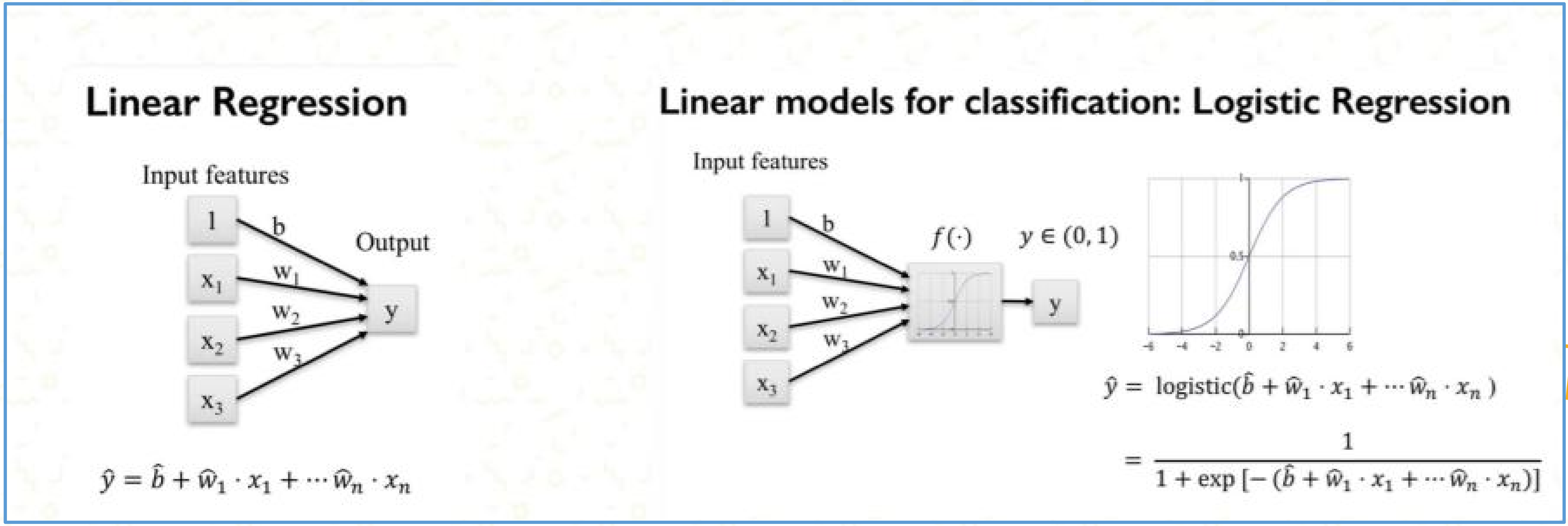
$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Or

$$P = \frac{1}{1+e^{-(a+bX)}}$$

Logistic Regression

- Merupakan sebuah kasus khusus regresi linier di mana responsnya adalah ‘log of odds’.
- Model Regresi Logistik memprediksi P(Y=1) dengan memasukkan data ke fungsi logit.



Linear Regression

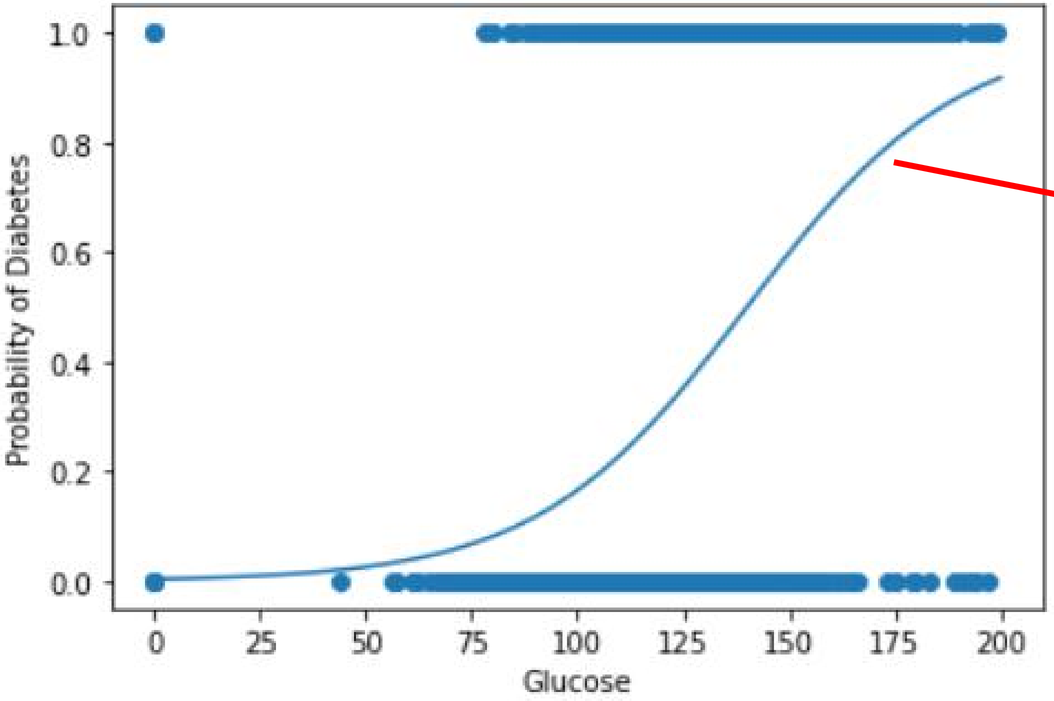
$$\hat{y} = \hat{b} + \hat{w}_1 \cdot x_1 + \cdots \hat{w}_n \cdot x_n$$

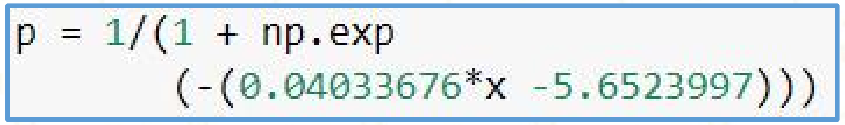
Linear models for classification: Logistic Regression

$$\hat{y} = \text{logistic}(\hat{b} + \hat{w}_1 \cdot x_1 + \cdots \hat{w}_n \cdot x_n)$$

$$= \frac{1}{1 + \exp\left[-(\hat{b} + \hat{w}_1 \cdot x_1 + \cdots \hat{w}_n \cdot x_n)\right]}$$

Logistic Regression

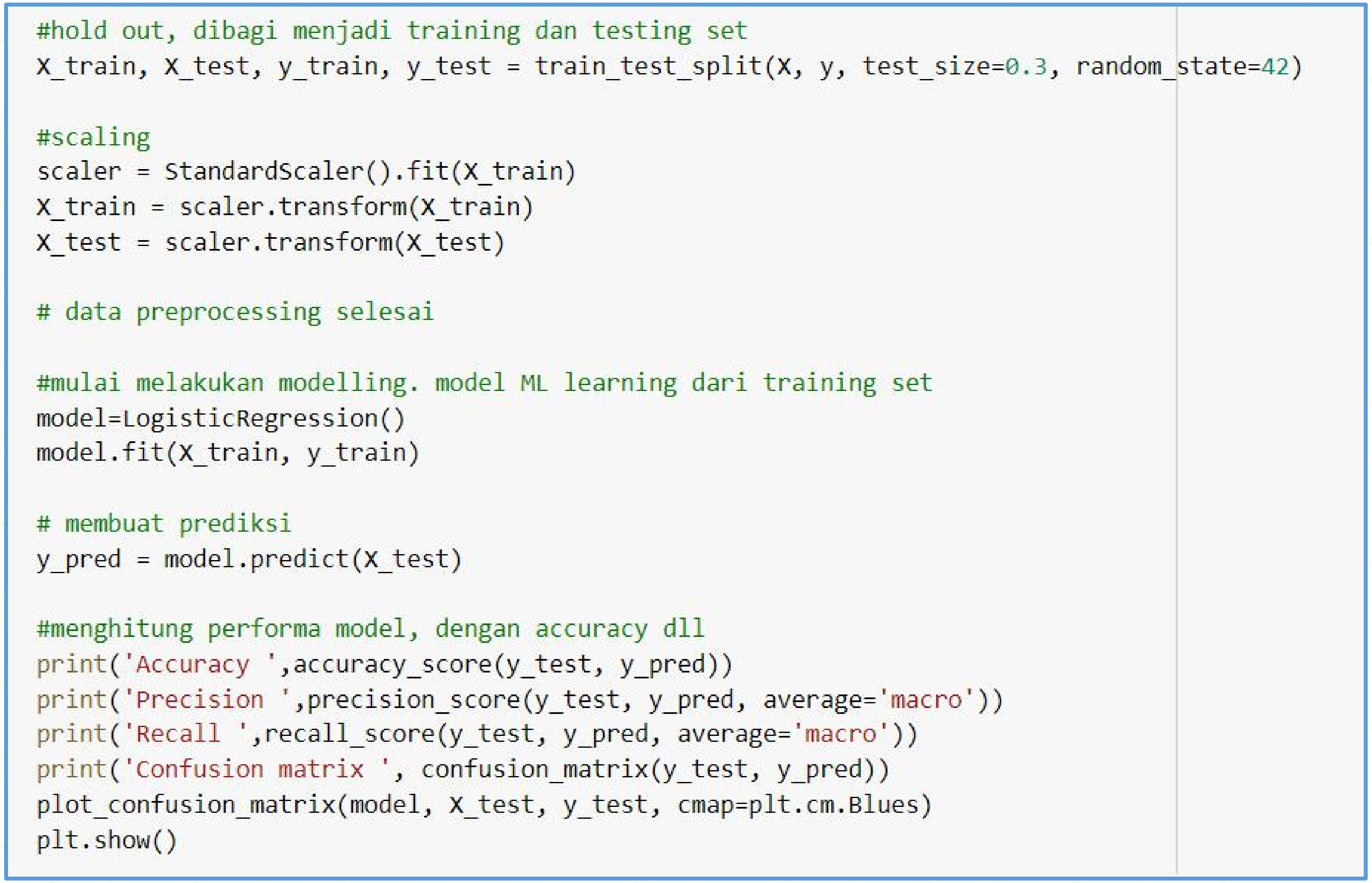




```
p = 1/(1 + np.exp
       (-(0.04033676*x -5.6523997)))
```

Q = Patient with BG 190 mg/dL, is it diagnosed as diabetes?
A = Probability diabetes is 0.882

Logistic Regression



```
#hold out, dibagi menjadi training dan testing set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

#scaling
scaler = StandardScaler().fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)

# data preprocessing selesai

#mulai melakukan modelling. model ML learning dari training set
model=LogisticRegression()
model.fit(X_train, y_train)

# membuat prediksi
y_pred = model.predict(X_test)

#menghitung performa model, dengan accuracy dll
print('Accuracy ',accuracy_score(y_test, y_pred))
print('Precision ',precision_score(y_test, y_pred, average='macro'))
print('Recall ',recall_score(y_test, y_pred, average='macro'))
print('Confusion matrix ', confusion_matrix(y_test, y_pred))
plot_confusion_matrix(model, X_test, y_test, cmap=plt.cm.Blues)
plt.show()
```

DigitalSkola
Uncover The World Of Digital Skills With Us

Thank YOU