Modelling SARIMA for Time series forecasting of investment

Presented by Maulana Aripianto

maulanaaripianto@gmail.com

https://www.linkedin.com/in/maulana-aripianto/

Experience

1 **Bridgestone Tire Indonesia (Jan 2022- Jun 2025)**
Full Stack Developer

- Analis data dan membuat aplikasi berbasis website menggunakan bahasa pemprograman ASP .NET, JavaScript, CSS, SQL Server & MySQL

2 **PT Pertamina Hulu Energi (Jul 2025 -Aug 2025)**
Full Stack Developer

- Projek tim aplikasi website CRM untuk manage data vendor Pertamina Hulu Energi
- ASP .NET Core, JavaScript, CSS, SQL Server, API

Content

1

DESCRIPTION

DESCRIPTION

Investasi berperan penting **bagi pertumbuhan ekonomi Indonesia,** baik dari **PMA** (Penanaman Modal Asing) maupun **PMDN** (Penanaman Modal Dalam Negeri). Data kuartalan **2010-2025** menunjukkan **tren jangka panjang** dengan **fluktuasi musiman.**

Dengan karakteristik tersebut, data ini sangat relevan untuk dijadikan **proyek Data Science**. Melalui penerapan model **Seasonal ARIMA (SARIMA),** kita dapat membangun sistem **peramalan investasi yang akurat.** Hasil prediksi ini bermanfaat untuk **mendukung kebijakan pemerintah, strategi investor, dan analisis ekonomi secara data-driven.**

2

Data Understanding

Data Understanding

```
<class 'pandas.core.frame.DataFrame'>
Index: 571979 entries, 0 to 604678
Data columns (total 17 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   period                   571979 non-null  object
 1   investment_status        571979 non-null  object
 2   region                   571979 non-null  object
 3   country                  571979 non-null  object
 4   main_sector              571979 non-null  object
 5   sector_name              571979 non-null  object
 6   kbli_2digit              571979 non-null  object
 7   province                 571979 non-null  object
 8   district_city            571979 non-null  object
 9   java_outside_java        571979 non-null  object
 10  island                   571979 non-null  object
 11  investment_idr_million   571979 non-null  float64
 12  investment_usd_thousand  571979 non-null  float64
 13  indonesian_workers       571979 non-null  int64
 14  year                     571979 non-null  int64
 15  quarter                  571979 non-null  object
 16  log_investment           571977 non-null  float64
dtypes: float64(3), int64(2), object(12)
memory usage: 78.5+ MB
```

Time series forecasting of investment

Kami berupaya **menganalisis tren investasi di Indonesia (2010-2025)** dan mengembangkan model peramalan deret waktu **menggunakan SARIMA.** Model ini digunakan u**ntuk memprediksi nilai** penanaman modal asing (**PMA**) dan penanaman modal dalam negeri (**PMDN**) **di masa depan.** Hasil peramalan diharapkan dapat membantu pemerintah dan investor dalam mengantisipasi dinamika ekonomi serta merancang strategi kebijakan yang lebih tepat.

571979
Rows
17
Column

Kaggle : http://bit.ly/46e8Qmb

3

Data Preprocessing

Data Preprocessing

Duplicated Data

```
# Hapus Data Duplikat
print("Jumlah duplikat:", df.duplicated().sum())

# hapus duplikat
df = df.drop_duplicates()
```

```
Jumlah duplikat: 32700
```

```
print("Jumlah duplikat:", df.duplicated().sum())
```

```
Jumlah duplikat: 0
```

Missing Values

	0
period	0
investment_status	0
region	0
country	0
main_sector	0
sector_name	0
kbli_2digit	0
province	0
district_city	0
java_outside_java	0
island	0
investment_idr_million	0
investment_usd_thousand	0
indonesian_workers	0
year	0
quarter	0

dtype: int64

Agregasi Data

```
df_group = df.groupby("period")["investment_idr_million"].sum().reset_index()
df_group['date'] = pd.PeriodIndex(df_group['period'], freq='Q').to_timestamp()
df_group = df_group.set_index('date')
df_group
```

date	period	investment_idr_million
2010-01-01	2010-Q1	42,198,184
2010-04-01	2010-Q2	50,826,129
2010-07-01	2010-Q3	56,762,575
2010-10-01	2010-Q4	58,979,466
2011-01-01	2011-Q1	53,627,588
...	...	...
2024-04-01	2024-Q2	428,409,235
2024-07-01	2024-Q3	431,476,757
2024-10-01	2024-Q4	452,796,223
2025-01-01	2025-Q1	465,211,584
2025-04-01	2025-Q2	477,658,737

62 rows × 2 columns

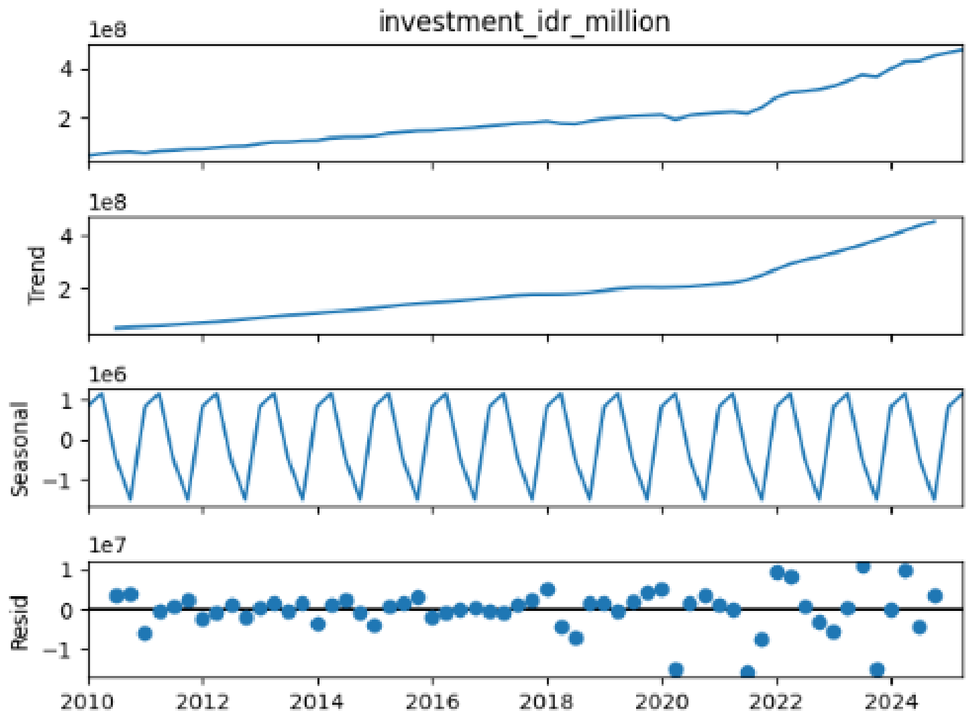
4

EDA
(Exploratory Data Analysis)

Trend Analysis Investasi 2010-2025

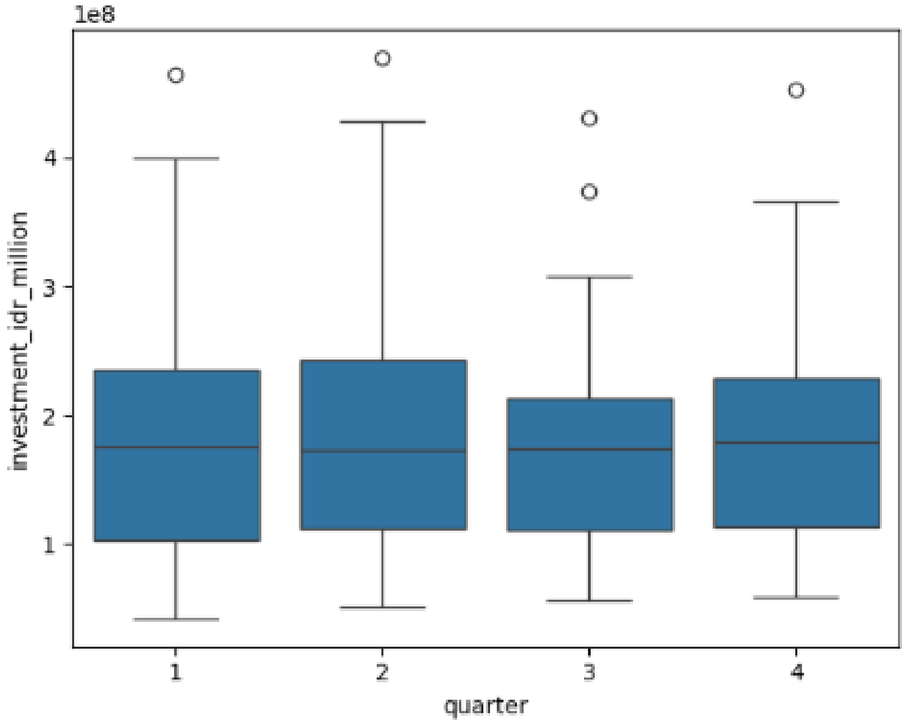
Situasi Modal Terkini: Investasi menunjukkan **tren naik kuat** sepanjang **2010-2025**, mencapai **level tertinggi** pada **2024-2025**. Setelah sempat **melambat di 2018-2021**, momentum kembali **menguat sejak 2022** dengan peningkatan yang konsisten.

Seasonality Analysis



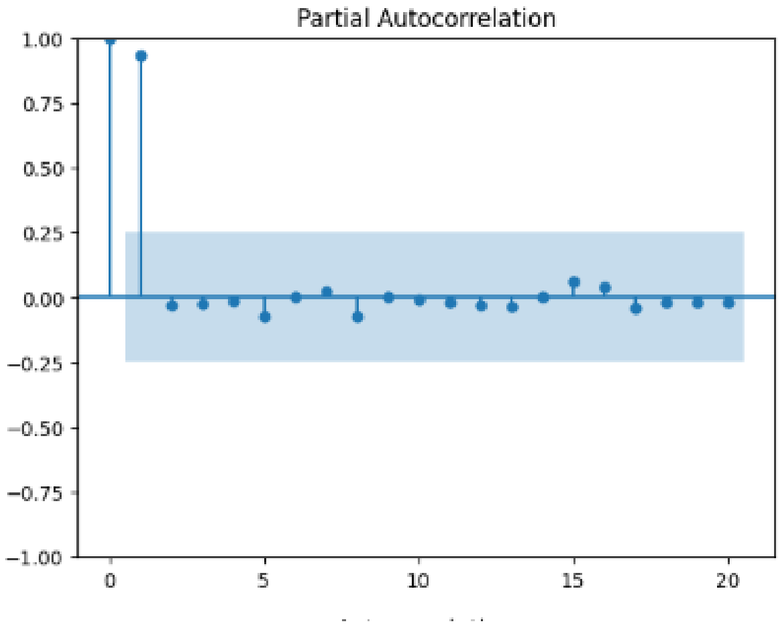
Investasi naik konsisten **sepanjang 2010-2025** dengan **akselerasi sejak 2022**. Pola musiman kuartalan stabil, sementara nilai residu pergerakan acak dengan beberapa lonjakan sesaat (sekitar 2020-2021 dan 2023).

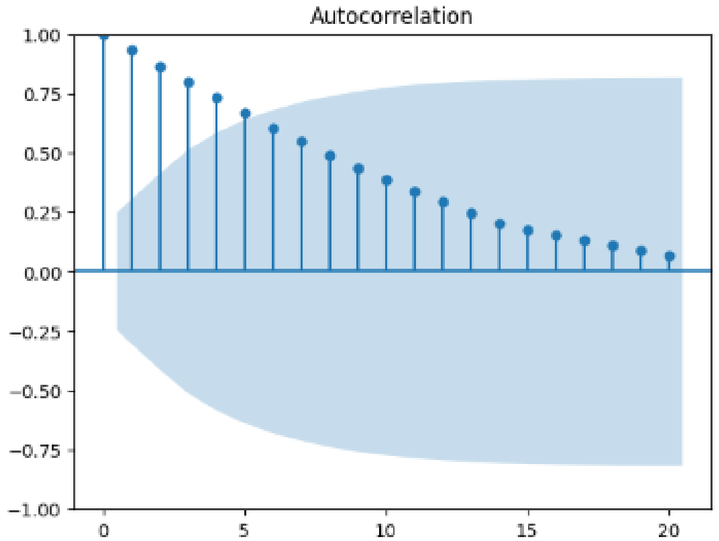
Quarterly Pattern

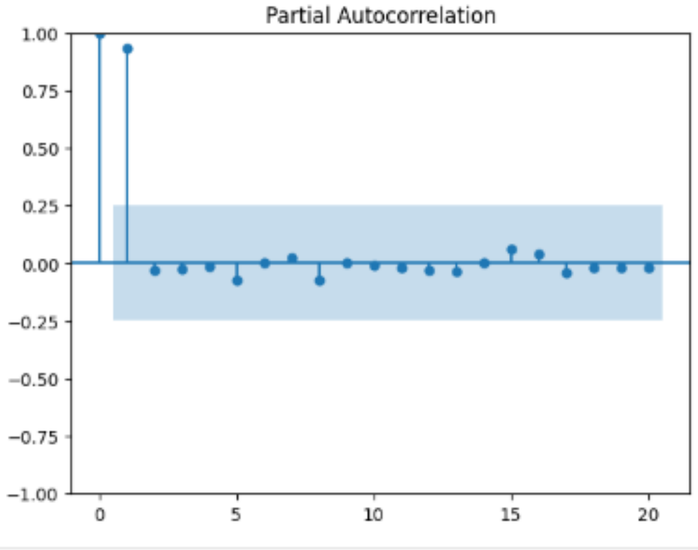


Median investasi per kuartal relatif mirip, dengan **Q2 dan Q4 sedikit lebih tinggi dibanding Q1-Q3**. Sebaran (IQR) antar kuartal juga serupa, sementara **Q3 cenderung sedikit lebih rendah.** Terdapat beberapa **outlier tinggi di semua kuartal**, terutama pada Q1, Q2, dan Q4.

Autocorrelation (ACF & PACF)







- ACF tinggi di lag-1 lalu turun perlahan → deret sangat persisten/bertren.
- PACF hanya spike besar di lag-1 → indikasi AR(1); AR lebih tinggi tidak kuat.

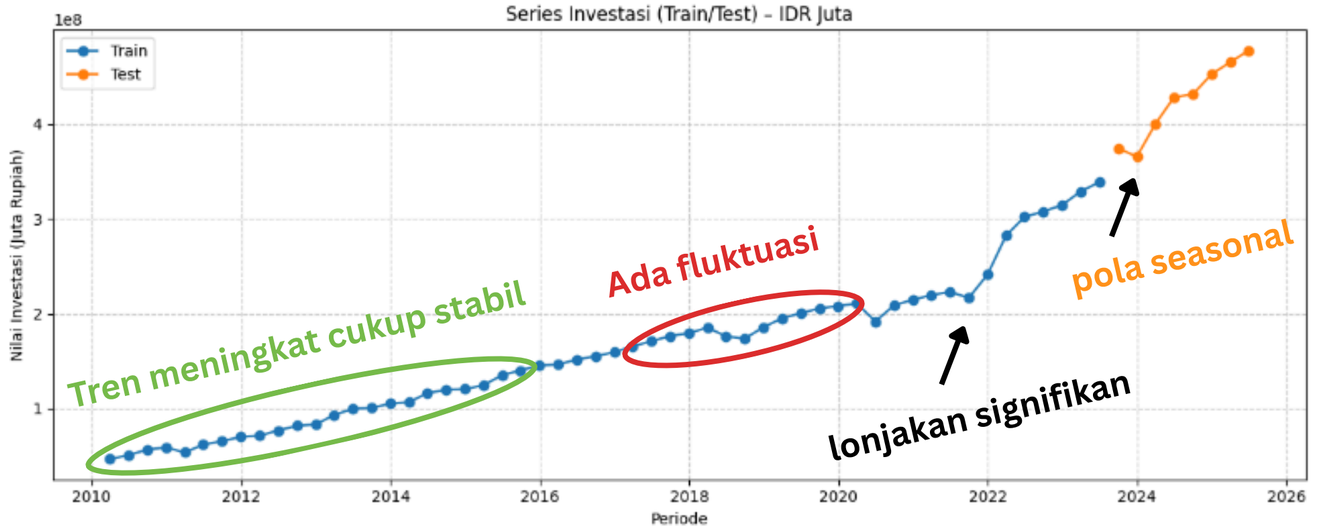
6 Model Preparation

Model Preparation

1. Split Train & Test Data
2. Hyperparameter Tuning (p,q,P,Q)
3. Fit Best Model & Evaluation
4. Residual Diagnostic
5. Forecast Future

Model Preparation

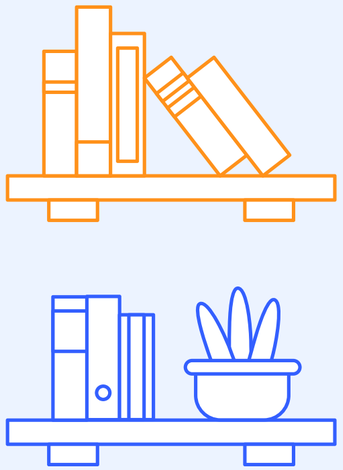
```
Setelah reindex: len=62 | NaN=0 | Range: 2010-03-31 → 2025-06-30
Train: 2010-03-31 → 2023-06-30 (n=54)
Test : 2023-09-30 → 2025-06-30 (n=8)
Missing di TRAIN sebelum imputasi: 0
Missing di TEST : 0

=== ADF (TRAIN) ===
[ADF] raw: stat=1.912 | p=0.999
[ADF] diff(1): stat=-5.391 | p=0.000
[ADF] diff(4): stat=-2.255 | p=0.187
[ADF] diff(1)+diff(4): stat=-3.887 | p=0.002
```

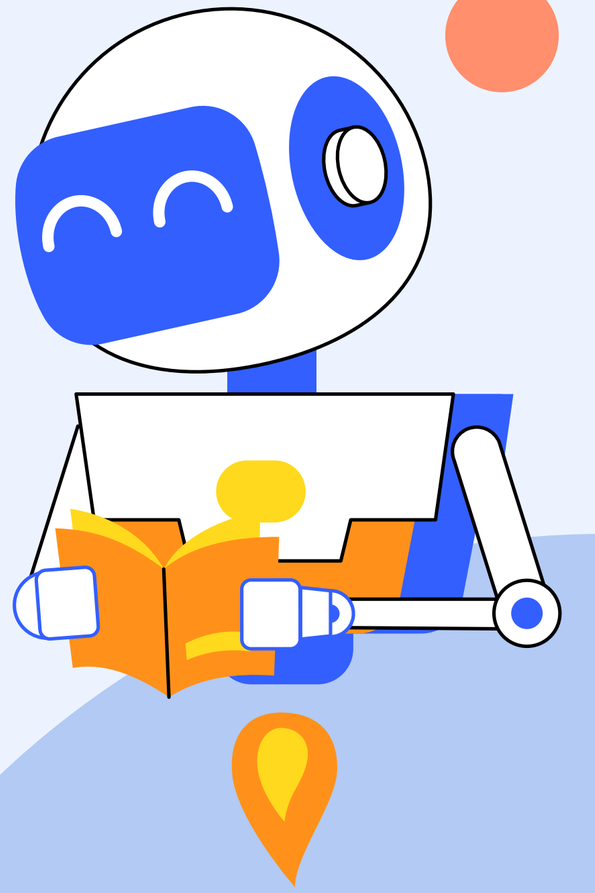


- **2010-2016: Tren meningkat cukup stabil,** kenaikan perlahan tapi konsisten.
- **2017-2020: Ada fluktuasi (sedikit naik-turun),** tapi tren jangka panjang tetap naik.
- **2021-2022: Terjadi lonjakan signifikan** (kemungkinan shock eksternal/kenaikan investasi besar**).**
- **2023-2025 (Test):** Data test **menunjukkan pertumbuhan berlanjut**, dengan **pola seasonal "zig-zag"** kuartalan yang lebih jelas.

7

Machine Learning Model & Evaluate





Baseline SARIMA + EVALUASI

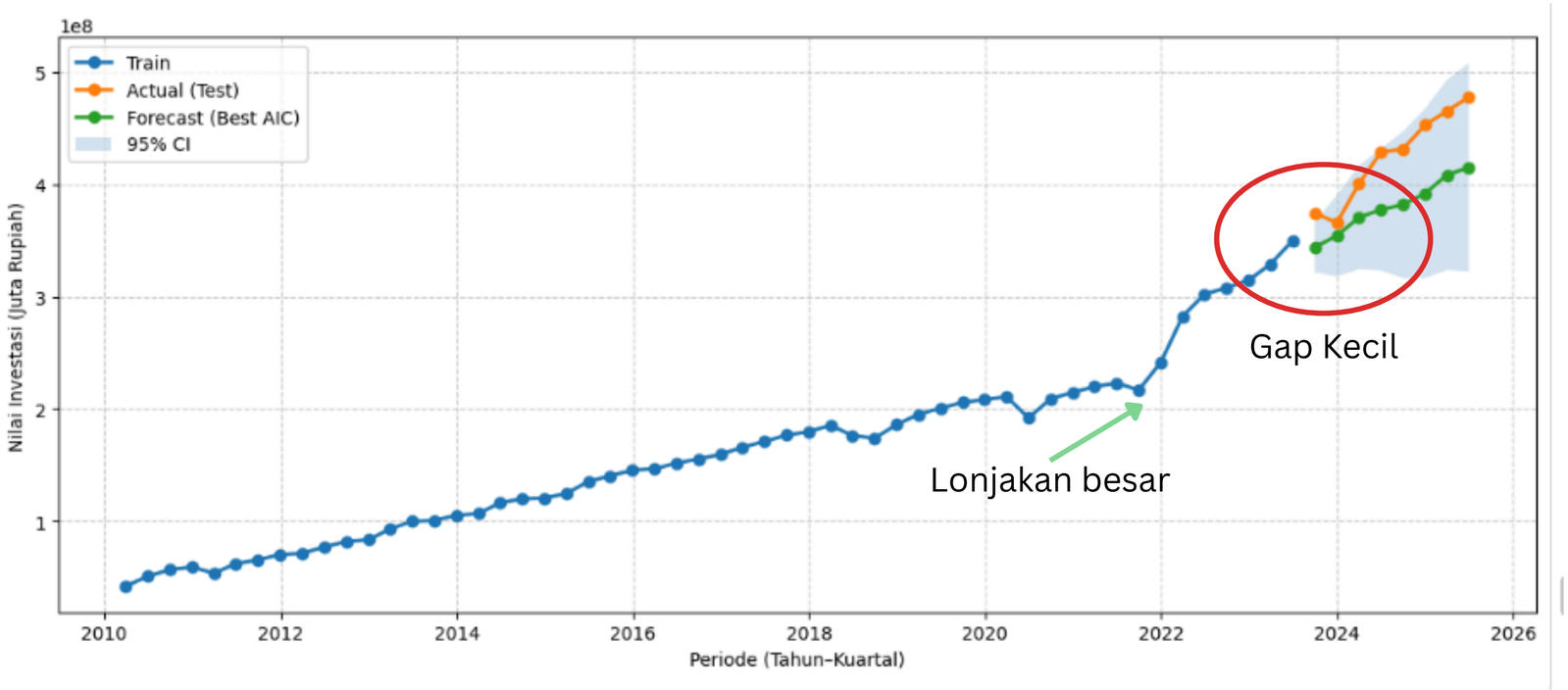
```
order=(1, 1, 0) seasonal=(1, 1, 0, 4)  AIC=1546.82
order=(0, 1, 1) seasonal=(0, 1, 1, 4)  AIC=1509.71
order=(1, 1, 1) seasonal=(0, 1, 1, 4)  AIC=1511.66
order=(1, 1, 0) seasonal=(0, 1, 1, 4)  AIC=1543.25

Best: (0, 1, 1) (0, 1, 1, 4) AIC= 1509.7086219512678
```

```
MAE : 18,637,851.98
RMSE: 19,929,006.72
MAPE: 4.32%
```

rata-rata error relatif hanya sekitar 4% dari nilai aktual, cukup rendah → model punya akurasi yang bagus.

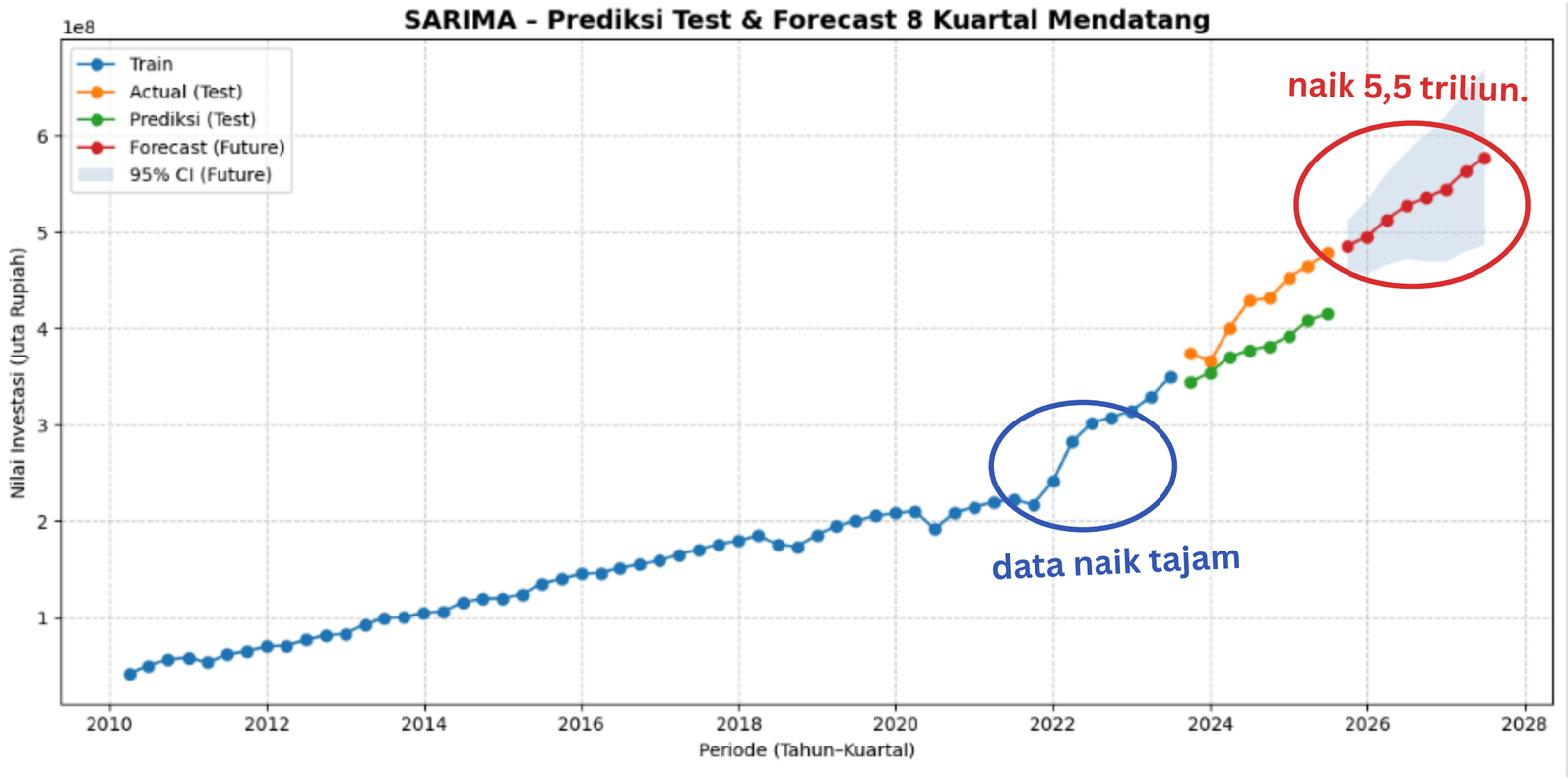
Hyndman & Koehler (2006), International Journal of **Forecasting: MAPE** populer karena mudah dipahami, dan **angka rendah (misalnya <5%**) menunjukkan **deviasi rata-rata yang kecil,** sehingga **dianggap excellent fit.**

Best SARIMA (AIC): Test-Set Forecast & Evaluation



- **Lonjakan besar 2021-2022** masih berpengaruh pada baseline model; SARIMA tidak sepenuhnya menangkap kenaikan ekstrem ini.
- **Gap antara forecast vs aktual 2023-2025** relatif kecil → membuktikan model cukup robust.
- Proyeksi jangka panjang menunjukkan **kenaikan bertahap** ke kisaran **480-500 triliun.**
- Model bisa jadi underestimate di fase pertumbuhan pesat, sehingga **SARIMAX/fundamental analysis** disarankan untuk validasi tambahan.

Future Forecast (mis. 8 kuartal ke depan)



- Grafik memperlihatkan historis investasi **naik tajam hingga 2022**, lalu **nilai aktual 2024 relatif datar**. **Proyeksi 8 kuartal ke depan (garis hijau)** Investasi **naik stabil sejak 2010,** dengan **lonjakan besar di 2021-2022.**
- **Prediksi SARIMA untuk 2023-2025** cukup **akurat**, meski sedikit **di bawah angka aktual.**
- **Proyeksi 8 kuartal ke depan (2026-2027)** menunjukkan tren masih **naik**, bisa tembus di atas **5,5 triliun**.

Artinya, investasi diperkirakan **terus tumbuh** dalam **2 tahun ke depan**.
kenaikan bertahap menuju **kisaran ~480-500 triliun,**

Recomendation and Action

Hasil Model SARIMA (Prediksi Investasi)

- Model punya akurasi yang bagus → rata-rata salahnya hanya sekitar 18,6 juta per kuartal (MAE),
- kesalahan rata-ratanya 19,9 juta (RMSE), dan kesalahan persentasenya kecil, hanya 4,3% (MAPE).
- Artinya, prediksi model bisa dipercaya untuk melihat tren ke depan.

Apa yang Terlihat dari Data:

1. Investasi terus naik dari tahun ke tahun.
2. Prediksi model sedikit lebih rendah dari kenyataan, artinya pertumbuhan bisa lebih cepat.
3. Semakin jauh ke depan, ketidakpastian makin besar.
4. Ada pola musiman tiap kuartal.
5. Pernah ada lonjakan besar di 2021–2022 akibat faktor luar.

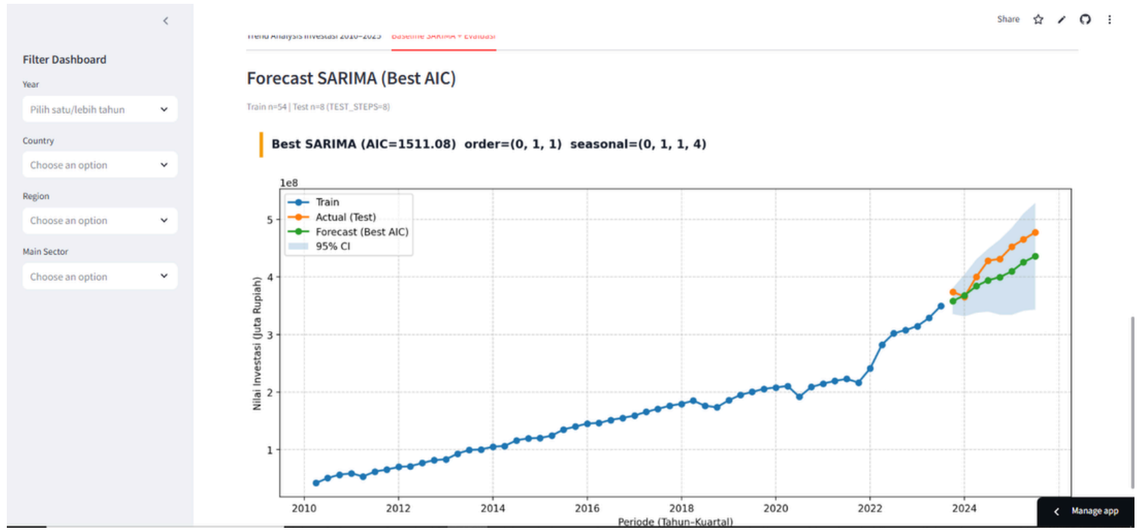
Recomendation and Action

Apa yang Sebaiknya Dilakukan:

- Pasang target pertumbuhan lebih tinggi untuk 2 tahun ke depan.
- Siapkan rencana cadangan kalau pertumbuhan lebih cepat dari prediksi.
- Gunakan model ini sebagai alat pantau, perbarui setiap kuartal.
- Buat 3 rencana: hati-hati, normal, dan optimis untuk mengantisipasi ketidakpastian.
- Atur anggaran sesuai musim/kuartal agar lebih efisien.
- Pelajari penyebab lonjakan 2021-2022 dan siapkan langkah cepat kalau kondisi serupa terulang.
- Tambahkan faktor luar (misalnya kebijakan atau kondisi ekonomi) agar prediksi makin kuat.

Singkatnya: Model SARIMA sudah cukup akurat (MAPE 4,3%), tren investasi tetap naik, dan perlu strategi fleksibel agar siap menghadapi pertumbuhan lebih cepat maupun ketidakpastian di masa depan.

Streamlit



http://bit.ly/3VsYIzJ

Terima Kasih