Mission: Movie Analytics for Data Scientist

Supervised Machine Learning

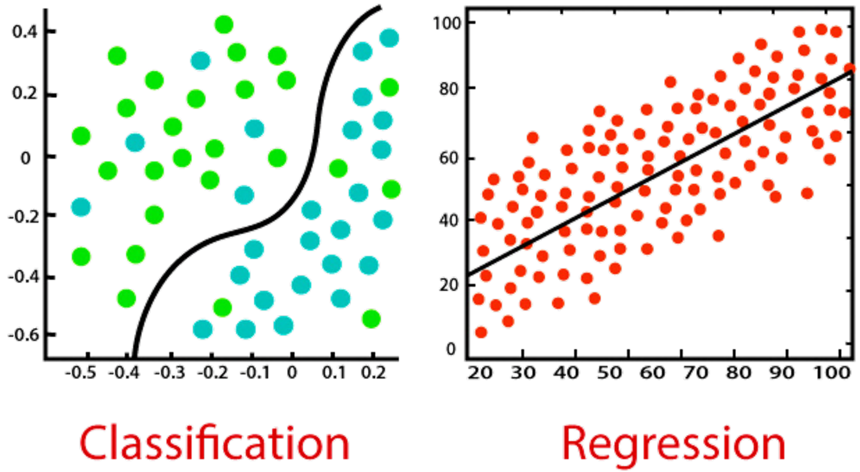
Regression vs Classification

Regression

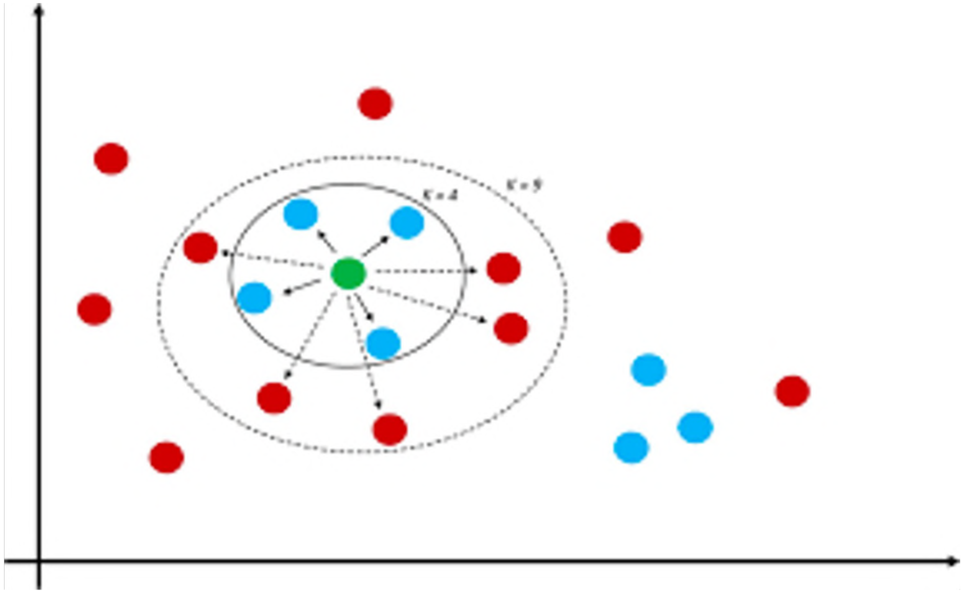
Prediction when the target feature is a continuous numerical variable

Classification

Prediction when the target feature is a categorical variable



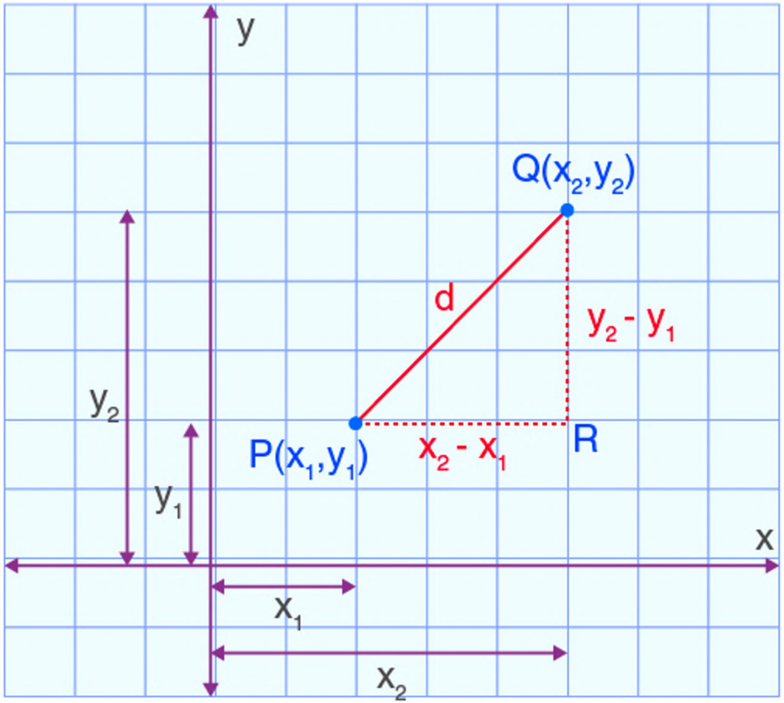
K Nearest Neighbor



Konsep

1. Menghitung jarak antara data baru dengan data yang ada
2. Menghitung *k* data terdekat
3. Memprediksi berdasarkan *majority voting*

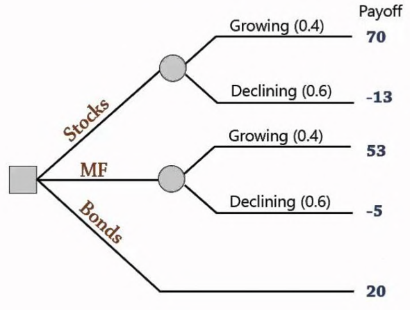
Distance



Decision Tree

Decision Tree

Alternatives	Growing	Declining
Stocks	70	-13
Mutual Funds	53	-5
Bonds	20	20
Probability	0.4	0.6



Konsep

1. Membagi data berdasarkan target
2. Menghitung *loss function* yang ada
3. Memilih *feature* dengan loss function terkecil

Model Evaluation

Confusion Matrix

		Actual	
		0 (Negative)	1 (Positive)
Prediction	0 (Negative)	True Negative (1-α)	False Negative (**β**)
	1 (Positive)	False Positive (α)	True Positive (1-**β**)

Accuracy

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

		Actual	
		0 (Negative)	1 (Positive)
Prediction	0 (Negative)	88	9
	1 (Positive)	2	1

Accuracy = **89/100 (89%)**

Recall

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

		Actual	
		0 (Negative)	1 (Positive)
Prediction	0 (Negative)	88	9
	1 (Positive)	2	1

Recall = **1/10 (10%)**

Precision

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$

		Actual	
		0 (Negative)	1 (Positive)
Prediction	0 (Negative)	88	9
	1 (Positive)	2	1

Precision = **1/3 (33%)**

F1 Score

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$


Threshold Tuning

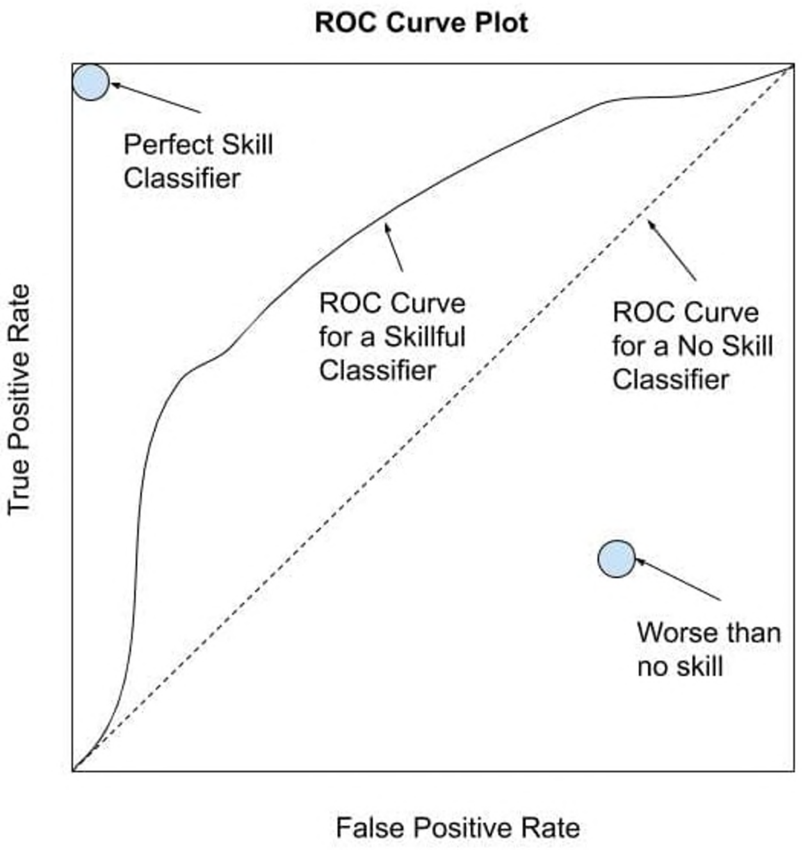
Probability

```python
proba = pd.DataFrame(lr.predict_proba(X))
```

fx 0,91

A	B	C
prob kelas 0	prob kelas 1	actual
0,09	0,91	1
0,08	0,92	1
0,65	0,35	0
0,09	0,91	1
0,18	0,82	1
0,63	0,37	1
0,43	0,57	1
0,31	0,69	0
0,13	0,87	1
0,73	0,27	0
0,57	0,43	0
0,12	0,88	1
0,69	0,31	0
0,05	0,95	1

ROC - AUC


ROC Curve Plot
Perfect Skill Classifier
ROC Curve for a Skillful Classifier
ROC Curve for a No Skill Classifier
Worse than no skill
True Positive Rate
False Positive Rate


Thank You!