DASAR-DASAR DATA SCIENCE

Panduan Lengkap untuk Pemula

Memahami konsep fundamental, tools, dan praktik terbaik dalam dunia Data Science

B OneCorp

2025 Edition

Daftar Isi

Bab 1: Pengenalan Data Science

1.1 Apa itu Data Science?

Data Science adalah bidang interdisipliner yang menggunakan metode ilmiah, proses, algoritma, dan sistem untuk mengekstrak pengetahuan dan wawasan dari data terstruktur maupun tidak terstruktur. Data Science menggabungkan berbagai elemen dari statistika, matematika, pemrograman komputer, dan domain expertise.

> **Definisi Kunci:** Data Science adalah seni dan ilmu untuk mengubah data mentah menjadi insight yang dapat ditindaklanjuti untuk membuat keputusan bisnis yang lebih baik.

1.2 Komponen Utama Data Science

Data Science terdiri dari beberapa komponen utama yang saling berkaitan:

- **Statistika dan Matematika:** Fondasi untuk memahami dan menganalisis data
- **Pemrograman:** Kemampuan untuk mengimplementasikan analisis dan model
- **Domain Knowledge:** Pemahaman tentang konteks bisnis atau industri
- **Data Visualization:** Kemampuan menyajikan hasil analisis secara visual
- **Machine Learning:** Teknik untuk membuat prediksi dan pattern recognition

1.3 Peran Data Scientist

Seorang Data Scientist memiliki tanggung jawab yang beragam dalam organisasi:

1. Mengumpulkan dan membersihkan data dari berbagai sumber

2. Melakukan analisis eksploratif untuk menemukan pola dan trend
3. Membangun model prediktif dan machine learning
4. Mengkomunikasikan hasil temuan kepada stakeholder
5. Mengimplementasikan solusi data-driven

1.4 Lifecycle Data Science Project

Setiap proyek Data Science mengikuti siklus yang sistematis:

Tahap	Deskripsi	Output
Problem Definition	Mendefinisikan masalah bisnis	Problem statement yang jelas
Data Collection	Mengumpulkan data relevan	Dataset mentah
Data Cleaning	Membersihkan dan mempersiapkan data	Data yang bersih
Exploratory Analysis	Mengeksplorasi dan memahami data	Insight awal
Modeling	Membangun model prediktif	Model yang terlatih
Evaluation	Mengevaluasi performa model	Metrics dan validasi
Deployment	Implementasi ke production	Sistem yang berjalan

1.5 Aplikasi Data Science di Industri

Data Science memiliki aplikasi yang luas di berbagai industri:

- **E-commerce:** Sistem rekomendasi produk, personalisasi pengalaman pengguna
- **Healthcare:** Prediksi penyakit, analisis genomik, drug discovery
- **Finance:** Deteksi fraud, credit scoring, algorithmic trading
- **Marketing:** Customer segmentation, churn prediction, campaign optimization
- **Transportation:** Route optimization, demand forecasting

Bab 2: Matematika untuk Data Science

2.1 Aljabar Linear

Aljabar linear adalah fondasi matematika dalam Data Science, terutama untuk machine learning dan deep learning.

Konsep Kunci:

- **Vektor:** Array satu dimensi yang merepresentasikan magnitude dan direction
- **Matriks:** Array dua dimensi yang digunakan untuk transformasi data
- **Operasi Matriks:** Penjumlahan, perkalian, transpose, inverse
- **Eigenvalues dan Eigenvectors:** Penting untuk PCA dan dimensionality reduction

> **Aplikasi Praktis:** Aljabar linear digunakan dalam neural networks, image processing, recommendation systems, dan natural language processing.

2.2 Kalkulus

Kalkulus membantu kita memahami bagaimana fungsi berubah dan digunakan dalam optimasi model.

Topik Penting:

- **Derivatif:** Mengukur rate of change suatu fungsi

- **Gradient:** Vektor derivatif parsial, digunakan dalam gradient descent
- **Chain Rule:** Fundamental dalam backpropagation neural networks
- **Optimisasi:** Mencari minimum atau maximum suatu fungsi

2.3 Probabilitas

Probabilitas adalah bahasa ketidakpastian dalam Data Science.

Konsep Dasar:

- **Probability Distribution:** Normal, Binomial, Poisson
- **Conditional Probability:** P(A|B) - probabilitas A given B
- **Bayes Theorem:** Dasar untuk Naive Bayes classifier
- **Expected Value:** Nilai rata-rata yang diharapkan

2.4 Logika dan Set Theory

Fundamental untuk memahami struktur data dan query.

- Operations: Union, Intersection, Complement
- Boolean Logic: AND, OR, NOT
- Venn Diagrams untuk visualisasi

Bab 3: Python untuk Data Science

3.1 Mengapa Python?

Python adalah bahasa pemrograman paling populer untuk Data Science karena:

- Sintaks yang mudah dipahami dan dipelajari
- Ekosistem library yang kaya untuk Data Science
- Komunitas yang besar dan aktif
- Integrasi yang baik dengan tools lain
- Open source dan gratis

3.2 Library Essential untuk Data Science

NumPy - Numerical Python

Library fundamental untuk komputasi numerik.

```
import numpy as np # Membuat array arr = np.array([1, 2, 3, 4, 5]) # Operasi matematika mean = np.mean(arr) std = np.std(arr) # Array multidimensi matrix = np.array([[1, 2], [3, 4]])
```

Pandas - Data Manipulation

Library untuk manipulasi dan analisis data terstruktur.

```
import pandas as pd # Membuat DataFrame df = pd.DataFrame({ 'nama': ['Alice', 'Bob', 'Charlie'], 'usia': [25, 30, 35],
```

```
'gaji': [50000, 60000, 70000] }) # Operasi dasar
print(df.head()) print(df.describe())
print(df.groupby('usia').mean())
```

Matplotlib & Seaborn - Visualisasi

Library untuk membuat visualisasi data.

```
import matplotlib.pyplot as plt import seaborn as sns # Plot
sederhana plt.plot([1, 2, 3, 4], [1, 4, 9, 16])
plt.xlabel('X') plt.ylabel('Y') plt.title('Plot Sederhana')
plt.show() # Seaborn untuk visualisasi statistik
sns.boxplot(data=df, x='kategori', y='nilai')
```

Scikit-learn - Machine Learning

Library utama untuk machine learning di Python.

```
from sklearn.model_selection import train_test_split from
sklearn.linear_model import LinearRegression # Split data
X_train, X_test, y_train, y_test = train_test_split( X, y,
test_size=0.2, random_state=42 ) # Train model model =
LinearRegression() model.fit(X_train, y_train) # Prediksi
predictions = model.predict(X_test)
```

3.3 Best Practices dalam Coding

- Gunakan nama variabel yang deskriptif
- Tulis komentar untuk kode yang kompleks
- Ikuti PEP 8 style guide
- Buat fungsi untuk kode yang reusable

- Handle errors dengan try-except
- Gunakan virtual environment untuk project

Bab 4: Statistika Dasar

4.1 Descriptive Statistics

Statistika deskriptif membantu merangkum dan mendeskripsikan karakteristik data.

Measures of Central Tendency:

- **Mean (Rata-rata):** Jumlah semua nilai dibagi jumlah observasi
- **Median:** Nilai tengah ketika data diurutkan
- **Mode:** Nilai yang paling sering muncul

Measures of Dispersion:

- **Range:** Selisih antara nilai maksimum dan minimum
- **Variance:** Rata-rata dari kuadrat deviasi dari mean
- **Standard Deviation:** Akar kuadrat dari variance
- **Interquartile Range (IQR):** Range dari Q1 ke Q3

4.2 Probability Distributions

Normal Distribution

Distribusi paling penting dalam statistika, berbentuk bell curve.

> **Karakteristik:**
>
> - Simetris terhadap mean
> - Mean = Median = Mode

> - 68% data dalam 1 standard deviation
> - 95% data dalam 2 standard deviations

Other Distributions:

- **Binomial:** Untuk discrete events dengan dua outcomes
- **Poisson:** Untuk menghitung events dalam interval waktu
- **Exponential:** Untuk waktu antar events

4.3 Hypothesis Testing

Metode untuk membuat keputusan berdasarkan data sample.

Langkah-langkah:

1. **Null Hypothesis (H0):** Asumsi default (tidak ada effect)
2. **Alternative Hypothesis (H1):** Yang ingin kita buktikan
3. **Significance Level (α):** Biasanya 0.05 atau 5%
4. **Calculate Test Statistic:** t-test, z-test, chi-square
5. **P-value:** Probabilitas mendapat hasil ekstrem
6. **Conclusion:** Reject atau fail to reject H0

4.4 Correlation dan Causation

Penting untuk membedakan hubungan vs sebab-akibat.

> **Ingat:** Correlation does not imply causation. Dua variabel bisa berkorelasi tanpa ada hubungan kausal.

Correlation Coefficient (r):

- Range: -1 hingga +1
- +1: Perfect positive correlation
- 0: No correlation
- -1: Perfect negative correlation

Bab 5: Eksplorasi dan Visualisasi Data

5.1 Exploratory Data Analysis (EDA)

EDA adalah proses investigasi awal untuk memahami karakteristik data, menemukan pola, anomali, dan menguji asumsi.

Langkah-langkah EDA:

1. **Load dan Inspect Data**
    - Lihat dimensi data (shape)
    - Cek tipe data setiap kolom
    - Identifikasi missing values
2. **Statistical Summary**
    - Calculate descriptive statistics
    - Identify outliers
    - Check distributions
3. **Visualize Relationships**
    - Correlation heatmaps
    - Scatter plots untuk relationships
    - Distribution plots

5.2 Jenis-jenis Visualisasi

Univariate Analysis:

- **Histogram:** Distribusi variabel numerik
- **Box Plot:** Menunjukkan median, quartiles, dan outliers
- **Bar Chart:** Untuk data kategorikal
- **Density Plot:** Smooth version dari histogram

Bivariate Analysis:

- **Scatter Plot:** Hubungan antara dua variabel numerik
- **Line Plot:** Trends over time
- **Heatmap:** Correlation matrix
- **Box Plot by Category:** Perbandingan distribusi

Multivariate Analysis:

- **Pair Plot:** Multiple scatter plots
- **3D Scatter Plot:** Tiga variabel sekaligus
- **Parallel Coordinates:** Multiple dimensions

5.3 Data Cleaning

Data cleaning adalah proses memperbaiki atau menghapus data yang tidak akurat, corrupt, atau tidak relevan.

Common Issues:

- **Missing Values:**
    - Drop rows/columns dengan missing values
    - Imputation (mean, median, mode)
    - Forward fill atau backward fill

    - Predictive imputation
- **Duplicates:** Identifikasi dan remove duplicate rows
- **Outliers:**
    - Detection: IQR method, Z-score
    - Treatment: Remove, cap, atau transform
- **Data Type Issues:** Convert ke tipe yang sesuai
- **Inconsistent Formatting:** Standardize format

5.4 Feature Engineering

Proses membuat features baru dari data existing untuk meningkatkan performa model.

Techniques:

- **Binning:** Convert continuous ke categorical
- **Encoding:** One-hot, label, target encoding
- **Scaling:** Normalization, standardization
- **Polynomial Features:** Create interaction terms
- **Date/Time Features:** Extract year, month, day, etc.
- **Text Features:** TF-IDF, word embeddings

> **Golden Rule:** Feature engineering sering memberikan improvement yang lebih besar daripada menggunakan algoritma yang lebih sophisticated.

Bab 6: Machine Learning Dasar

6.1 Apa itu Machine Learning?

Machine Learning adalah subset dari Artificial Intelligence yang memberikan sistem kemampuan untuk belajar dan improve dari experience tanpa explicitly programmed.

6.2 Tipe Machine Learning

Supervised Learning

Model belajar dari labeled data (input-output pairs).

- **Classification:** Prediksi kategori (spam/not spam, disease/healthy)
- **Regression:** Prediksi nilai kontinyu (harga rumah, temperature)

Unsupervised Learning

Model menemukan pola dari unlabeled data.

- **Clustering:** Grouping similar data points (customer segmentation)
- **Dimensionality Reduction:** Reduce features (PCA, t-SNE)
- **Association:** Find rules (market basket analysis)

Reinforcement Learning

Model belajar melalui trial and error dengan rewards.

- Game playing (AlphaGo)
- Robotics
- Autonomous vehicles

6.3 Algoritma Supervised Learning

Linear Regression

Algoritma paling sederhana untuk regression problems.

> **Formula:** $y = mx + b$
> **Use Case:** Sales forecasting, price prediction

Logistic Regression

Untuk binary classification problems.

> **Output:** Probability between 0 and 1
> **Use Case:** Email spam detection, disease diagnosis

Decision Trees

Model berbentuk tree untuk classification dan regression.

- Easy to interpret dan visualize
- Handle non-linear relationships
- Prone to overfitting

Random Forest

Ensemble dari multiple decision trees.

- Reduces overfitting
- More accurate than single tree
- Feature importance

Support Vector Machines (SVM)

Finds optimal hyperplane untuk classification.

- Effective dalam high-dimensional spaces
- Memory efficient
- Kernel trick untuk non-linear problems

K-Nearest Neighbors (KNN)

Classifies based on k nearest neighbors.

- Simple dan intuitive
- No training phase
- Slow prediction time

6.4 Model Evaluation

Regression Metrics:

- **Mean Absolute Error (MAE):** Average absolute difference
- **Mean Squared Error (MSE):** Average squared difference
- **R-squared (R^2):** Proportion of variance explained
- **Root Mean Squared Error (RMSE):** Square root of MSE

Classification Metrics:

- **Accuracy:** Percentage correct predictions
- **Precision:** True positives / (True positives + False positives)
- **Recall:** True positives / (True positives + False negatives)
- **F1-Score:** Harmonic mean of precision and recall
- **Confusion Matrix:** Visual representation of predictions

- **ROC-AUC:** Area under ROC curve

6.5 Overfitting dan Underfitting

Overfitting

Model terlalu complex dan "menghafal" training data.

- High accuracy pada training data
- Low accuracy pada test data
- **Solution:** Regularization, more data, cross-validation

Underfitting

Model terlalu simple dan tidak capture patterns.

- Low accuracy pada training dan test data
- **Solution:** More features, more complex model

> **Goal:** Find sweet spot dengan good generalization pada unseen data.

6.6 Cross-Validation

Technique untuk evaluate model performance lebih reliably.

K-Fold Cross-Validation:

1. Split data menjadi K folds
2. Train pada K-1 folds, test pada 1 fold
3. Repeat K times
4. Average results

6.7 Hyperparameter Tuning

Process mencari kombinasi hyperparameter terbaik.

Methods:

- **Grid Search:** Try all combinations
- **Random Search:** Random sampling dari parameter space
- **Bayesian Optimization:** Smart search based on previous results

Bab 7: Tools dan Teknologi

7.1 Development Environment

Jupyter Notebook

Interactive environment untuk data analysis dan prototyping.

- Combine code, visualization, dan markdown
- Easy to share dan collaborate
- Support untuk multiple languages
- Great untuk exploratory analysis

IDE Options:

- **VS Code:** Lightweight, extensible, popular
- **PyCharm:** Powerful IDE khusus Python
- **Spyder:** Scientific Python development
- **Google Colab:** Free cloud-based Jupyter

7.2 Version Control

Git dan GitHub

Essential untuk collaboration dan tracking changes.

```
# Basic Git commands git init # Initialize repository git add
. # Stage changes git commit -m "message" # Commit changes
git push # Push to remote git pull # Pull from remote git
```

```
branch # List branches git checkout -b new # Create new
branch
```

7.3 Big Data Technologies

Apache Spark

Distributed computing framework untuk big data processing.

- Process large datasets in parallel
- In-memory computation
- Support untuk ML dengan MLlib
- Real-time streaming

Hadoop Ecosystem

- **HDFS:** Distributed file system
- **MapReduce:** Programming model
- **Hive:** SQL-like queries
- **Pig:** Data flow language

7.4 Cloud Platforms

AWS (Amazon Web Services)

- **S3:** Object storage
- **EC2:** Virtual servers
- **SageMaker:** ML platform
- **Redshift:** Data warehouse

Google Cloud Platform

- **BigQuery:** Data warehouse
- **Cloud Storage:** Object storage
- **Vertex AI:** ML platform
- **Dataflow:** Stream/batch processing

Microsoft Azure

- **Azure ML:** ML platform
- **Blob Storage:** Object storage
- **Synapse Analytics:** Data warehouse
- **Databricks:** Spark platform

7.5 Databases

SQL Databases

- **PostgreSQL:** Advanced open-source RDBMS
- **MySQL:** Popular open-source database
- **SQLite:** Lightweight embedded database

NoSQL Databases

- **MongoDB:** Document database
- **Cassandra:** Wide-column store
- **Redis:** In-memory key-value store
- **Neo4j:** Graph database

7.6 Data Pipeline Tools

Apache Airflow

Platform untuk orchestrate complex workflows.

- Define workflows as code (DAGs)
- Schedule dan monitor pipelines
- Rich UI untuk tracking
- Extensible dengan custom operators

ETL Tools:

- **Apache NiFi:** Data integration
- **Talend:** Data integration platform
- **dbt:** Transform data in warehouse

7.7 Deployment dan MLOps

Model Deployment:

- **Flask/FastAPI:** Create REST APIs
- **Docker:** Containerize applications
- **Kubernetes:** Orchestrate containers
- **MLflow:** ML lifecycle management

Monitoring:

- Track model performance
- Detect data drift
- A/B testing
- Logging dan alerting

Bab 8: Proyek Data Science

8.1 Best Practices

Project Structure

```
project/ │ ├── data/ │ ├── raw/ # Original data │ ├── processed/ # Cleaned data │ └── external/ # External datasets │ ├── notebooks/ # Jupyter notebooks │ ├── 01_exploration.ipynb │ ├── 02_modeling.ipynb │ └── 03_evaluation.ipynb │ ├── src/ # Source code │ ├── data/ # Data processing │ ├── features/ # Feature engineering │ ├── models/ # Model training │ └── visualization/ # Plotting functions │ ├── tests/ # Unit tests ├── models/ # Saved models ├── reports/ # Analysis reports ├── requirements.txt # Dependencies └── README.md # Documentation
```

8.2 Documentation

Good documentation adalah kunci untuk collaboration dan maintenance.

What to Document:

- **README:** Project overview, setup instructions
- **Code Comments:** Explain complex logic
- **Docstrings:** Function/class descriptions
- **Data Dictionary:** Describe all features
- **Model Card:** Model specifications, limitations

- **Analysis Report:** Findings dan recommendations

8.3 Ethical Considerations

Bias in Data

- Historical bias in training data
- Selection bias dalam sampling
- Measurement bias dalam data collection
- **Mitigation:** Diverse data, fairness metrics, bias testing

Privacy dan Security

- Protect personal identifiable information (PII)
- Comply dengan regulations (GDPR, CCPA)
- Anonymize sensitive data
- Secure data storage dan transmission

Transparency dan Explainability

- Make models interpretable
- Document model decisions
- Communicate limitations
- Enable audit trails

> **Remember:** Dengan great power comes great responsibility. Data Scientists harus consider ethical implications dari their work.

8.4 Communication Skills

Storytelling dengan Data

Transform technical findings menjadi compelling narratives.

- Know your audience
- Start dengan business question
- Use clear visualizations
- Highlight key insights
- Provide actionable recommendations

Presentation Tips:

- Avoid technical jargon untuk non-technical audience
- Use analogies untuk explain complex concepts
- Focus pada business impact, bukan technical details
- Anticipate questions dan prepare answers
- Practice delivery

8.5 Continuous Learning

Staying Current

Data Science adalah fast-evolving field. Keep learning:

- **Read Papers:** ArXiv, Google Scholar, conference proceedings
- **Follow Blogs:** Towards Data Science, KDnuggets, Analytics Vidhya
- **Take Courses:** Coursera, edX, DataCamp, Fast.ai
- **Join Communities:** Kaggle, GitHub, Stack Overflow
- **Attend Conferences:** NeurIPS, ICML, KDD, local meetups
- **Practice:** Kaggle competitions, personal projects

8.6 Career Path

Roles dalam Data Science:

- **Data Analyst:** Focus pada reporting dan insights
- **Data Scientist:** Build predictive models
- **Machine Learning Engineer:** Deploy dan scale models
- **Data Engineer:** Build data infrastructure
- **Research Scientist:** Advance state-of-the-art
- **Product Data Scientist:** Product-focused analytics

Skills untuk Success:

Technical Skills	Soft Skills
Programming (Python, R, SQL)	Communication
Statistics dan Mathematics	Problem Solving
Machine Learning	Critical Thinking
Data Visualization	Business Acumen
Big Data Technologies	Collaboration
Cloud Platforms	Curiosity

8.7 Sample Project: Customer Churn Prediction

Problem Statement

Predict which customers are likely to churn untuk enable proactive retention.

Step-by-Step Approach:

1. **Business Understanding**
    - Define churn (tidak aktif selama 3 bulan)
    - Understand business impact
    - Set success criteria
2. **Data Collection**
    - Customer demographics
    - Transaction history
    - Customer service interactions
    - Product usage data
3. **Data Exploration**
    - Check data quality
    - Analyze churn rate
    - Identify patterns
    - Feature correlations
4. **Feature Engineering**
    - Recency, frequency, monetary value
    - Customer lifetime value
    - Engagement metrics
    - Seasonal features
5. **Model Building**
    - Train-test split
    - Try multiple algorithms

- Handle class imbalance
- Hyperparameter tuning

6. **Model Evaluation**
   - Focus on recall (catch churners)
   - ROC-AUC score
   - Feature importance
   - Business metrics (cost-benefit)
7. **Deployment**
   - Create API endpoint
   - Build dashboard
   - Setup monitoring
   - Document model
8. **Business Impact**
   - Prioritize high-risk customers
   - Personalized retention campaigns
   - Track retention rate improvement
   - Calculate ROI

8.8 Key Takeaways

> **10 Principles untuk Data Science Success:**
>
> 1. Always start dengan clear problem definition
> 2. Spend significant time pada data quality
> 3. Simple models sering outperform complex ones
> 4. Feature engineering > Algorithm selection

> 5. Always validate dengan out-of-sample data
> 6. Communicate findings clearly
> 7. Consider ethical implications
> 8. Document everything
> 9. Collaborate dengan domain experts
> 10. Never stop learning

Penutup

Selamat! Anda telah menyelesaikan panduan "Dasar-Dasar Data Science". Buku ini telah membahas fundamental concepts yang Anda butuhkan untuk memulai journey Anda dalam dunia Data Science.

Journey Anda Baru Dimulai

Data Science adalah field yang luas dan terus berkembang. Pengetahuan yang Anda peroleh dari buku ini adalah foundation yang solid, namun practical experience adalah kunci untuk mastery.

Next Steps

- **Practice:** Work on real projects dan Kaggle competitions
- **Build Portfolio:** Showcase your work di GitHub
- **Network:** Join communities dan attend events
- **Specialize:** Deep dive ke area yang Anda minati
- **Contribute:** Share knowledge melalui blog atau open source

Recommended Resources

- **Books:** "Python for Data Analysis" by Wes McKinney, "Hands-On Machine Learning" by Aurélien Géron
- **Courses:** Andrew Ng's Machine Learning, Fast.ai Practical Deep Learning
- **Platforms:** Kaggle, DataCamp, Coursera, edX
- **Communities:** r/datascience, Kaggle Forums, Stack Overflow

> **"Data is the new oil, but like oil, it needs to be refined to be valuable."**
>
> Keep learning, stay curious, and use your skills untuk create positive impact.

Happy Data Science Journey!

Dasar-Dasar Data Science

Untuk informasi lebih lanjut, kunjungi website kami atau hubungi tim kami.

"Empowering the next generation of data scientists"