

E-mail Classification using Machine Learning & Python



Gönül Aycı
Python Saati #94
July 16, 2019





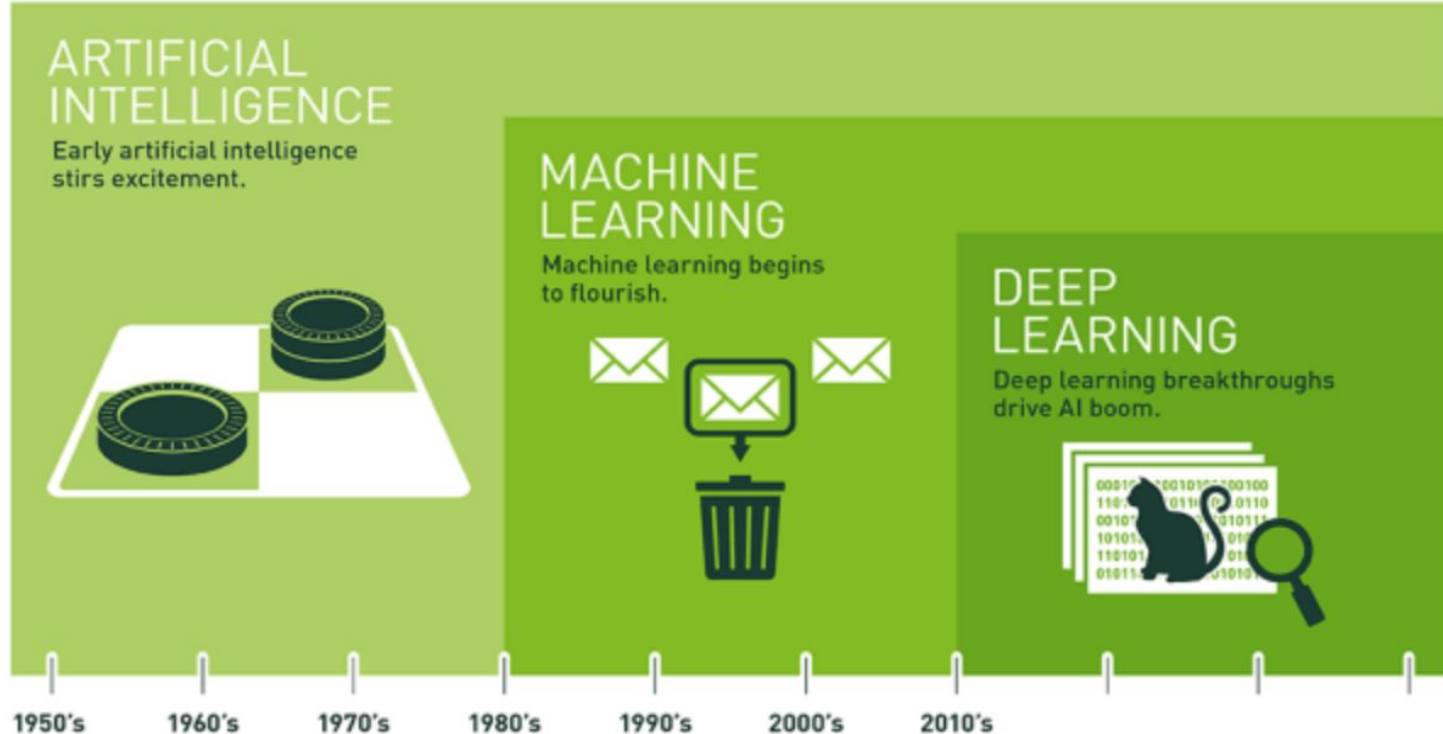
Spam

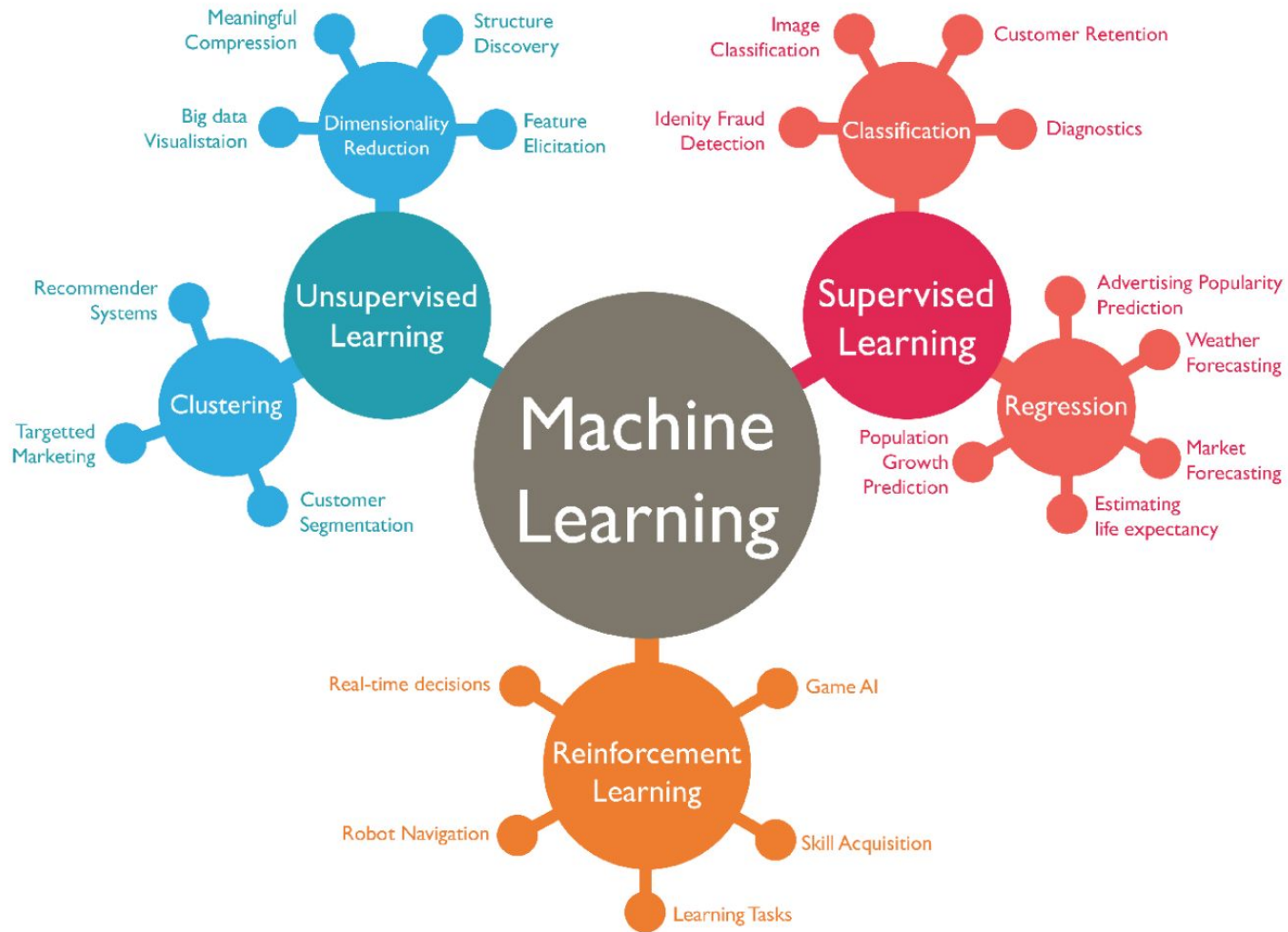


Legitimate

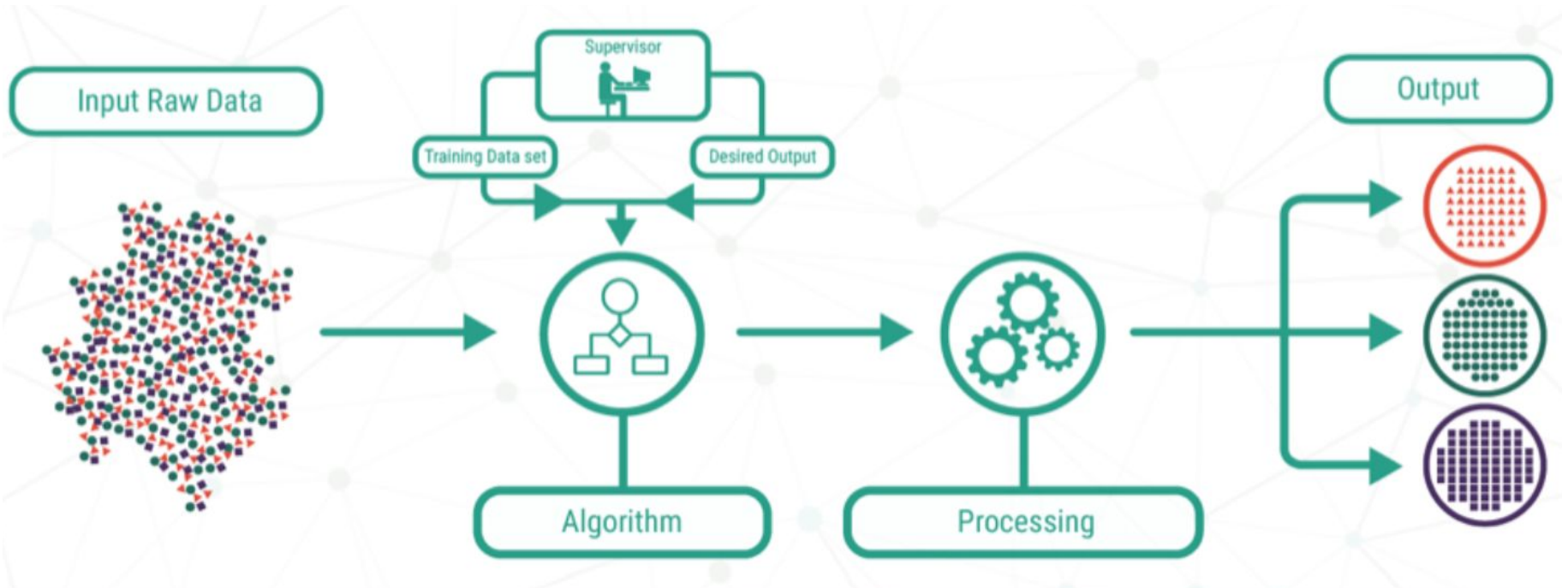


- ‘Yapay öğrenme, veriye dayalı öğrenimini olanaklı kılan algoritmaların tasarım ve geliştirme süreçlerini konu edinen bir bilim dalıdır.’



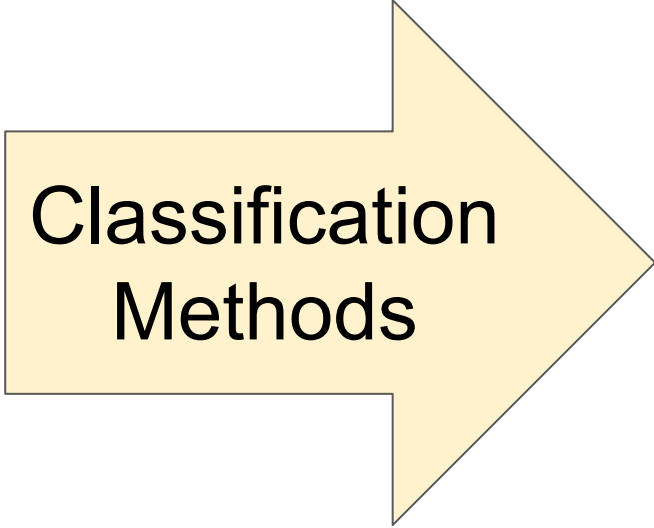


Supervised Learning



Classification

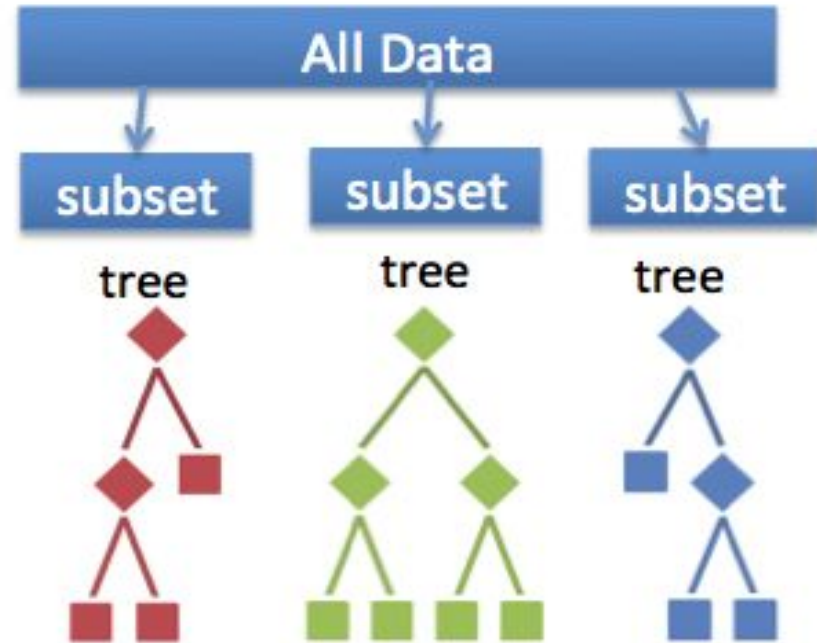


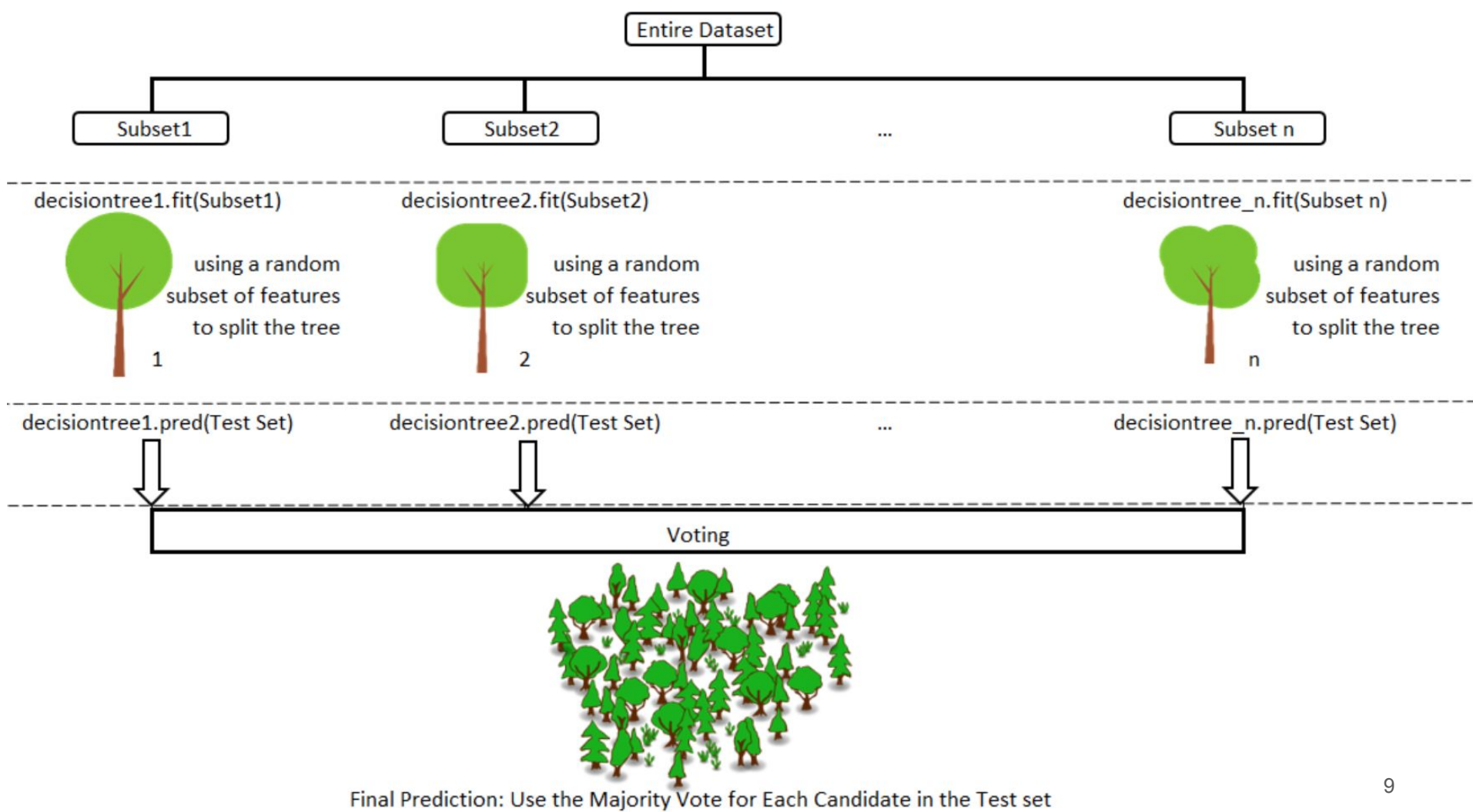


Classification Methods



- Linear Classifiers
 - Logistic Regression
 - Naive Bayes Classifiers
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Decision Tree
- Ensemble Methods
 - Random Forest
 - Gradient Boosting
- Neural Networks

Random Forest





Prepare Input Matrix

	order	address	our	mail	report	one	send	program	list
	0	5	3	1	0	0	1	0	0
	0	0	8	1	0	0	1	5	0

Dataset



Pre-processing

Natural
Language
ToolKit



- Remove Stopwords
- Lemmatization
- Stemming
- Tokenization
- Remove Punctuations
- ...

Lemmatization

- *am, are, is* → *be*
- *car, cars, car's, cars'* → *car*

the boy's cars are different colors → *the boy car be different color*

Train Dataset

Subject: franz boa

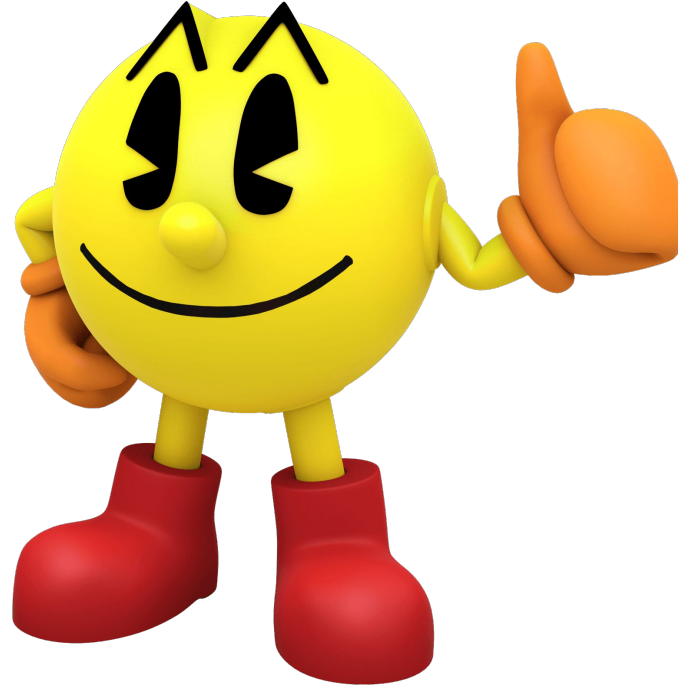
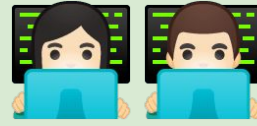
dear fellow linguist . graduate student taichung ,
taiwan . recently , interest boa ' life , personality ,
work . easy material here . anyone suggestion ,
(maybe) material . thank much advice . rose huang

Test Dataset

Subject: re : information requested

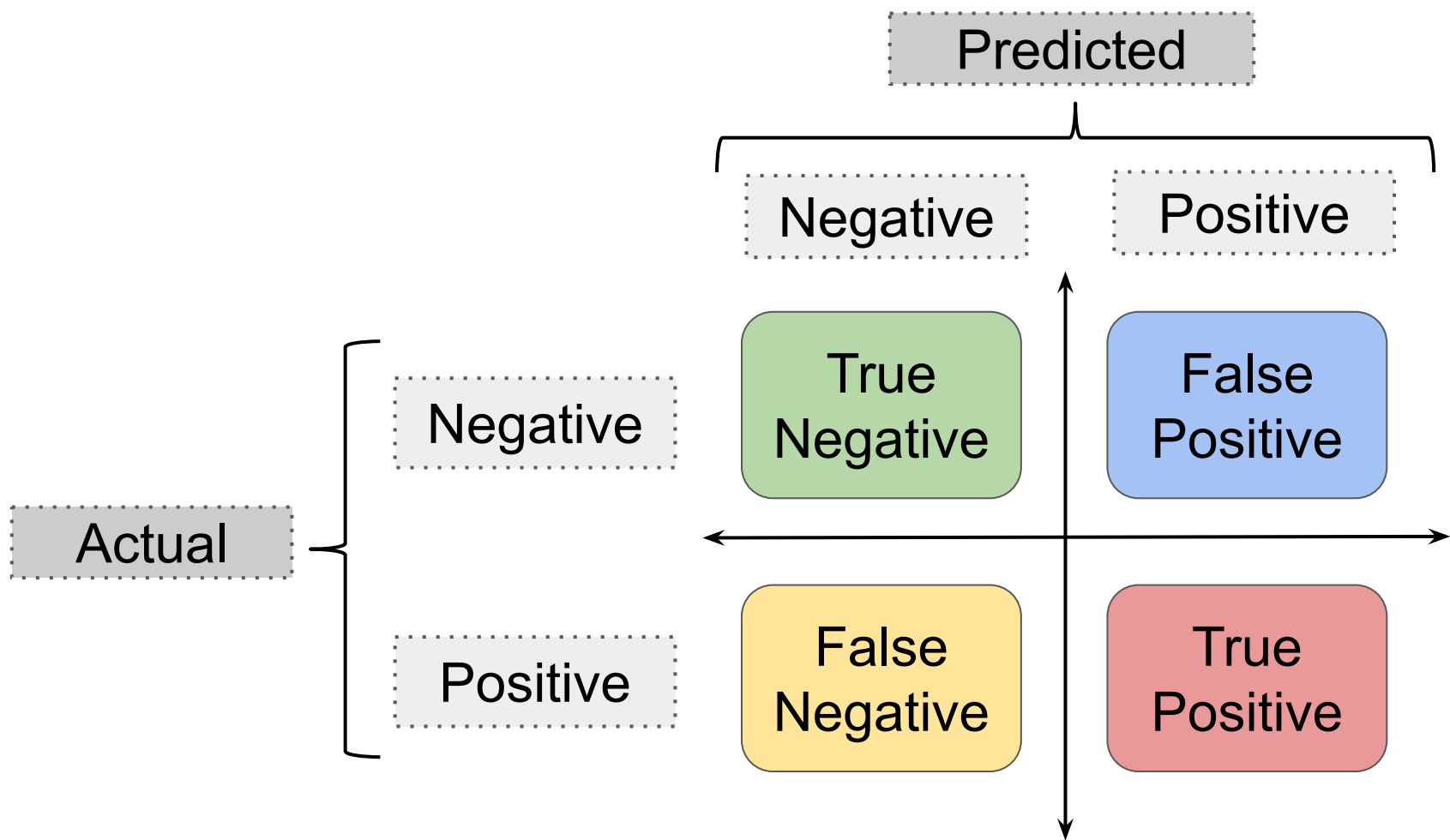
hi , name john ' m 27 old . able \$ 2 million work home ,
'd share . please few moment busy life listen short
message tell ! call listen , 1-800 - 764-6203 change
life !

Demo

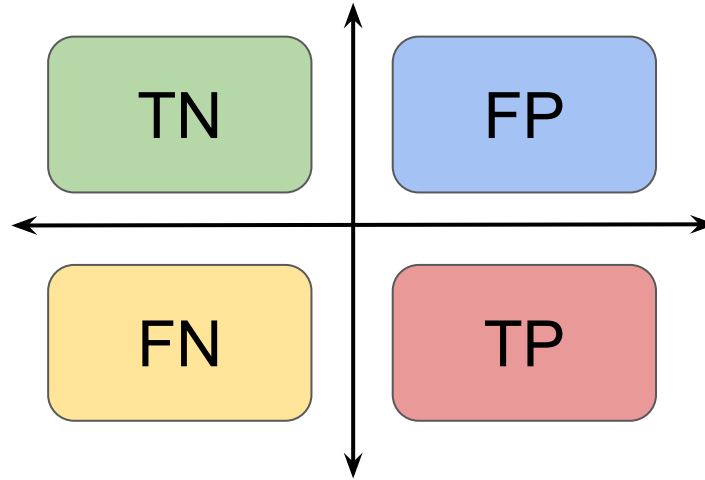


Evaluate Performance of Random Forest Classifier





$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}}$$



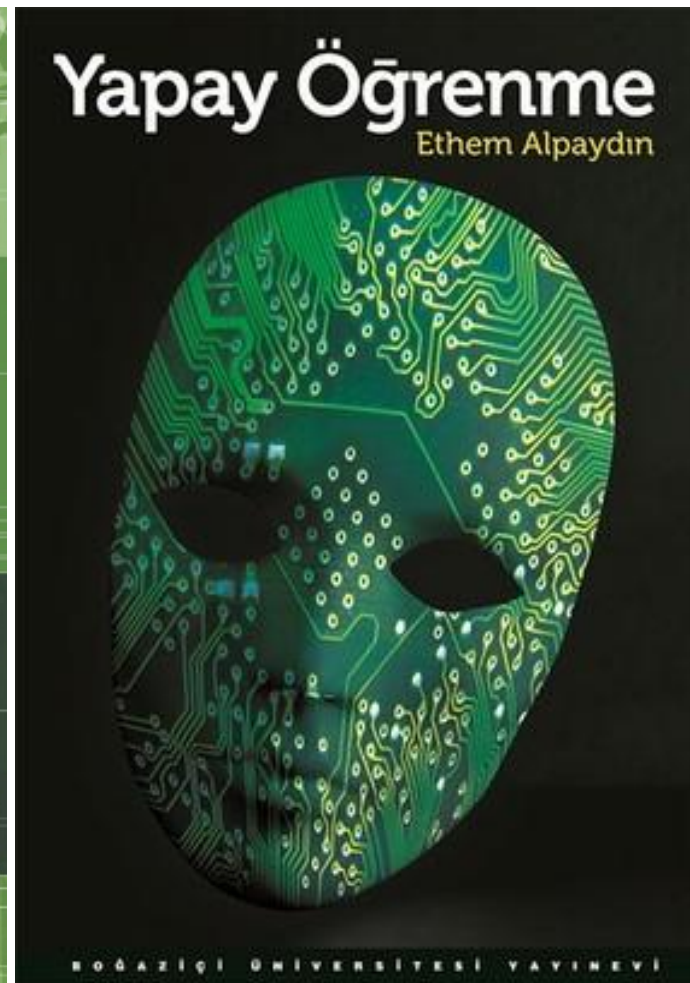
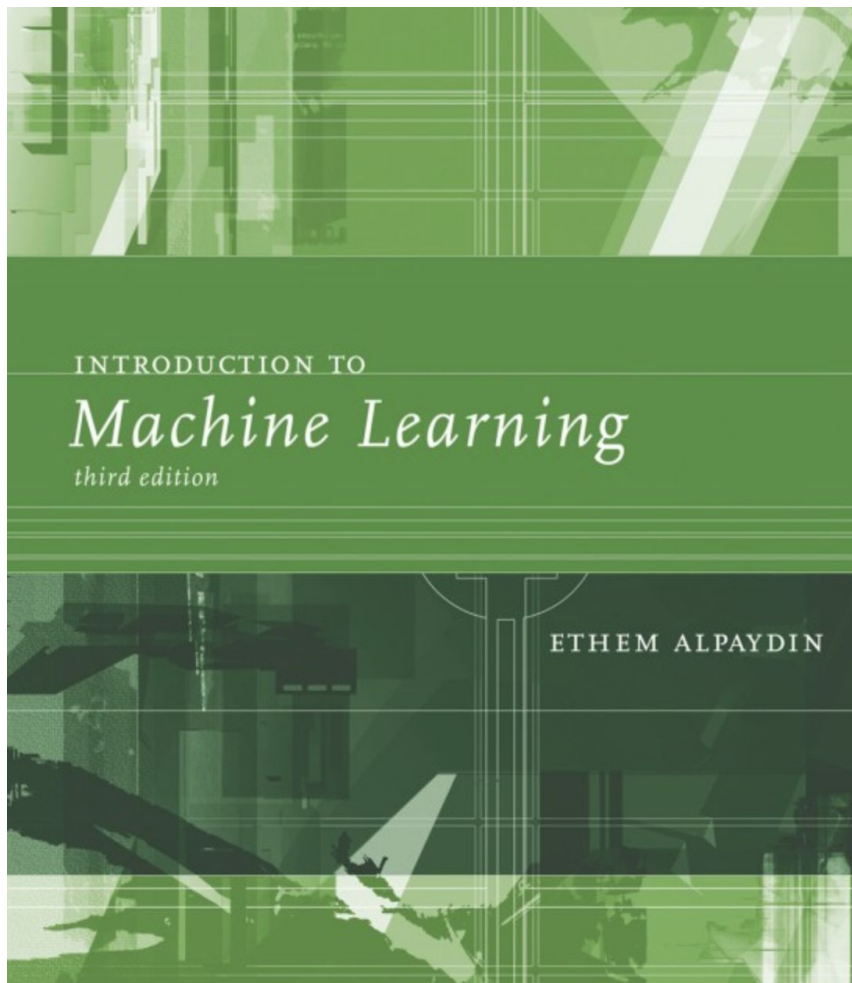
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

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

Demo







Machine Learning



Machine Learning

by Andrew Ng

References



- I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos. ***“An Evaluation of Naive Bayesian Anti-Spam Filtering.”*** In Potamias, G., Moustakis, V. and van Someren, M. (Eds.), Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000), Barcelona, Spain, pp. 9-17, 2000.
- <http://tiny.cc/swep9y>, <http://tiny.cc/29hr9y>, <http://tiny.cc/4wep9y>, <http://tiny.cc/28ip9y>, <http://tiny.cc/o1ep9y>, <http://tiny.cc/0n9r9y>, <http://tiny.cc/jses9y>

Thank You For Listening 🤗
Any Questions? 🤔



aycignl



gonul_ayci



aycignl

