scikit-learn

Praktik Data Mining Menggunakan scikit-learn

Ahmad Aulia Wiguna

Ahmad Aulia Wiguna, M.Kom

Kendal, 18 April 1992

- SDN 4 Langenharjo
- SMPN 2 Kendal
- SMKN 2 Kendal jurusan Teknik Mesin
- D3 - UDINUS
- S1 – UDINUS
- S2 – UDINUS
- S3 – UGM (Usaha Gagean Married)

- Asisten lab Udinus Semarang 2010-2013
- IT Support PT Petraya Mitrajaya 2011-2012
- Senior Back End Programmer PT Dinustek Semarang (2014-now)

Phone/WA/Telegram : 085-640-243-001
https://github.com/auliawiguna
https://id.linkedin.com/in/wiguna-ahmad-aulia

What is learning, anyway? What is *machine* learning?

Machine Learning

Seperangkat ~~alat sholat~~ alat dan metode yang mencoba menyimpulkan pola dan mengambil pengetahuan dari *record-record* yang diamati (Conway, 2012)

Ilmu yang mempelajari bagaimana computer dapat belajar berdasar data yang diberikan *(Han et al., 2012)*

Machine Learning

- Supervised learning
- Unsupervised learning

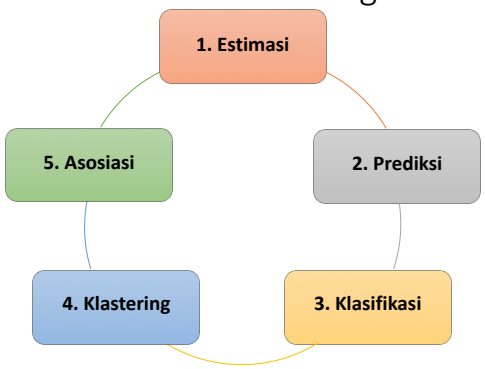
Data Mining

- Dapat didefinisikan dengan banyak sebutan
- Sebuah topik yang melibatkan praktek *learning (Witten et al., 2011)*
- Melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data *(Witten et al., 2011)*
- Kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar *(Santosa, 2007)*

Data? Informasi?

1. Data: fakta yang terekam dan tidak membawa arti
2. Pengetahuan: pola, rumus, aturan atau model yang muncul dari data

Peran Utama Data Mining



Contoh Pemanfaatan Data Mining

PREDIKSI KEMENANGAN BOT DOTA 2

MENGGUNAKAN METODE NAIVE BAYES

DOTA 2 BOT VICTORY PREDICTION USING NAIVE BAYES

Nanang Budi Kurniawan[1], Pulung Nurtantio Andono[2]
Universitas Dian Nuswantoro Semarang
Jl. Nakula I No 5-11 Semarang 50131
Telp : (024) 35176361, Fax : (024) 3520165
[1; 2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer
e-mail: [*1] nb.kurniawan@gmail.com, [2] pulung@dsn.dinus.ac.id

Abstrak

Dota 2 adalah salah satu game multiplayer yang banyak dimainkan saat ini. Dalam memainkan game dota 2, pemain dibagi menjadi dua kelompok yang saling bertempur, terdiri dari Radiant dan Dire. Masing-masing kelompok terdiri dari lima pemain. Didalam game dota 2 terdapat jenis permainan custom scenario dimana hero musuh dijalankan oleh artificial intelligences (AI). Ketika memainkan versi scenario, pemain bisa dibantu oleh hero dota sekutu yang dijalankan oleh AI game. Masing-masing team memiliki markas utama yang harus dilindungi, markas ini disebut Ancient. Ketika ancient salah satu team hancur, maka pertandingan selesai. Naive bayes adalah algoritma data mining klasifikasi yang digunakan untuk prediksi kemenangan bot dalam pertandingan dota 2. Dengan pembobotan statistik pada masing-masing bot akan dilakukan prediksi kemenangan bot pada game dota 2 sehingga akan diketahui team radiant atau dire yang akan memenangkan pertandingan dalam game dota 2. Akurasi yang dihasilkan naive bayes untuk memprediksi kemenangan team dota 2 adalah sebesar 54.17%.

Kata kunci: *dota 2, artificial intelligences (AI), team, naive bayes, prediksi.*

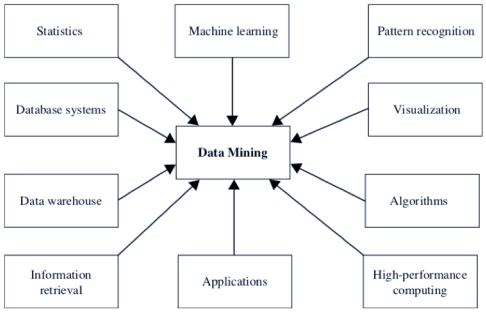
Contoh Pemanfaatan Data Mining

PERTANDINGAN 6			VS			
	DIRE	BOUNTY HUNTER	57	MELEE	AGI	MENANG
		NECROPOS	53	RANGE	INT	
		RAZOR	50	RANGE	AGI	
		DRAGON KNIGHT	53	MELEE	STR	
	RADIANT	VENGEFULL SPIRIT	68	RANGE	AGI	MENANG
		JUGGERNAUT	60	MELEE	AGI	
		DRAGON KNIGHT	53	MELEE	STR	
		WARLOCK	56	RANGE	INT	
PERTANDINGAN 7			VS			
	DIRE	BANE	66	RANGE	INT	KALAH
		SHADOW FIEND	53	RANGE	AGI	
		TINY	52	MELEE	STR	
		KUNKA	56	MELEE	STR	
	RADIANT	PHANTOM ASSASSIN	58	MELEE	AGI	MENANG
		WINDRANGER	54	RANGE	INT	
		CHAOS KNIGHT	50	MELEE	STR	
		SVEN	60	MELEE	STR	
PERTANDINGAN 8			VS			
	DIRE	PUDGE	53	MELEE	STR	KALAH
		LICH	51	RANGE	INT	
		LUNA	49	RANGE	AGI	
		JAKIRO	63	RANGE	INT	
	RADIANT	BANE	66	RANGE	INT	KALAH
		DRAGON KNIGHT	53	MELEE	STR	
		JUGGERNAUT	60	MELEE	AGI	
		SNIPER	52	RANGE	AGI	
PERTANDINGAN 9			VS			

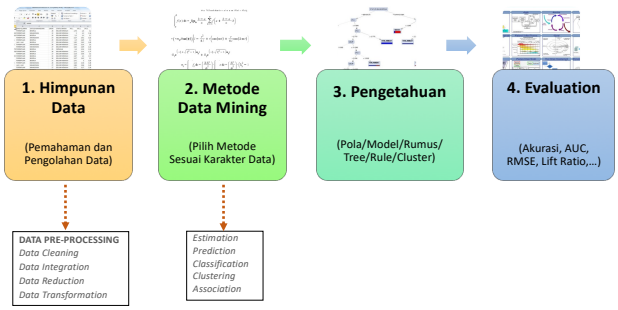
Teknologi Apa Sih Yang Dipakai?

- Statistik
- ML (Machine Learning)
- Database system/ data warehouse
- IR (Information Retrieval)

Teknologi Apa Sih Yang Dipake?



Proses Data Mining



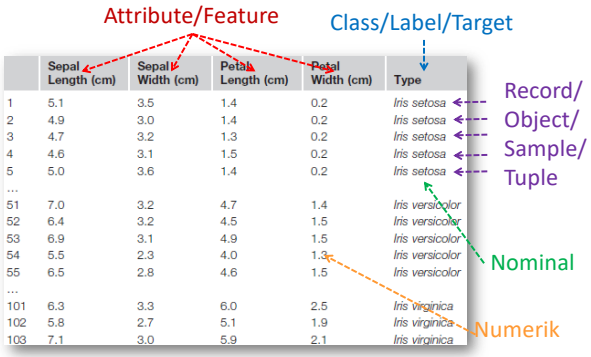
13

Dataset (Himpunan Data)

- Atribut adalah faktor atau parameter yang menyebabkan class/label/target terjadi
- Jenis dataset ada dua: Private dan Public
- Private Dataset: data set dapat diambil dari organisasi yang kita jadikan obyek penelitian
  - Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc
- Public Dataset: data set dapat diambil dari repositori pubik yang disepakati oleh para peneliti data mining
  - UCI Repository (*http://www.ics.uci.edu/~mlearn/MLRepository.html*)
  - ACM KDD Cup (*http://www.sigkdd.org/kddcup/*)
  - PredictionIO (*http://docs.prediction.io/datacollection/sample/*)
- Trend penelitian data mining saat ini adalah menguji metode yang dikembangkan oleh peneliti dengan public dataset, sehingga penelitian dapat bersifat: comparable, repeatable dan verifiable

14

Dataset (Himpunan Data)



	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
...					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*

15

Top Programming Language for Data Mining

1. R
2. Python
3. JULIA Lang
4. JAVA
5. KAFKA

Source : https://www.freelancinggig.com/blog/2017/06/19/top-5-analyticsdata-mining-programming-languages/

1. Python
2. JAVA
3. R
4. JULIA Lang
5. SAS
6. SQL
7. MATLAB
8. SCALA
9. C
10. F#

Source : https://analyticsindiamag.com/top-10-programming-languages-data-scientists-learn-2018/

Pemakaian Dalam *Coding*

- R
- PHP-ML (https://php-ml.readthedocs.io/en/latest/)
- Scikit-learn *(http://scikit-learn.org/stable/)*

Scikit-learn

- Load datasets
- Preprocessing
- Klasifikasi
- Clustering
- Document Classification

Scikit-learn

- Windows user? Saya merekomendasikan pakai WinPython (sudah include NumPy, SciPy, Matplotlib, dll)
- Install pydotplus (pip2.7 install pydotplus)
- Install python-graphviz untuk membuat tree (pip2.7 install graphviz)
- Source code dan dataset sudah saya upload di github https://github.com/auliawiguna/sabtuceria

macOS or even Linux user?

Selamat bersenang-senang install manual T.T

Scikit-learn

- Load datasets
- Preprocessing
- Klasifikasi
- Validasi (pemisahan data training dan data testing)
- Clustering
- Document Classification

Load Datasets

- numPy dasar
- Load dataset bawaan
- Load datasetmu sebagai numPy array
- Pisahkan mana atribut dan mana label

Load Datasets

- *IRIS (klasifikasi)*
- *DIGITS (klasifikasi)*
- *WINE (klasifikasi)*
- *BREAST CANCER (klasifikasi)*
- *BOSTON (regresi)*
- *DIABETES (regresi)*
- *LINNERUD (regresi)*

Load Your Own Datasets

- Siapkan file teks (terserah mau csv, txt, atau apapun)
- Separator antar kolom usahakan **titik koma**, atau **koma**
- Boleh menggunakan private datasets
- Boleh juga download di public repository

Scikit-learn

- Load datasets
- Preprocessing
- Klasifikasi
- Validasi (pemisahan data training dan data testing)
- Clustering
- Document Classification

Preprocessing

- Bagaimana jika ada data numerik di dataset yang mau diklasifikasi?
- Bagaimana jika masih ada data "teks" di datasets padahal mau diklasifikasikan menggunakan algoritma yang menggunakan perhitungan jarak (eg KNN) ?
- Bagaimana jika ada missing value?

Preprocessing

- Bagaimana jika ada data numerik di dataset yang mau diklasifikasi?

Tinggi Badan	Berat Badan	Potensi Jantungan
176	56	1
156	78	1
145	55	0
177	89	0
188	96	0
143	45	1
150	58	1

What if… tinggi badan 144 dan berat badan 42?

Preprocessing

- Bagaimana jika masih ada data "teks" di datasets padahal mau diklasifikasikan menggunakan algoritma yang menggunakan perhitungan jarak (eg KNN) ?

Tinggi Badan	Berat Badan	Potensi Jantungan
tinggi	kurus	1
sedang	gapura kecamatan	1
agak sedang	kurus	0
tinggi	gemuk	0
tinggi	obesitas	0
rada-rada	kurang gizi	1
pendek	Kurang gizi	1

Preprocessing

- Bagaimana jika ada missing value?

Tinggi Badan	Berat Badan	Potensi Jantungan
tinggi	?	1
?	gapura kecamatan	1
agak sedang	?	0
tinggi	gemuk	0
?	obesitas	0
rada-rada	kurang gizi	1
pendek	Kurang gizi	1

Scikit-learn

- Load datasets
- Preprocessing
- Klasifikasi
- Validasi (pemisahan data training dan data testing)
- Clustering
- Document Classification

Klasifikasi

- Gaussian Naïve Bayes
- Multinomial Naïve Bayes
- Bernoulli Naïve Bayes
- Decision Tree
- Nearest Neighbors

Scikit-learn

- Load datasets
- Preprocessing
- Klasifikasi
- Validasi (pemisahan data training dan data testing)
- Clustering
- Document Classification

Validasi

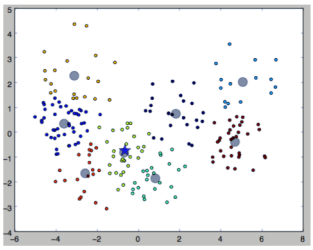
- Pemisahan berdasar rasio (train_test_split)
- K-Fold Cross Validation (cross_val_score & KFold)
- Akurasi Klasifikasi

Scikit-learn

- Load datasets
- Preprocessing
- Klasifikasi
- Validasi (pemisahan data training dan data testing)
- Clustering
- Document Classification

Clustering

- Dataset apa yang bisa dipakai
- Simulasi clustering menggunakan K-Means
- Visualisasi cluster



Scikit-learn

- Load datasets
- Preprocessing
- Klasifikasi
- Validasi (pemisahan data training dan data testing)
- Clustering
- Document Classification

Document Classification

- Menggunakan Naïve Bayes Multinomial
- Makin banyak file contoh, makin jos klasifikasi yang dihasilkan