TUGAS 1 MAKALAH ESTIMASI

KELOMPOK 4

MK. IFB-307 DATA MINING DAN INFORMATION RETRIEVAL

MODEL ESTIMASI DALAM DATA MINING MENGGUNAKAN STUDI KASUS USA MONTHLY RETAIL SALES

Disusun oleh :

Kelompok 4

15-2023-143	-	Maudy Amalia
15-2023-146	-	Rizki Saepul Aziz
15-2023-155	-	Pasha Muhammad Nashwan
15-2023-168	-	Ridayanti Wardani

Prodi Informatika

Fakultas Teknologi Industri

Institut Teknologi Nasional

Bandung

2025

1. Pendahuluan

Perkembangan teknologi informasi telah memungkinkan berbagai bidang untuk memanfaatkan data dalam jumlah besar guna mendukung pengambilan keputusan. Salah satu teknik penting dalam Data Mining dan Information Retrieval adalah model estimasi, yaitu model yang digunakan untuk memprediksi nilai atau pola berdasarkan data historis. Estimasi sangat dibutuhkan dalam dunia bisnis, ekonomi, kesehatan, dan berbagai sektor lainnya untuk membantu perencanaan masa depan.

Makalah ini akan membahas definisi dan sejarah model estimasi, formula matematis yang digunakan, studi kasus dengan data penjualan ritel bulanan di Amerika Serikat, serta implementasi menggunakan Microsoft Excel dan Python.

2. Definisi dan Sejarah Model Estimasi

2.1 Definisi

Model estimasi adalah metode statistik atau komputasi yang digunakan untuk memperkirakan nilai suatu variabel berdasarkan data sebelumnya. Dalam data mining, estimasi bertujuan untuk mengetahui kemungkinan nilai di masa depan, sehingga dapat membantu perencanaan dan pengambilan keputusan.

Secara umum, estimasi adalah suatu proses untuk memperkirakan nilai parameter atau variabel berdasarkan data yang tersedia. Dalam bidang statistika, estimasi digunakan untuk menghitung nilai parameter populasi (misalnya rata-rata atau variansi) dengan menggunakan data sampel. Sementara dalam data mining dan information retrieval, estimasi lebih difokuskan pada pembuatan model prediktif yang dapat memperkirakan nilai variabel di masa depan berdasarkan data historis.

Beberapa ciri utama dari model estimasi dalam data mining adalah:

a) Berbasis Data Historis → prediksi dilakukan dengan memanfaatkan pola data masa lalu.
b) Menggunakan Metode Matematis atau Komputasional → dapat berupa model sederhana seperti regresi linear, hingga metode kompleks seperti machine learning.
c) Mengandung Unsur Ketidakpastian (Error) → hasil estimasi tidak pernah 100% tepat, sehingga ada residual atau error yang perlu diminimalisasi.
d) Digunakan untuk Keputusan → hasil estimasi dimanfaatkan oleh perusahaan, pemerintah, maupun individu dalam membuat perencanaan jangka pendek maupun panjang.

Contoh penggunaan model estimasi dalam kehidupan nyata:

a) Perusahaan ritel menggunakan estimasi penjualan untuk menentukan stok barang.
b) Bank menggunakan estimasi risiko untuk memberikan pinjaman.
c) Pemerintah menggunakan estimasi inflasi dan pertumbuhan ekonomi untuk merumuskan kebijakan.

2.2 Sejarah Singkat

Konsep estimasi berkembang dari ilmu statistika sejak abad ke-18 dan 19. Beberapa tonggak penting dalam sejarah model estimasi adalah:

1) Awal Perkembangan (1700–1800-an)
    a) Estimasi pertama kali muncul dalam konteks statistik deskriptif dan inferensial.
    b) Carl Friedrich Gauss (1809) memperkenalkan metode kuadrat terkecil (least squares method) untuk memperkirakan parameter dalam model regresi. Metode ini hingga kini menjadi dasar dari analisis regresi modern.
2) Regresi dan Korelasi (Akhir 1800-an)
    a) Sir Francis Galton (1886) memperkenalkan konsep regresi menuju rata-rata (regression to the mean) saat meneliti hubungan antara tinggi badan orang tua dan anak.
    b) Karl Pearson kemudian mengembangkan konsep korelasi yang menjadi dasar pengukuran hubungan antar variabel.
3) Perkembangan Statistik Modern (1900–1950)
    a) Ronald A. Fisher mengembangkan teori inferensi statistik, termasuk konsep estimasi maksimum likelihood (Maximum Likelihood Estimation, MLE).
    b) Estimasi mulai banyak digunakan dalam bidang ekonomi, biologi, dan ilmu sosial.
4) Komputerisasi dan Model Ekonometrik (1950–1980)
    a) Dengan adanya komputer, metode estimasi lebih kompleks mulai dikembangkan, termasuk model ekonometrik untuk memprediksi indikator ekonomi seperti inflasi, pengangguran, dan PDB.
    b) Estimasi deret waktu (time series forecasting) dengan metode ARIMA (Auto-Regressive Integrated Moving Average) dikembangkan oleh Box dan Jenkins (1970).

5) Era Data Mining dan Machine Learning (1990–sekarang)
    a) Perkembangan data mining di tahun 1990-an membawa estimasi ke ranah baru dengan algoritma machine learning seperti decision tree, support vector machine, random forest, neural networks, dan deep learning.
    b) Estimasi tidak hanya digunakan untuk data numerik, tetapi juga pada data teks, citra, hingga big data.

3. Formula Estimasi

Pada kasus penjualan ritel bulanan di Amerika Serikat, data yang digunakan berbentuk deret waktu (time series), artinya data dicatat berdasarkan urutan waktu (tahun dan bulan). Oleh karena itu, ada beberapa formula estimasi yang relevan:

3.1 Regresi Linear Sederhana

Jika hanya melihat hubungan tahun → total penjualan tahunan, kita bisa gunakan model regresi linear:

$$Yt = a + bXt + \epsilon t$$

Keterangan:

1) YtY_tYt = total penjualan pada tahun ke-ttt
2) XtX_tXt = tahun ke-ttt
3) aaa = intercept (penjualan awal)
4) bbb = slope (tingkat pertumbuhan tahunan)
5) ϵt\epsilon_tϵt = error

Contoh penerapan ke dataset:

a. XXX = 1992, 1993, …, 2020
b. YYY = total penjualan tahunan (dalam juta USD)
c. Dari model regresi akan didapatkan nilai aaa dan bbb yang menunjukkan tren pertumbuhan.

3.2 Growth Rate (Laju Pertumbuhan)

Untuk melihat pertumbuhan dari tahun ke tahun, digunakan formula:

$$Growth\ Rate_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \times 100\%$$

Keterangan:

1) YtY_tYt = total penjualan pada tahun ttt
2) Yt−1Y_{t-1}Yt−1 = total penjualan pada tahun sebelumnya

3.3 Model Time Series (Moving Average)

Karena dataset bulanan, salah satu formula sederhana adalah moving average (rata-rata bergerak), digunakan untuk meratakan fluktuasi jangka pendek.

$$MA_n = \frac{Y_t + Y_{t-1} + \cdots + Y_{t-(n-1)}}{n}$$

Keterangan:

1) MAnMA_nMAn = nilai estimasi dengan moving average periode nnn
2) YtY_tYt = nilai aktual pada periode ttt

4. Studi Kasus dengan Excel

4.1 Dataset

Dataset yang digunakan adalah USA Monthly Retail Sales dari tahun 1992 hingga 2020. Data berisi penjualan ritel bulanan serta total tahunan.

4.2 Tujuan Analisis

a) Menganalisis tren total penjualan ritel dan food services di AS dari tahun ke tahun (1992–2020).
b) Membandingkan pertumbuhan tahunan (Growth %) untuk melihat periode kenaikan dan penurunan.
c) Memberikan gambaran visual (grafik) untuk memudahkan interpretasi tren jangka panjang.

4.3 Langkah Pengolahan Data (Excel)

1) Import Data

Dataset dari Kaggle dalam format .xls dibuka di Excel dan Data berisi penjualan bulanan dari tahun 1992–2020.

2) Menambahkan Kolom Total

Menghitung Total Sales per Year dengan formula:

```
=SUM(Januari:Desember)
```

3) Menghitung Growth %

Growth dihitung dengan rumus:

= (Total Tahun Ini - Total Tahun Sebelumnya) / Total Tahun Sebelumnya

Format hasil dalam persentase (%)

4) Membuat Grafik

- Pilih kolom Tahun, Total Sales, Growth %.
- Gunakan Insert → Combo Chart:
    - *Total Sales → Clustered Column*
    - *Growth % → Line (Secondary Axis)*

4.4 Hasil Grafik Analisis

Grafik Combo Chart berjudul "Total Sales & Growth Rate per Year (1992–2020)"

- Batang (Column) → menunjukkan total penjualan tahunan.
- Garis (Line, axis kanan) → menunjukkan persentase pertumbuhan tahunan.

4.5 Interpretasi / Insight

a) Tren Jangka Panjang

Total penjualan ritel AS cenderung meningkat dari tahun ke tahun (dari ±2 juta di 1992 menjadi lebih dari 6 juta di 2020).

b) Fluktuasi Growth %

- Ada tahun dengan pertumbuhan tinggi (7–9%), terutama awal periode.
- Ada periode negatif (misalnya 2009 & 2010 sekitar -1% s.d -7%) yang kemungkinan terkait dengan krisis finansial global 2008–2009.
- Setelah 2010, pertumbuhan stabil di kisaran 3–6%, menunjukkan pasar yang lebih matang.

c) Insight

- Retail & food services tetap menjadi sektor yang stabil dengan tren jangka panjang positif.
- Namun, tetap rentan terhadap kondisi ekonomi makro (krisis, pandemi, dll).

5. Implementasi Program Python

Untuk memperkuat hasil analisis, dilakukan implementasi menggunakan Python dengan library pandas, matplotlib, dll.

5.1 Deskripsi Umum Program

Program ini dibuat untuk menganalisis data "USA Monthly Retail and Food Services Sales" dalam format Excel (.xls) yang berisi data penjualan bulanan dari tahun 1992 hingga 2020. Tujuan utama dari implementasi ini adalah untuk:

1) Mengolah data mentah menjadi format yang siap dianalisis,
2) Menghitung tren pertumbuhan tahunan (Growth Rate),
3) Menganalisis pola musiman dan rata-rata penjualan per bulan,
4) Menampilkan grafik interaktif seperti bar chart, line chart, dan moving average trend.

Analisis dilakukan menggunakan Python dengan bantuan beberapa pustaka data science populer seperti:

- pandas → untuk manajemen dan pembersihan data,
- numpy → untuk operasi numerik,
- matplotlib dan seaborn → untuk visualisasi data,
- datetime → untuk memproses data tanggal.

5.2 Langkah Implementasi

1) Import Library

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
```

Bagian ini mengimpor semua pustaka yang dibutuhkan untuk analisis data dan visualisasi.

- pandas digunakan untuk membaca dan memanipulasi dataset Excel.
- numpy untuk perhitungan matematis dan array.
- matplotlib.pyplot dan seaborn untuk membuat grafik tren dan perbandingan.
- datetime membantu konversi data tahun-bulan menjadi format tanggal Python

2) Membaca Dataset

```
file_path = "mrtssales92-present.xls"
sheets_dict = pd.read_excel(file_path, sheet_name=None)
```

Dataset terdiri dari **banyak sheet (lembar kerja)**, masing-masing mewakili **satu tahun** (misalnya: "1992", "1993", dst).

Dengan sheet_name=None, semua sheet dibaca sekaligus ke dalam dictionary.

3) Fungsi Pemrosesan Data Tiap Sheet

```
def process_sheet_data(sheet_name, sheet_data):
    """Memproses data dari setiap sheet tahunan"""
     retail_sales_row = sheet_data[sheet_data.iloc[:, 1] == "Retail and food services sales, total"]
    if retail_sales_row.empty:
        return None

    monthly_data = retail_sales_row.iloc[:, 2:14].values.flatten()
    months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun',
              'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec']

    year_data = []
    for i, month in enumerate(months):
        if i < len(monthly_data) and not pd.isna(monthly_data[i]):
            year_data.append({
                'Year': sheet_name,
                'Month': month,
                'Sales': monthly_data[i]
            })

    return pd.DataFrame(year_data)
```

- Fungsi ini bertugas mengambil baris data yang mengandung teks *"Retail and food services sales, total"*, kemudian mengekstrak **nilai penjualan bulanan (Jan–Dec)** untuk tiap tahun.

  Hasilnya dikonversi menjadi Data Frame berisi kolom:

- Year
- Month
- Sales

4) Penggabungan Seluruh Tahun

```
all_data = []

for sheet_name, sheet_data in sheets_dict.items():
    if sheet_name.isdigit() and 1992 <= int(sheet_name) <= 2020:
        processed_data = process_sheet_data(sheet_name, sheet_data)
        if processed_data is not None:
            all_data.append(processed_data)
df = pd.concat(all_data, ignore_index=True)
```

- Fungsi Kode di atas menyatukan semua hasil ekstraksi dari setiap tahun (sheet) menjadi satu tabel besar bernama df.

5) Pra-pemrosesan Data

```
df['Year'] = df['Year'].astype(int)
df['Sales'] = pd.to_numeric(df['Sales'], errors='coerce')
df['Date'] = pd.to_datetime(df['Year'].astype(str) + '-' + df['Month'], format='%Y-%b')
```

- Mengubah tipe data tahun menjadi integer,
- Konversi kolom Sales menjadi numerik,
- Gabungkan kolom Year dan Month menjadi kolom waktu (Date) agar bisa divisualisasikan sebagai deret waktu (time series).

5.3 Analisis

1) Analisis 1 - Total Tahunan & Growth Rate

```
annual_sales = df.groupby('Year')['Sales'].sum().reset_index()
annual_sales['Growth_Rate'] = annual_sales['Sales'].pct_change() * 100
```

- Mengelompokkan data berdasarkan tahun dan menghitung total penjualan per tahun.
- Menggunakan rumus Growth Rate (%):

$$Growth\ Rate = \frac{Sales_t - Sales_{t-1}}{Sales_{t-1}} \times 100$$

- Visualisasi dilakukan dengan bar chart (total penjualan) dan line chart (growth rate) untuk melihat fluktuasi setiap tahun.

2) **Analisis 2 – Perbandingan Antar Bulan**

```
monthly_avg = df.groupby('Month')['Sales'].mean().reset_index()
```

- Analisis ini menghitung rata-rata penjualan tiap bulan dari seluruh periode 1992–2020.
- Hasilnya divisualisasikan dengan diagram batang (bar chart) untuk melihat bulan mana yang paling tinggi penjualannya.

3) **Analisis 3 – Tren Penjualan Bulanan**

```
plt.plot(df['Date'], df['Sales']/1e6, linewidth=2, color='blue', alpha=0.7)
```

- Bagian ini menampilkan tren penjualan bulanan (time series).
- Program juga menandai titik penjualan tertinggi dan terendah menggunakan anotasi panah (plt.annotate).

4) **Analisis 4 – Moving Average (Trend Jangka Panjang)**

```
df['MA_12'] = df['Sales'].rolling(window=12).mean()
```

- Moving Average (MA) 12 bulan digunakan untuk menghaluskan fluktuasi data bulanan dan memperlihatkan arah tren jangka panjang.
- Rumusnya :

$$MA_{12} = \frac{X_t + X_{t-1} + \ldots + X_{t-11}}{12}$$

5) **Statistik dan Analisis Musiman**

```
seasonal_analysis = df.groupby('Month').agg({
    'Sales': ['mean', 'std', 'min', 'max']})
```

a. Kode ini menghitung:

- Rata-rata (mean) penjualan per bulan,
- Standar deviasi (std) untuk melihat variasi,
- Nilai minimum dan maksimum tiap bulan.

b. Dari hasil ini dapat diketahui bulan-bulan dengan penjualan tertinggi dan terendah secara konsisten.

6) Ekspor Hasil ke CSV

```
df.to_csv('retail_sales_analysis.csv', index=False)
annual_sales.to_csv('annual_retail_sales.csv', index=False)
```

- Hasil analisis disimpan ke file .csv agar dapat digunakan kembali untuk visualisasi tambahan atau laporan analitik lebih lanjut.

5.4 Hasil Output

Program menghasilkan:

a. Grafik Total Tahunan & Growth Rate
b. Rata-rata Penjualan Bulanan
c. Tren Penjualan Bulanan (1992–2020)
d. Moving Average 12 Bulan
e. Statistik dan Analisis Musiman
f. File CSV hasil analisis.

6. Kesimpulan dan Penutup

6.1 Kesimpulan

Berdasarkan hasil analisis dan implementasi program Python terhadap dataset *USA Monthly Retail and Food Services Sales (1992–2020)*, diperoleh beberapa kesimpulan utama sebagai berikut:

1) Pertumbuhan Penjualan Secara Umum Positif

   Dari hasil perhitungan *annual growth rate*, sektor retail dan layanan makanan di Amerika Serikat menunjukkan tren pertumbuhan positif selama hampir tiga dekade. Nilai pertumbuhan kumulatif mencapai lebih dari +260%, yang mengindikasikan peningkatan konsumsi masyarakat dari waktu ke waktu.

2) Adanya Pola Musiman (Seasonal Pattern)

   Berdasarkan rata-rata penjualan bulanan, ditemukan bahwa bulan November dan Desember secara konsisten memiliki penjualan tertinggi setiap tahunnya. Hal ini berkaitan dengan periode liburan akhir tahun seperti Thanksgiving, Black Friday, dan Natal.

   Sebaliknya, bulan Januari dan Februari menunjukkan nilai terendah akibat penurunan konsumsi pasca-liburan.

3) Tren Jangka Panjang yang Meningkat

   Hasil visualisasi menggunakan Moving Average 12 bulan menunjukkan bahwa meskipun terdapat fluktuasi kecil, secara umum tren penjualan terus meningkat dari

tahun 1992 hingga 2020. Hal ini menandakan bahwa sektor retail memiliki daya tahan yang kuat terhadap perubahan ekonomi global.

4) Dampak Ekonomi terhadap Pola Penjualan

   Beberapa periode seperti tahun 2008–2009 mengalami perlambatan signifikan akibat krisis keuangan global, sementara tahun 2020 menunjukkan efek penurunan akibat pandemi COVID-19. Meski demikian, secara keseluruhan industri tetap menunjukkan pemulihan yang baik.

5) Pentingnya Analisis Data Mining dalam Pengambilan Keputusan

   Penggunaan pendekatan *data mining* dengan bahasa pemrograman Python terbukti efektif untuk:

   a) Mengidentifikasi tren dan pola penjualan,
   b) Mendeteksi anomali atau fluktuasi,
   c) Menyediakan dasar bagi perencanaan bisnis, pemasaran, dan logistik secara berbasis data (*data-driven decision making*).

6.2 Penutup

Melalui penelitian ini, dapat disimpulkan bahwa penerapan *data mining* dengan dukungan bahasa pemrograman Python memberikan kontribusi nyata dalam memahami perilaku penjualan dan tren pasar jangka panjang.

Analisis berbasis data historis tidak hanya membantu pengambilan keputusan strategis, tetapi juga membuka peluang untuk melakukan prediksi dan inovasi di sektor ekonomi berbasis data.

Dengan demikian, penelitian ini menjadi dasar yang kuat untuk pengembangan sistem analitik penjualan yang lebih cerdas, akurat, dan adaptif terhadap perubahan pasar global.