

Dr. Wahyudi,S.T.,M.T.



# PENGANTAR SAINS DATA

Proses Sains Data dan Pembelajaran Mesin

# PENGANTAR SAINS DATA

## Proses Sains Data dan Pembelajaran Mesin

Buku Pengantar Sains Data: Proses Sains Data Dan Pembelajaran Mesin terdiri dari tiga bab. Bab pertama yaitu Sains Data Dalam Dunia Big Data menjelaskan tentang definisi sains data dan big data, proses sains data, dan ekosistem sains data. Bab kedua menjelaskan secara rinci proses sains data. Terdapat 6 langkah proses yang harus dilakukan di dalam sains data yaitu: Menetapkan tujuan penelitian, Mengambil data, Persiapan data, Eksplorasi data, Pemodelan data, Presentasi dan otomatisasi. Bab terakhir menjelaskan tentang pembelajaran mesin. Ilmuwan data sangat bergantung pada teknik-teknik dari statistik dan pembelajaran mesin untuk melakukan pemodelan. Ada 4 tahap proses pemodelan yaitu: Rekayasa fitur, Pelatihan model, Pemilihan dan validasi model, Penilaian model. Bab ini juga menjelaskan 3 teknik pembelajaran mesin yaitu: Pembelajaran Terawasi (supervised Learning), Pembelajaran tanpa pengawasan (Unsupervised learning), dan Pembelajaran semi terawasi (semi-supervised Learning) berada di antara kedua teknik tersebut dan digunakan ketika hanya sebahagian kecil data yang diberi label.



☎ 0858 5343 1992  
✉ eurekamediaaksara@gmail.com  
📍 Jl. Banjaran RT.20 RW.10  
Bojongsari - Purbalingga 53362

ISBN 978-623-151-503-2



**PENGANTAR SAINS DATA:**  
**Proses Sains Data dan Pembelajaran Mesin**

Dr. Wahyudi,S.T.,M.T.



**PENERBIT CV.EUREKA MEDIA AKSARA**

**PENGANTAR SAINS DATA:  
Proses Sains Data dan Pembelajaran Mesin**

**Penulis** : Dr. Wahyudi,S.T.,M.T.

**Desain Sampul** : Ardyan Arya Hayuwaskita

**Tata Letak** : Herlina Sukma

**ISBN** : 978-623-151-503-2

Diterbitkan oleh : **EUREKA MEDIA AKSARA, SEPTEMBER 2023**  
**ANGGOTA IKAPI JAWA TENGAH**  
**NO. 225/JTE/2021**

**Redaksi:**

Jalan Banjaran, Desa Banjaran RT 20 RW 10 Kecamatan Bojongsari  
Kabupaten Purbalingga Telp. 0858-5343-1992

Surel : eurekaediaaksara@gmail.com

Cetakan Pertama : 2023

**All right reserved**

Hak Cipta dilindungi undang-undang  
Dilarang memperbanyak atau memindahkan sebagian atau seluruh  
isi buku ini dalam bentuk apapun dan dengan cara apapun,  
termasuk memfotokopi, merekam, atau dengan teknik perekaman  
lainnya tanpa seizin tertulis dari penerbit.

## KATA PENGANTAR

Alhamdulillahirobbil'alamin. Penulis bersyukur kehadiran Allah SWT berkat rahmat, karunia dan pertolonganNya, penulis dapat menyelesaikan buku berjudul " **Pengantar Sains Data: proses sains data dan pembelajaran mesin**". Shalawat serta salam semoga senantiasa tercurah atas Nabi Muhammad SAW, para kerabat, serta pengikutnya hingga hari kiamat nanti.

Buku ini hadir untuk menambah literasi tentang teknologi informasi di lingkungan Universitas Andalas. Buku ini merupakan seri pertama dari beberapa buku sains data. Buku ini menjelaskan tentang munculnya bidang sains data, hubungan sains data dan big data, proses yang dilakukan di sains data dan pembelajaran mesin yang merupakan bagian paling penting di sains data.

Penulis menyadari bahwa dalam penulisan buku ini masih banyak terdapat kekurangan, untuk itu penulis mengharapkan kritik dan sarannya guna penyempurnaan buku ini di masa mendatang

Padang, Agustus 2023

Penulis

## DAFTAR ISI

<b>KATA PENGANTAR.....</b>	<b>iii</b>
<b>DAFTAR ISI.....</b>	<b>iv</b>
<b>DAFTAR GAMBAR.....</b>	<b>vi</b>
<b>DAFTAR TABEL .....</b>	<b>x</b>
<b>BAB 1 SAINS DATA DALAM DUNIA BIG DATA.....</b>	<b>1</b>
A. Manfaat dan Penggunaan Sains Data dan Data Besar ...	3
B. Aspek-Aspek Data.....	6
C. Proses Sains Data.....	13
D. Ekosistem Big Data dan Sains Data .....	17
<b>BAB 2 PROSES SAINS DATA.....</b>	<b>27</b>
A. Gambaran Umum Proses Sains Data.....	27
B. Langkah 1: Menentukan Tujuan Penelitian dan Membuat Perjanjian Proyek .....	33
1. Luangkan waktu untuk memahami tujuan dan konteks penelitian Anda .....	34
2. Buat perjanjian proyek .....	35
C. Langkah 2: Mengambil Data .....	36
1. Mulailah dengan data yang tersimpan di dalam perusahaan .....	37
2. Jangan takut untuk berbelanja.....	39
3. Pemeriksaan Kualitas Data Untuk Mencegah Masalah.....	40
D. Langkah 3: Membersihkan, Mengintegrasikan, Dan Mengubah Data .....	41
E. Langkah 4: Analisis dan eksplorasi Data .....	66
F. Langkah 5: Membangun Model.....	72
G. Langkah 6: Menyajikan temuan dan membangun aplikasi di atasnya.....	84
<b>BAB 3 PEMBELAJARAN MESIN.....</b>	<b>86</b>
A. Apa Itu Pembelajaran Mesin dan Mengapa Anda Harus Tahu? .....	87
B. Proses Pemodelan.....	95
C. Jenis-Jenis Pembelajaran Mesin.....	103
D. Pembelajaran Semi-Pengawasan.....	133
E. Rangkuman.....	136

**DAFTAR PUSTAKA ..... 139**

## DAFTAR GAMBAR

Gambar 1	Tabel Excel adalah contoh data terstruktur.....	7
Gambar 2	Email merupakan contoh data tidak terstruktur dan data bahasa alami .....	8
Gambar 3	Contoh data yang dihasilkan mesin.....	10
Gambar 4	Teman di jejaring sosial adalah contoh data berbasis graf .....	12
Gambar 5	Proses Sains data.....	14
Gambar 6	Teknologi big data dapat diklasifikasikan ke dalam beberapa komponen utama .....	19
Gambar 7	Enam langkah proses sains data.....	29
Gambar 8	Langkah 1: Menetapkan tujuan penelitian.....	34
Gambar 9	Langkah 2: Mengambil data .....	37
Gambar 10	Langkah 3: Persiapan data .....	43
Gambar 11	Titik yang dilingkari sangat mempengaruhi model dan perlu diselidiki karena titik tersebut dapat menunjukkan wilayah di mana Anda tidak memiliki data yang cukup atau mungkin mengindikasikan adanya kesalahan pada data, tetapi titik tersebut juga dapat menjadi titik data yang valid .....	45
Gambar 12	Plot distribusi sangat membantu dalam mendeteksi pencilan dan membantu Anda memahami variabel	49
Gambar 13	Menggabungkan dua tabel pada kunci Item dan Wilayah.....	57
Gambar 14	Menambahkan data dari tabel adalah operasi yang umum dilakukan, namun membutuhkan struktur yang sama pada tabel yang ditambahkan.....	59
Gambar 15	View membantu Anda menggabungkan data tanpa replikasi .....	60
Gambar 16	Pertumbuhan, penjualan berdasarkan kelas produk, dan peringkat penjualan adalah contoh ukuran turunan dan agregat.....	61
Gambar 17	Mentransformasikan $x$ menjadi $\log x$ membuat hubungan antara $x$ dan $y$ menjadi linier (kanan), dibandingkan dengan non- $\log x$ (kiri).....	62



Gambar 18	Variabel Pengurangan memungkinkan Anda untuk mengurangi jumlah variabel, dan tetap mempertahankan banyak informasi sebanyak mungkin.....	64
Gambar 19	Mengubah variabel menjadi dummy adalah transformasi data yang memecah sebuah variabel yang memiliki banyak kelas menjadi beberapa variabel, yang masing-masing hanya memiliki dua kemungkinan nilai: 0 atau 1.....	65
Gambar 20	Langkah 4: Eksplorasi data.....	66
Gambar 21	Dari atas ke bawah, diagram batang, plot garis, dan distribusi adalah beberapa grafik yang digunakan dalam analisis eksplorasi.....	68
Gambar 22	Menggambar beberapa plot secara bersamaan dapat membantu Anda memahami struktur data Anda atas beberapa variabel.....	69
Gambar 23	Diagram Pareto.....	70
Gambar 24	Tautan dan kuas memungkinkan Anda untuk memilih pengamatan pada satu plot dan menyorot pengamatan yang sama pada plot lainnya.....	70
Gambar 25	Contoh histogram: jumlah orang dalam kelompok umur interval 5 tahun.....	71
Gambar 26	Contoh boxplot: setiap kategori pengguna memiliki distribusi apresiasi yang dimiliki masing-masing untuk gambar tertentu pada situs web fotografi. Teknik-teknik yang kami jelaskan pada tahap ini sebagian besar bersifat visual, namun dalam praktiknya teknik-teknik tersebut tidak terbatas pada teknik visualisasi. Tabulasi, pengelompokan, dan teknik pemodelan lainnya juga dapat menjadi bagian dari analisis eksplorasi. Bahkan membangun model sederhana dapat menjadi bagian dari langkah ini. ..	71
Gambar 27	Langkah 5: Pemodelan data.....	73
Gambar 28	Regresi linier mencoba untuk menyesuaikan sebuah garis sambil meminimalkan jarak ke setiap titik.....	75

Gambar 29	Model Output Informasi Regresi linier.....	76
Gambar 30	Teknik k-nearest neighbor melihat titik k-terdekat untuk membuat prediksi.....	79
Gambar 31	Confusion matrix: menunjukkan berapa banyak kasus yang diklasifikasikan dengan benar dan salah dengan membandingkan prediksi dengan nilai sebenarnya. Catatan: kelas (0,1,2) ditambahkan pada gambar untuk memperjelas. ....	80
Gambar 32	Rumus untuk mean square error .....	82
Gambar 33	Sampel penanggulangan membantu Anda membandingkan model dan memastikan bahwa Anda dapat menggeneralisasi hasil pada data yang belum dilihat oleh model. ....	83
Gambar 34	Langkah 6: Presentasi dan otomatisasi .....	84
Gambar 35	Proses sains data .....	90
Gambar 36	Gambaran umum paket Python yang digunakan selama fase pembelajaran mesin .....	92
Gambar 37	Kontrol Captcha sederhana dapat digunakan untuk mencegah spam otomatis dikirim melalui formulir web online .....	105
Gambar 38	Skala abu-abu buram representasi dari angka 0 dengan matriks yang sesuai. Semakin tinggi angkanya, semakin dekat angkanya ke warna putih; semakin rendah angkanya, semakin dekat dengan warna hitam .....	108
Gambar 39	Kita akan mengubah sebuah gambar menjadi sesuatu yang dapat digunakan oleh pengklasifikasi Naïve Bayes dengan mendapatkan nilai grayscale untuk setiap pikselnya (ditunjukkan di sebelah kanan) dan memasukkan nilai tersebut ke dalam daftar.....	109
Gambar 40	Matriks confusion yang dihasilkan dengan memprediksi nomor berapa yang digambarkan oleh gambar buram.....	112
Gambar 41	Untuk setiap gambar buram, sebuah angka diprediksi; hanya angka 2 yang disalahartikan	

	sebagai 8. Kemudian angka yang tidak jelas diprediksi sebagai 3, namun bisa juga sebagai 5; bahkan bagi mata manusia, hal ini tidak jelas.....	114
Gambar 42	Plot scree PCA yang menunjukkan jumlah marginal informasi dari setiap variabel baru yang dapat dihasilkan oleh PCA. Variabel pertama menjelaskan sekitar 28% dari varians dalam data, variabel kedua menyumbang 17% lainnya, variabel ketiga sekitar 15%, dan seterusnya.....	122
Gambar 43	Plot hasil menunjukkan bahwa menambahkan lebih banyak variabel laten ke dalam model (sumbu x) sangat meningkatkan daya prediksi (sumbu y) .....	128
Gambar 44	Tujuan dari pengelompokan adalah untuk membagi kumpulan data menjadi himpunan bagian yang "cukup himpunan bagian yang "cukup berbeda". Dalam plot ini misalnya misalnya, pengamatan telah dibagi menjadi tiga cluster.....	131
Gambar 45	Output dari klasifikasi iris.....	133
Gambar 46	Plot ini hanya memiliki dua pengamatan berlabel-terlalu sedikit untuk pengamatan yang diawasi, tetapi cukup untuk memulai dengan pendekatan yang tidak diawasi atau semi-diawasi .....	134
Gambar 47	Cara yang lebih baik membagi data dalam bentuk clustering.....	135

## DAFTAR TABEL

Tabel 1	Daftar penyedia data terbuka yang dapat membantu Anda .....	39
Tabel 2	Overview kesalahan umum .....	43
Tabel 3	Mendeteksi pencilan pada variabel sederhana dengan tabel frekuensi.....	47
Tabel 4	Gambaran teknik untuk menangani data yang hilang.....	50
Tabel 5	Tiga baris pertama dari Kumpulan Data Kualitas Anggur Merah.....	120
Tabel 6	Temuan-temuan PCA .....	123
Tabel 7	Bagaimana PCA menghitung korelasi 11 variabel asli dengan 5 variabel laten .....	125
Tabel 8	Interpretasi dari variabel-variabel yang dibuat oleh PCA untuk kualitas wine .....	126
Tabel 9	Tiga baris pertama dari Kumpulan Data Kualitas Anggur Merah yang dikodekan ulang dalam lima variabel laten .....	126

# BAB

# 1

## SAINS DATA DALAM DUNIA BIG DATA

Bab ini mencakup

1. Mendefinisikan Sains data dan big data
2. Mengenali berbagai jenis data
3. Memperoleh wawasan tentang proses sains data
4. Memperkenalkan bidang-bidang sains data dan big data

Big data adalah istilah umum untuk kumpulan data yang begitu besar atau kompleks sehingga sulit untuk diproses menggunakan teknik manajemen data tradisional seperti, misalnya, RDBMS (sistem manajemen basis data relasional). RDBMS yang diadopsi secara luas telah lama dianggap sebagai solusi yang cocok untuk semua, namun tuntutan untuk menangani data besar menunjukkan sebaliknya. Sains data melibatkan penggunaan metode untuk menganalisis data dalam jumlah besar dan mengekstrak pengetahuan yang dikandungnya. Anda dapat menganggap hubungan antara big data dan sains data seperti hubungan antara minyak mentah dan kilang minyak. sains data dan big data berevolusi dari statistik dan manajemen data tradisional, namun kini dianggap sebagai disiplin ilmu yang berbeda.

# BAB

# 2

## PROSES SAINS DATA

Tujuan dari bab ini adalah untuk memberikan gambaran umum tentang proses sains data tanpa harus menyelami big data. Anda akan mempelajari cara bekerja dengan kumpulan data besar, data streaming, dan data teks di bab-bab berikutnya.

### A. Gambaran Umum Proses Sains Data

Mengikuti pendekatan terstruktur untuk sains data membantu Anda memaksimalkan peluang keberhasilan dalam proyek sains data dengan biaya terendah. Pendekatan ini juga memungkinkan untuk mengerjakan proyek sebagai sebuah tim, dengan setiap anggota tim berfokus pada apa yang mereka lakukan dengan baik. Namun, berhati-hatilah: pendekatan ini mungkin tidak cocok untuk semua jenis proyek atau menjadi satu-satunya cara untuk melakukan sains data yang baik.

Proses sains data pada umumnya terdiri dari enam langkah yang akan Anda lakukan secara berulang, seperti yang ditunjukkan pada gambar 7.

Gambar 7 meringkas proses sains data dan menunjukkan langkah-langkah dan tindakan utama yang

# BAB 3

## PEMBELAJARAN MESIN

Bab ini mencakup

1. Memahami mengapa ilmuwan sains data menggunakan pembelajaran mesin
2. Mengidentifikasi pustaka Python yang paling penting untuk pembelajaran mesin
3. Mendiskusikan proses pembuatan model
4. Menggunakan teknik-teknik pembelajaran mesin
5. Mendapatkan pengalaman langsung dengan pembelajaran mesin

Tahukah Anda bagaimana komputer belajar untuk melindungi Anda dari orang yang berniat jahat? Komputer menyaring lebih dari 60% email Anda dan dapat belajar untuk melakukan pekerjaan yang lebih baik dalam melindungi Anda dari waktu ke waktu.

Dapatkah Anda secara eksplisit mengajari komputer untuk mengenali orang dalam sebuah gambar? Mungkin saja, tetapi tidak praktis untuk mengkodekan semua cara yang mungkin untuk mengenali seseorang, tetapi Anda akan segera melihat bahwa kemungkinannya hampir tak terbatas. Agar berhasil, Anda harus menambahkan keterampilan baru

## DAFTAR PUSTAKA

- Apache Spark. (2023). Apache Spark for Python. <https://spark.apache.org/>
- Cielen , Meysman, Ali. (2016). Introducing Data Science. Manning Publications Co. Shelter Island, New York
- Cielen. (2016). Overview of the data science process. <http://freecontent.manning.com/overview-of-the-data-science-process/>
- Cyton. (2023). C-Extensions For Python. <https://cython.org/>
- Mike Bostock. (2017). diagram Sankey. <https://bost.ocks.org/mike/>
- Pydata. (2023). NumExpr: Fast numerical expression evaluator for NumPy. <https://github.com/pydata/numexpr>
- Wikipedia. (2023). Time complexity. [https://en.wikipedia.org/wiki/Time\\_complexity](https://en.wikipedia.org/wiki/Time_complexity)