

STARTUP SUCCESS PREDICTION

Analyst



Johannes Christian

Data Scientist



Topik
Latar Belakang
Studi Literatur
Data
Metodologi



DISCUSSION POINTS



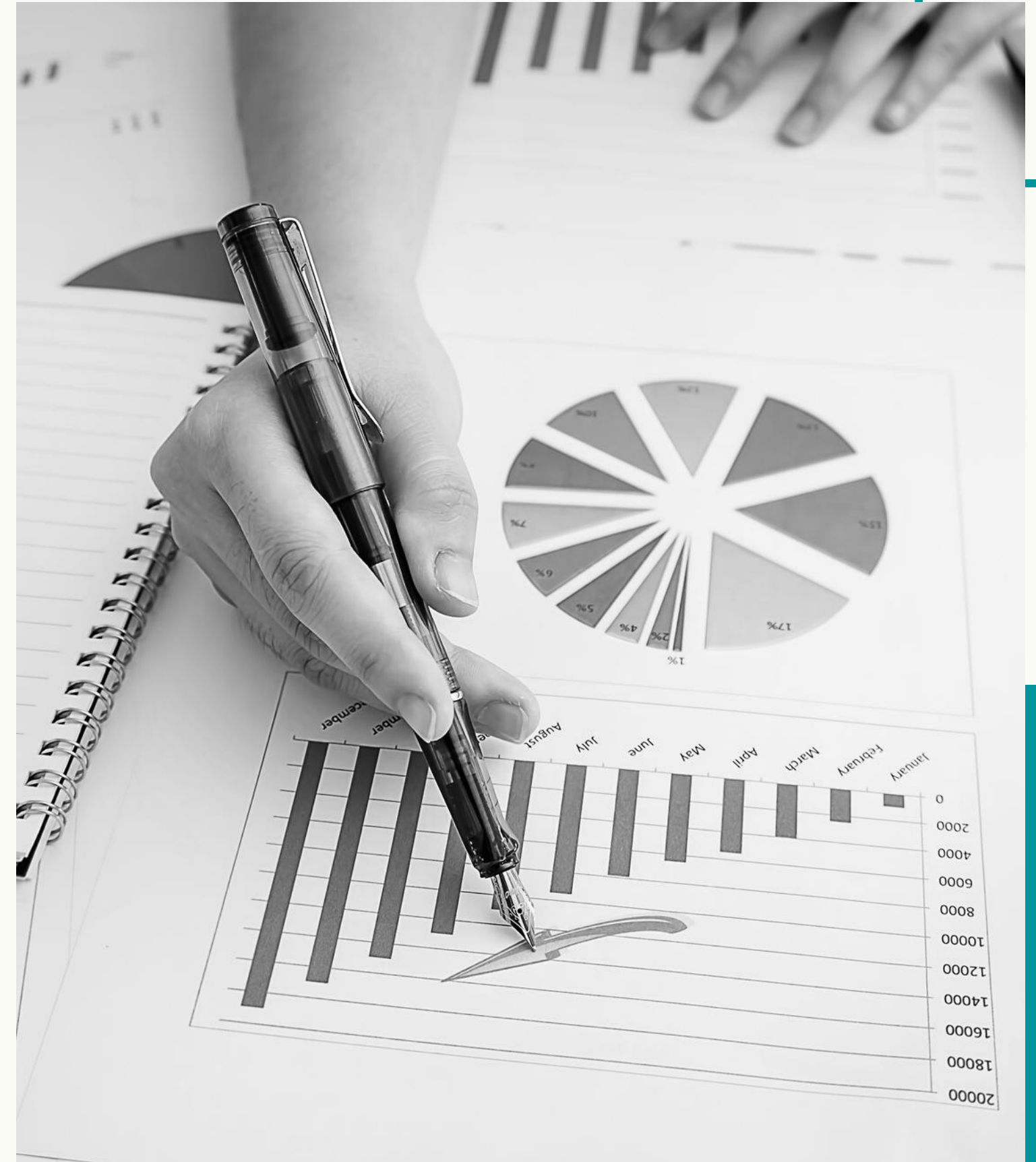
TOPIK

PREDIKSI KESUKSESAN STARTUP

Permasalahan : Bidang Bisnis

Topik yang akan digunakan :

- Exploratory Data Analysis
- Feature Engineering
- Preprocessing Data
- Classification



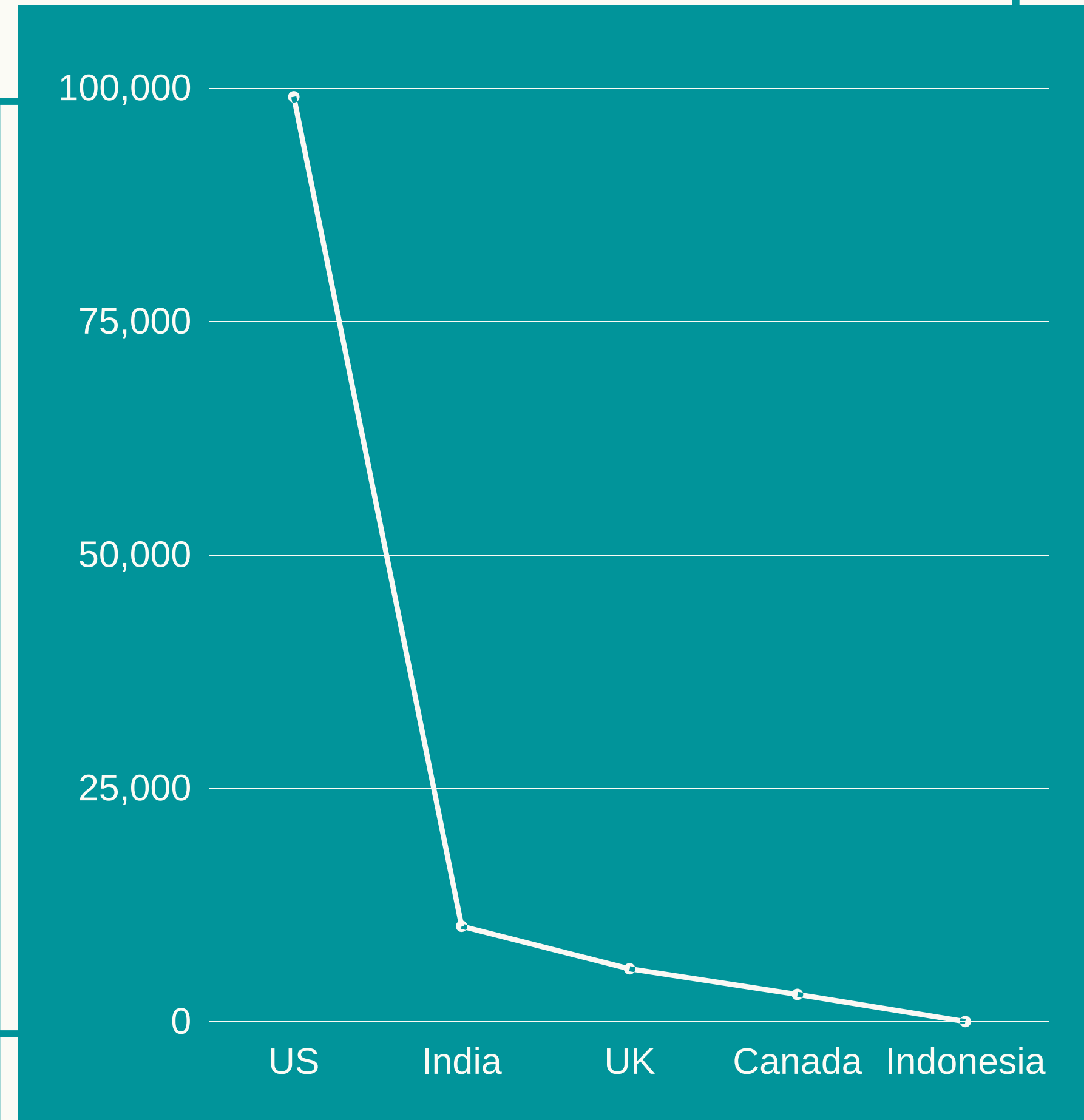
“

LATAR BELAKANG

Startup merupakan sebuah perusahaan yang dibangun oleh seorang wirausahawan yang mencari, mengembangkan, dan memverifikasi model ekonomi yang terukur. Pada saat ini startup memainkan peran penting dalam pertumbuhan ekonomi. Mereka membawa ide-ide baru, mendorong inovasi, dan menciptakan kesempatan kerja yang mendorong pembangunan ekonomi.

PERKEMBANGAN STARTUP DI DUNIA

Dalam beberapa tahun terakhir, perkembangan perusahaan startup telah tumbuh secara eksponensial. Berdasarkan data yang dimiliki oleh **StartupRanking** (3/6/2021), pada saat ini Amerika menduduki peringkat pertama dengan jumlah start up terbanyak di dunia.



Memprediksi keberhasilan start-up memungkinkan investor untuk menemukan perusahaan dengan potensi pertumbuhan yang cepat, sehingga memungkinkan mereka untuk tetap selangkah lebih maju dari persaingan.





| | | | | |
|--------|--------|--------|--------|--------|
| 00 | 1.0000 | .00000 | 1.0000 | .00000 |
| 52 | .90484 | .10017 | 1.0050 | .09967 |
| 14 | .81873 | .20134 | 1.0201 | .19738 |
| 99 | .74082 | .30452 | 1.0453 | .29131 |
| 18 | .67032 | .41075 | 1.0811 | .37995 |
| 87 | .60653 | .52110 | 1.1276 | .46212 |
| 21 | .54881 | .63665 | 1.1855 | .53705 |
| 188 | .49659 | .75858 | 1.2552 | .60437 |
| 2.2255 | .44933 | | 1.3374 | .66404 |
| 2.4596 | .4065 | | 1.4331 | .71630 |
| 2.7183 | | 1.1752 | | .76159 |
| 3.0042 | | 1.3356 | | .80050 |
| 3.3201 | | 1.5095 | | .83365 |
| 3.6693 | | 1.6984 | | .86172 |
| 4.0552 | | 1.9043 | | .88535 |
| 4.4817 | 2.2313 | 2.1293 | 2.352 | .90515 |
| 4.9530 | 2.0190 | 2.3756 | 2.577 | .92167 |
| 5.4739 | 1.8268 | 2.6456 | 2.828 | .93541 |
| 6.0496 | 1.6530 | 2.9422 | 3.107 | .94681 |
| 6.6859 | 1.4957 | 3.2682 | 3.41 | .95624 |
| 7.3891 | 1.34 | 3.6269 | 3.7 | .96403 |
| 8.1662 | 1.18 | 4.0219 | 4.0 | .97045 |
| 9.0250 | 1.02 | 4.4571 | 4.3 | .97574 |
| 9.9742 | 0.86 | 4.9370 | 4.6 | .98010 |
| 11.023 | 0.70 | 5.5569 | 5.0 | .98367 |
| 12.182 | 0.54 | 6.1323 | 5.4 | .98661 |
| 13.464 | 0.38 | 6.7690 | 5.8 | .98903 |
| 14.880 | 0.22 | 7.4735 | 6.2 | .99101 |
| 16.445 | 0.06 | 8.2527 | 6.6 | .99263 |
| 18.174 | | 9.1146 | 7.0 | .99396 |
| 20.086 | | 10.018 | 7.4 | .99505 |
| 22.198 | | 11.076 | 7.8 | .99595 |
| 24.533 | | 12.246 | 8.2 | .99668 |
| 27.113 | | 13.538 | 8.6 | .99728 |
| 29.964 | | 14.965 | 9.0 | .99777 |
| 33.115 | | 16.543 | 9.4 | .99818 |
| 36.598 | | 18.285 | 9.8 | .99851 |
| 40.447 | | 20.211 | 10.2 | .99878 |
| 44.701 | | 22.339 | 10.6 | .99900 |
| 49.402 | | 24.691 | 11.0 | .99918 |

DATASET

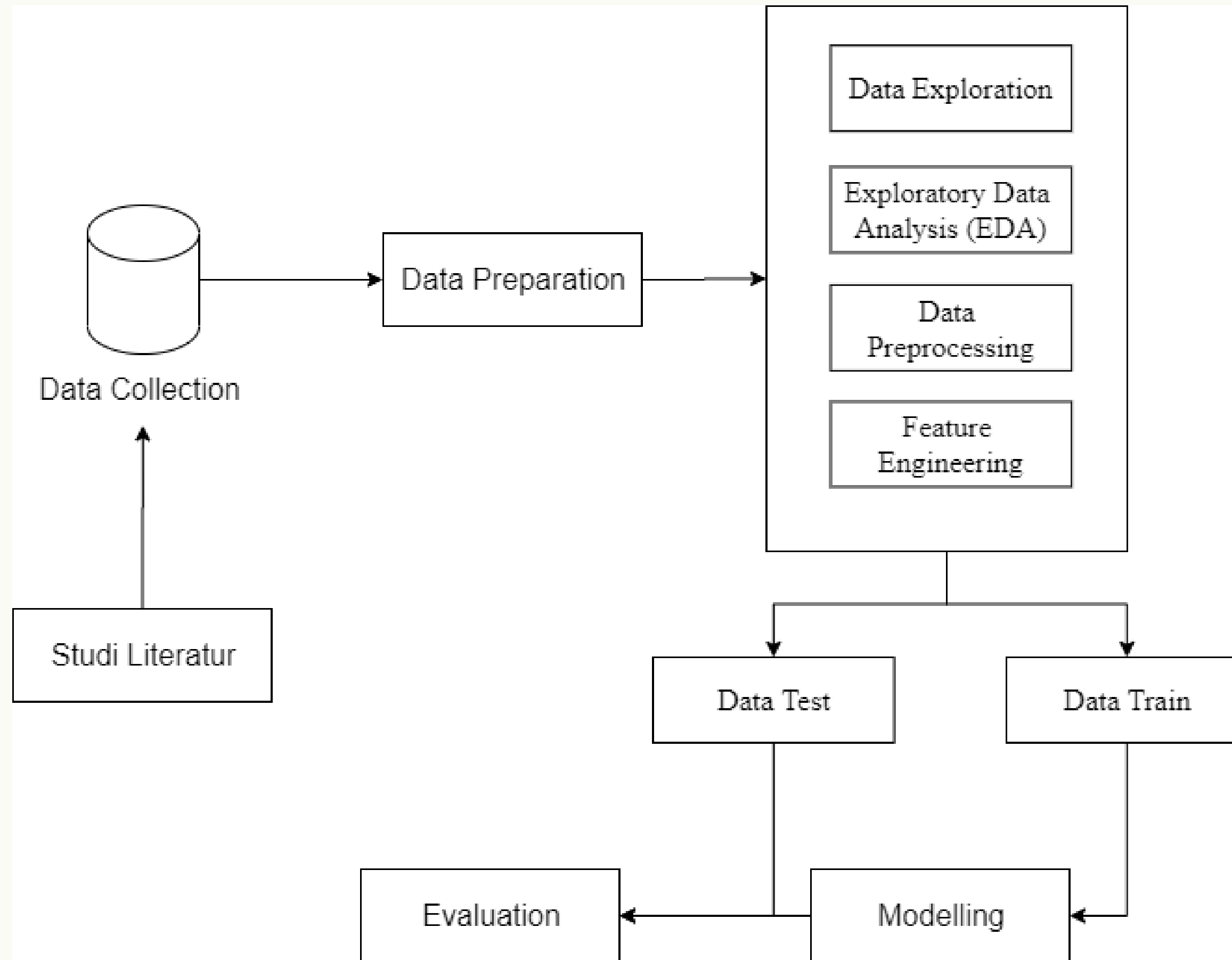
Kaggle :

(<https://www.kaggle.com/manishkco6/startup-success-prediction?select=startup+data.csv>)

Klasifikasi :

Support Vector Machine

METODOLOGI





DATA PREPARATION

Data Preparation merupakan proses/langkah yang perlu dilakukan untuk mengubah data mentah menjadi data yang berkualitas agar mendapatkan nilai akurasi yang terbaik. Tahapan dalam data preparation meliputi :

- Data Exploration
- Exploratory Data Analysis (EDA)
- Data Preprocessing
- Feature Engineering

Proses learning pada SVM adalah dengan mencari support vector untuk memperoleh hyperplane terbaik. Hyperplane terbaik diperoleh dengan memaksimalkan jarak antara hyperplane dan data training serta melewati pertengahan antara dua kelas

MODELLING : SUPPORT VECTOR MACHINE



HASIL DAN PEMBAHASAN

- Pengubahan bentuk nilai pada variabel “status” dimana acquired diubah menjadi 1 dan closed menjadi 0.
- Penghapusan kolom “label” karena adanya kesamaan antara kolom variabel “label” dengan “status”.
- Penghapusan kolom “state_code.1” karena antara kolom “state_code” dan “state_code.1” hanya memiliki satu value yang berbeda

EKSPLORASI DATA

ANALISIS DATA

Dari data startup yang digunakan juga dapat diambil beberapa gambaran baru untuk perkembangan startup di Indonesia kedepannya



DATA

| | Unnamed: 0 | state_code | latitude | longitude | zip_code | id | city | Unnamed: 6 | name | labels | founded_at | closed_at | first_funding_at | last_funding_at |
|---|------------|------------|-----------|-------------|----------|---------|-----------|-----------------------|----------------------|--------|------------|-----------|------------------|-----------------|
| 0 | 1005 | CA | 42.358880 | -71.056820 | 92101 | c:6669 | San Diego | NaN | Bandsintown | 1 | 1/1/2007 | NaN | 4/1/2009 | 1/1/2010 |
| 1 | 204 | CA | 37.238916 | -121.973718 | 95032 | c:16283 | Los Gatos | NaN | TriCipher | 1 | 1/1/2000 | NaN | 2/14/2005 | 12/28/2009 |
| 2 | 1001 | CA | 32.901049 | -117.192656 | 92121 | c:65620 | San Diego | San Diego CA 92121 | Plixi | 1 | 3/18/2009 | NaN | 3/30/2010 | 3/30/2010 |
| 3 | 738 | CA | 37.320309 | -122.050040 | 95014 | c:42668 | Cupertino | Cupertino CA 95014 | Solidcore Systems | 1 | 1/1/2002 | NaN | 2/17/2005 | 4/25/2007 |

| age_first_funding_year | age_last_funding_year | age_first_milestone_year | age_last_milestone_year | relationships | funding_rounds | funding_total_usd | milestones | state_code.1 | is_CA |
|------------------------|-----------------------|--------------------------|-------------------------|---------------|----------------|-------------------|------------|--------------|-------|
| 2.2493 | 3.0027 | 4.6685 | 6.7041 | 3 | 3 | 375000 | 3 | CA | 1 |
| 5.1260 | 9.9973 | 7.0055 | 7.0055 | 9 | 4 | 40100000 | 1 | CA | 1 |
| 1.0329 | 1.0329 | 1.4575 | 2.2055 | 5 | 1 | 2600000 | 2 | CA | 1 |

| is_NY | is_MA | is_TX | is_otherstate | category_code | is_software | is_web | is_mobile | is_enterprise | is_advertising | is_gamesvideo | is_ecommerce | is_biotech | is_consulting |
|-------|-------|-------|---------------|---------------|-------------|--------|-----------|---------------|----------------|---------------|--------------|------------|---------------|
| 0 | 0 | 0 | 0 | music | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | enterprise | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | web | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

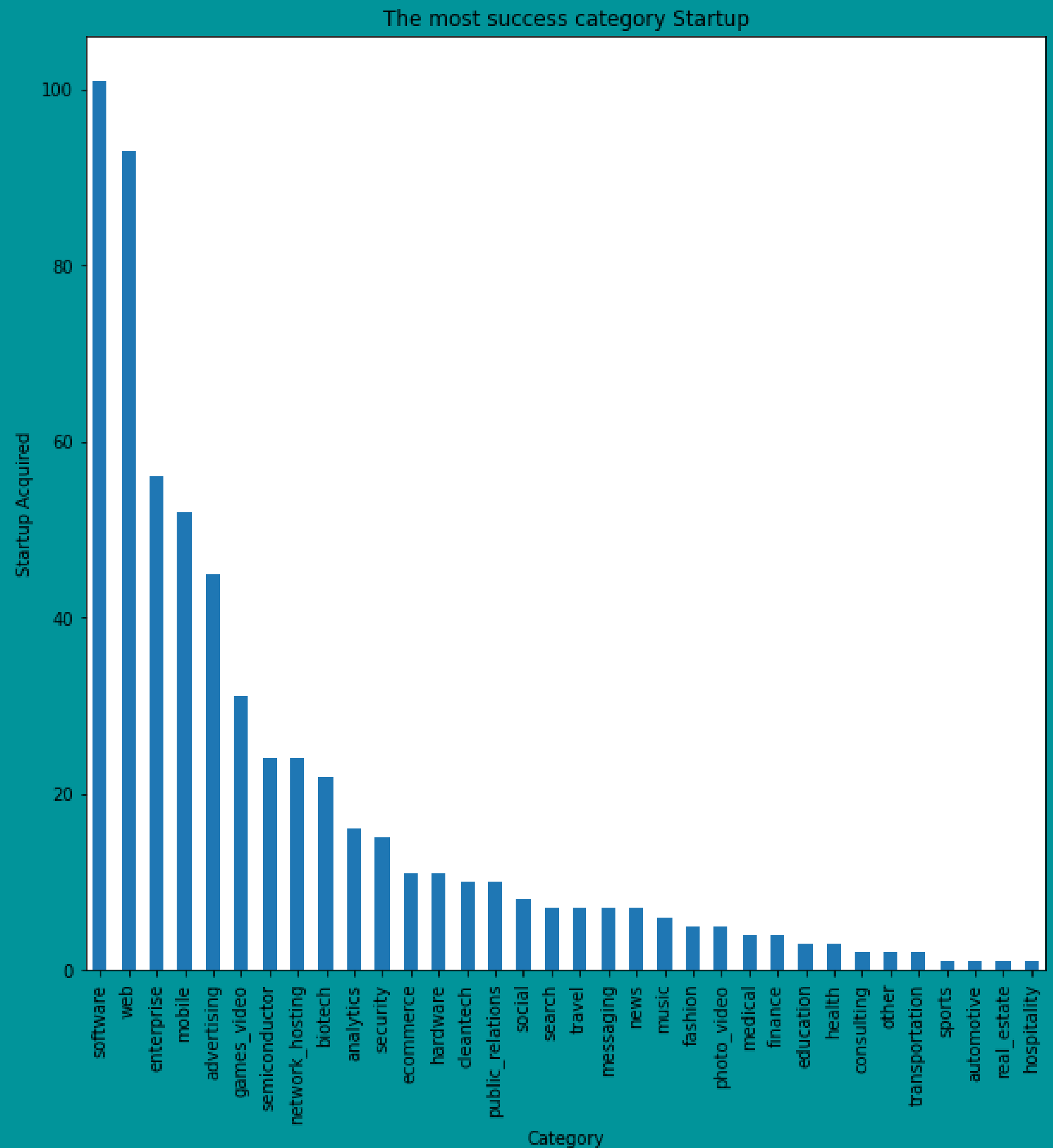
| is_othercategory | object_id | has_VC | has_angel | has_roundA | has_roundB | has_roundC | has_roundD | avg_participants | is_top500 | status |
|------------------|-----------|--------|-----------|------------|------------|------------|------------|------------------|-----------|----------|
| 1 | c:6669 | 0 | 1 | 0 | 0 | 0 | 0 | 1.0000 | 0 | acquired |
| 0 | c:16283 | 1 | 0 | 0 | 1 | 1 | 1 | 4.7500 | 1 | acquired |
| 0 | c:65620 | 0 | 0 | 1 | 0 | 0 | 0 | 4.0000 | 1 | acquired |

KATEGORI STARTUP DENGAN JUMLAH KEBERHASILAN TERTINGGI

TOP 3 :

- Software,
- Web dan
- Enterprise.

Terendah : Hospitality

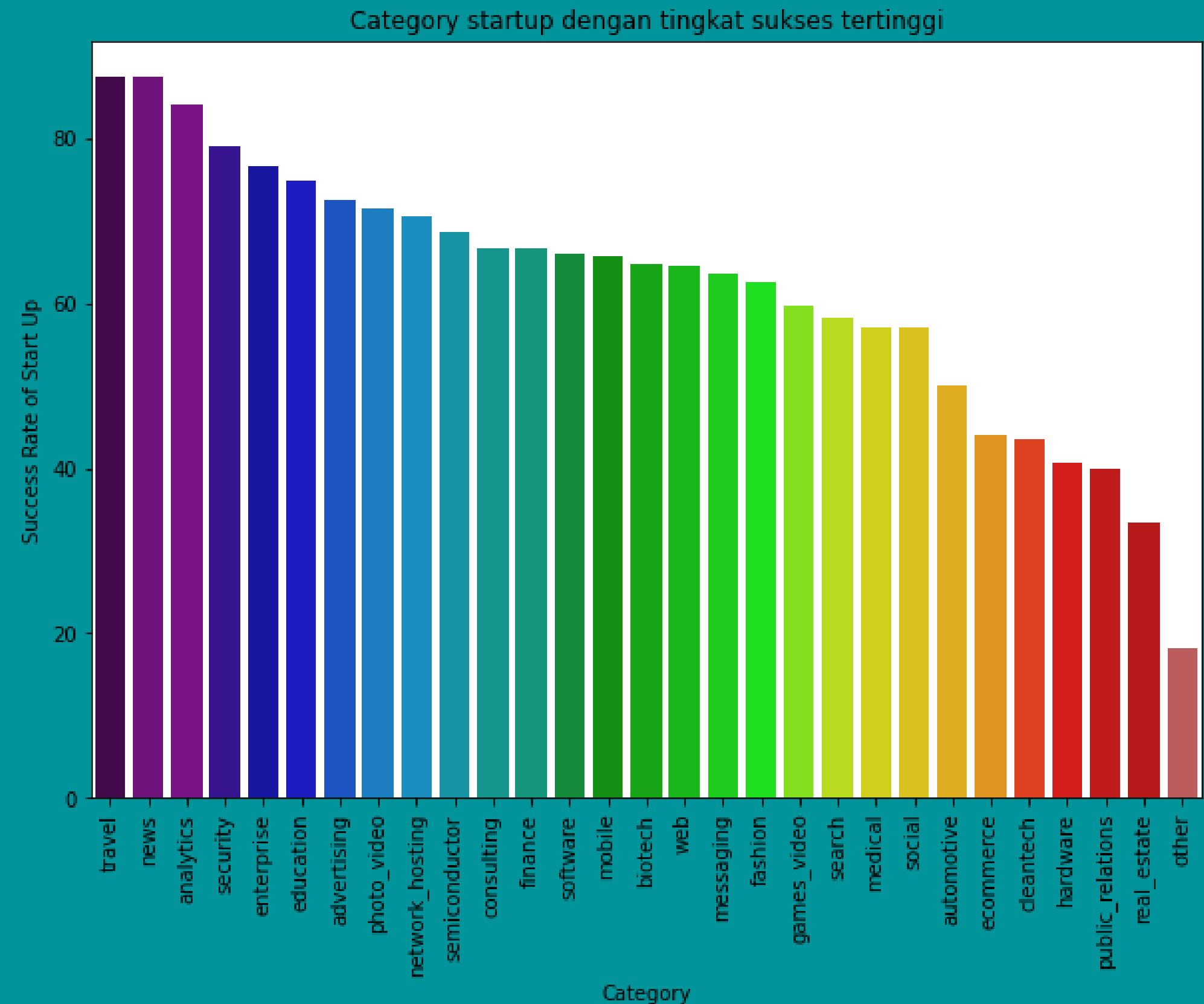


KATEGORI STARTUP DENGAN TINGKAT KESUKSESAN TERTINGGI

TOP 3 :

- Travel,
- News dan
- Analytics.

Terendah : Other

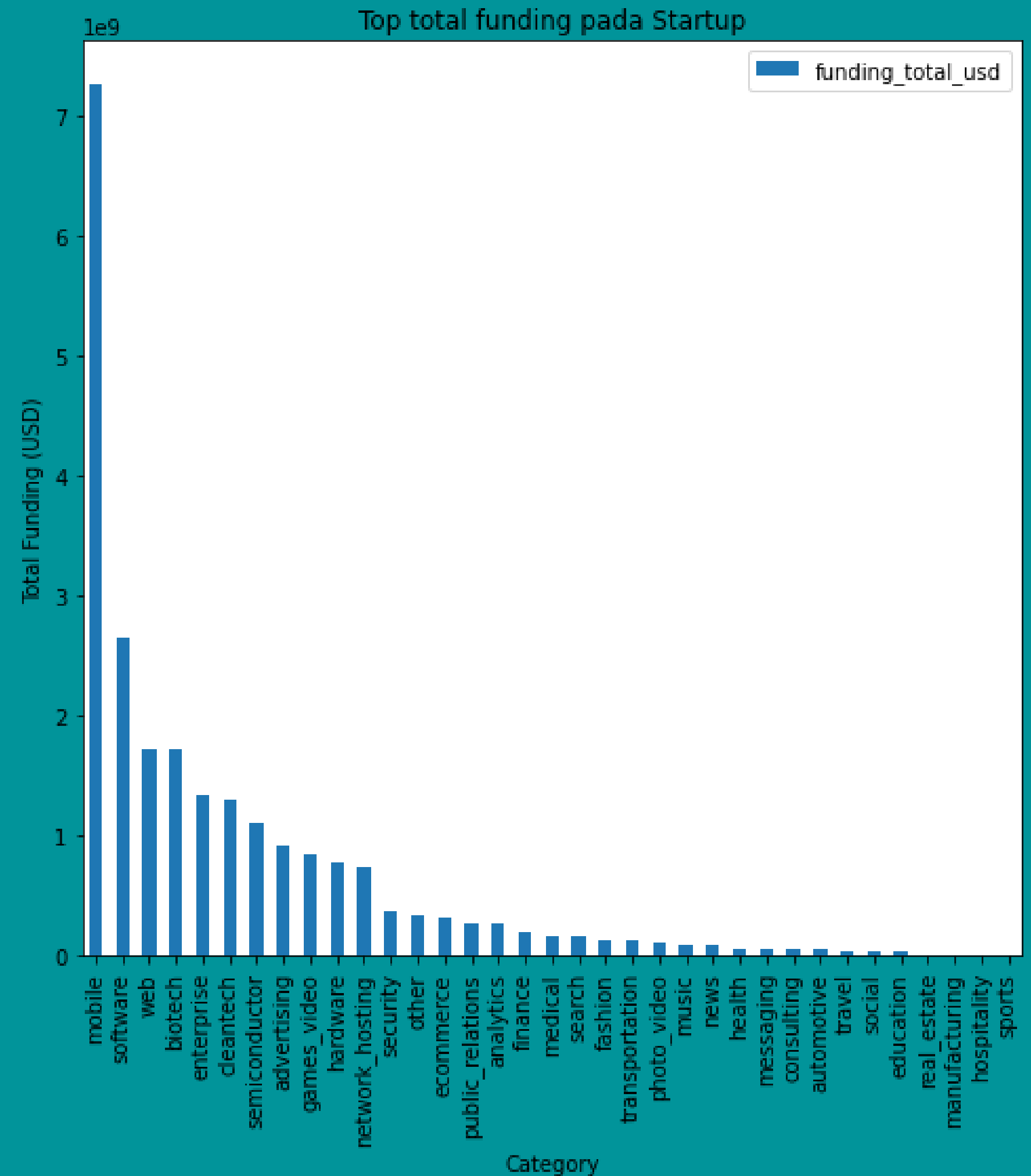


KATEGORI STARTUP DENGAN FUNDING TERBESAR

TOP 3 :

- Mobile,
- Software dan
- Web.

Terendah : Sports



DATA PREPROCESSING

IDENTIFIKASI MISSING VALUE

| Features | Null Values | Missing (%) |
|--------------------------|-------------|-------------|
| <u>Unnamed:6</u> | 493 | 53.41 |
| <u>closed_at</u> | 588 | 63.7 |
| age_first_milestone_year | 152 | 16.46 |
| age_last_milestone_year | 152 | 16.46 |

- Fitur “Unnamed:6” : fitur gabungan dari beberapa fitur diantaranya city, state_code, dan zip_code -> penghapusan kolom “Unnamed:6” dan kemudian diisi kembali dengan gabungan dari fitur city, state_code, dan zip_code.
- Fitur “closed_at” : data tanggal penutupan startup -> tidak mendukung dalam pemodelan sehingga dilakukan penghapusan kolom.
- Missing value pada variabel age first milestone year dan age last milestone year disebabkan startup tersebut belum ada milestones -> diisi dengan 0

DATA PREPROCESSING

DUPLICATE DATA

| No | Unnamed: 0 | state_code | latitude | longitude | zip_code | id | city | Unnamed: 6 | name | ... | status |
|-----|------------|------------|----------|-------------|----------|---------|---------|------------------|-----------------|-----|--------|
| 832 | 505 | CA | 37.48151 | -121.945328 | 94538 | c:28482 | Fremont | Fremont CA 94538 | Redwood Systems | ... | 1 |

Dalam pengecekan tidak ditemukan adanya duplikasi data, namun sebelumnya pada eksplorasi data telah dilakukan analisis kategorikal dan ditemukan fitur yang memiliki kesamaan record data diantaranya name, id, dan object_id -> penghapusan pada ketiga fitur tersebut

DATA PREPROCESSING

IDENTIFIKASI NEGATIVE VALUE

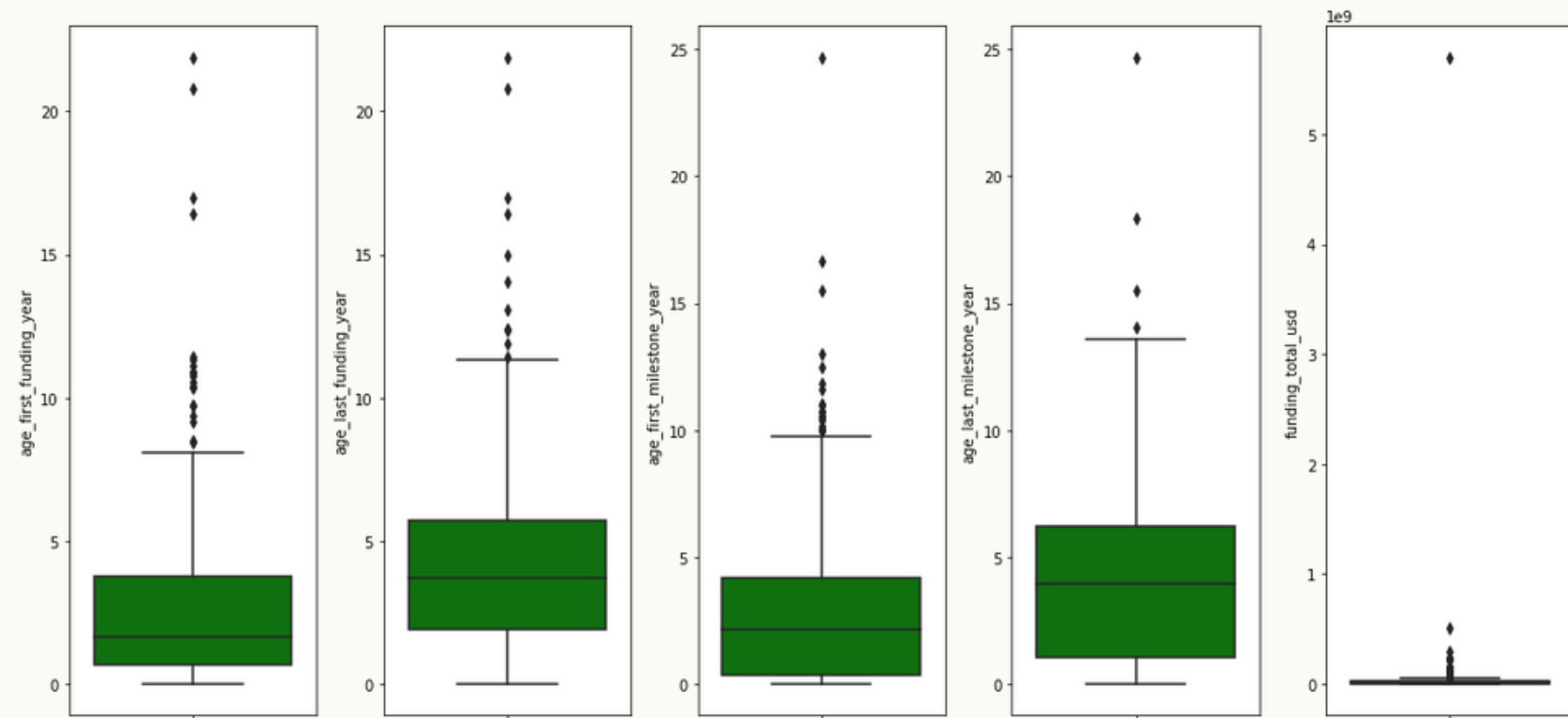
| Features | Negative Values |
|--------------------------|-----------------|
| age_first_funding_year | 46 |
| age_last_funding_year | 13 |
| age_first_milestone_year | 46 |
| age_last_milestone_year | 12 |

Ditemukan adanya value yang bernilai negatif pada kolom age first funding year, age last funding year, age first milestones year, dan age last milestones year -> penghapusan pada kolom yang memiliki nilai negatif

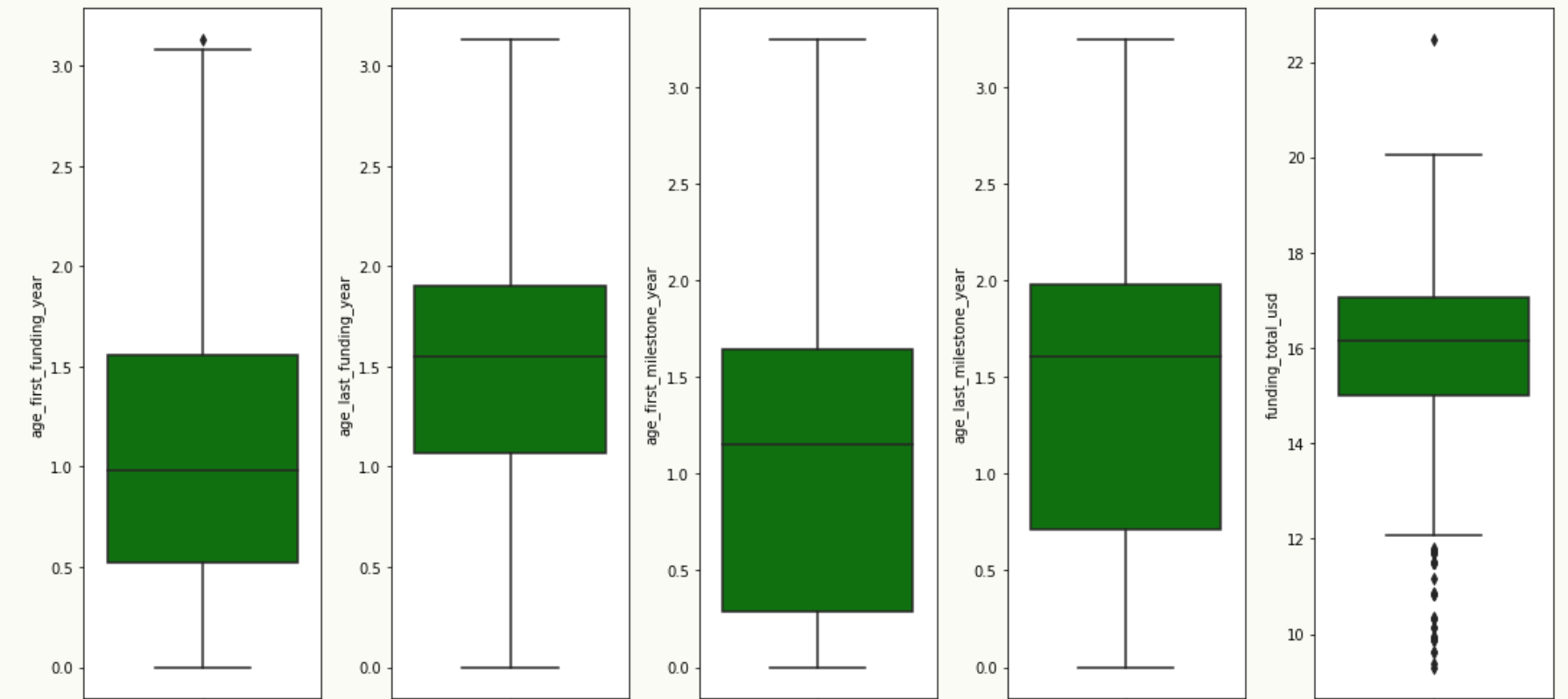
DATA PREPROCESSING

HANDLING OUTLIER

BEFORE



AFTER



Untuk mengatasi outlier, kami menggunakan log-transformation pada fitur age_first_funding_year, age_last_funding_year, age_first_milestone_year, age_last_milestone_year, dan funding_total_usd

FEATURE ENGINEERING

1

HAS_ROUNDABCD

membuat fitur baru untuk melihat apakah startup tersebut memiliki funding baik A, B, C ataupun D.

2

HAS_INVESTOR

membuat fitur baru untuk melihat apakah suatu perusahaan startup memiliki seorang investor atau tidak

3

HAS_SEED

membuat fitur baru untuk melihat apakah suatu perusahaan startup sudah memiliki pendanaan yang digunakan untuk penumbuhan startup atau tidak




DATA MODELLING

SUPPORT VECTOR
MACHINE (SVM)

Pembagian Data

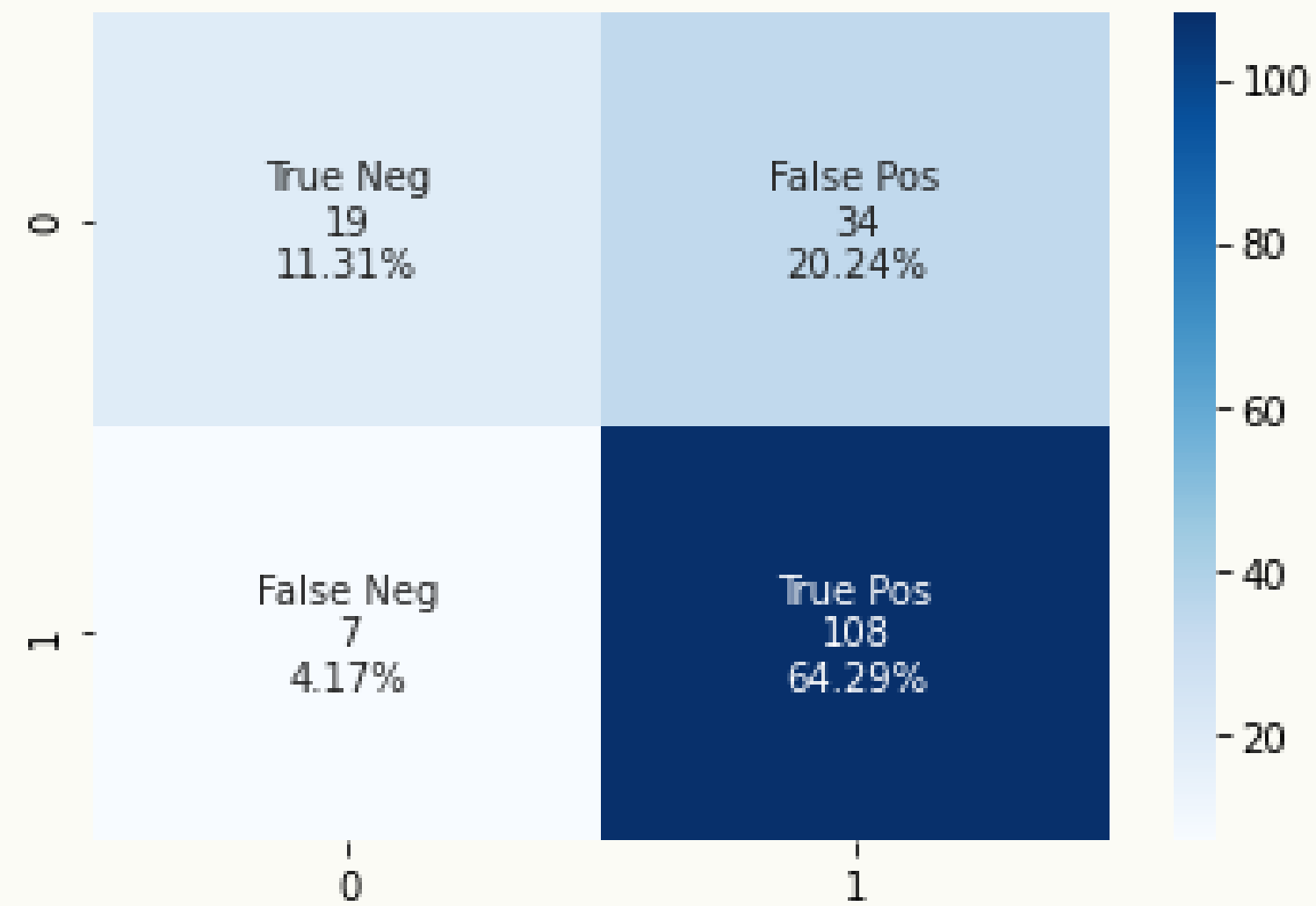
Data dibagi menjadi data training dan data testing dengan perbandingan 80% sebagai data training dan 20% sebagai data testing. Pembagian data ini menghasilkan pembagian untuk data training sejumlah 671 data dan data testing sejumlah 168 data.



EVALUASI

Evaluasi pada penelitian ini menggunakan metode evaluasi confusion matrix.

Tabel confusion matrix direpresentasikan dengan 4 nilai klasifikasi, yaitu : True Positive (TP), True Negative (TN), False Positif (FP) dan False Negative (FN).

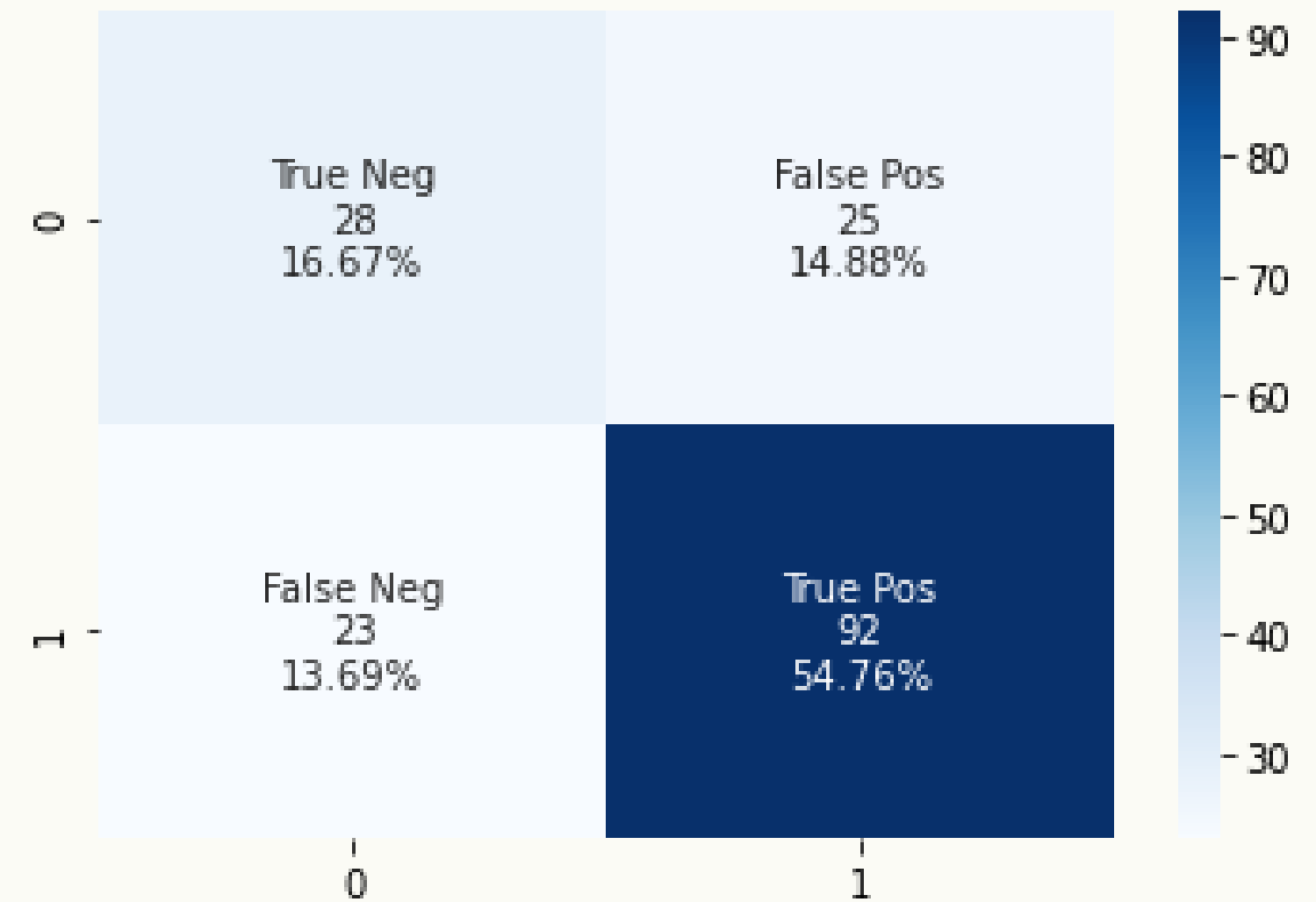


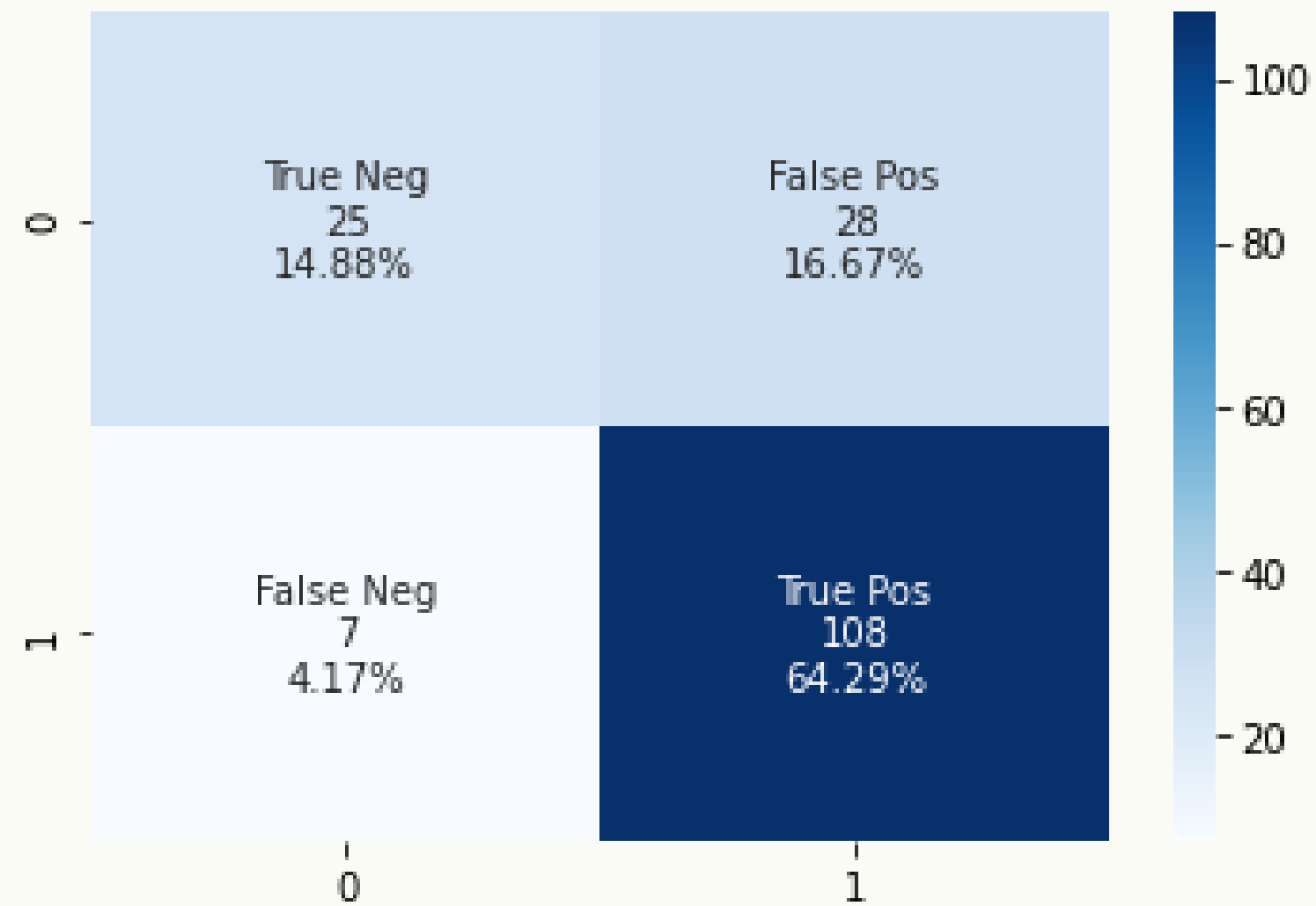
C=1.0, KERNEL=RBF

75.6%

C=100.0, KERNEL=RBF

71.43%



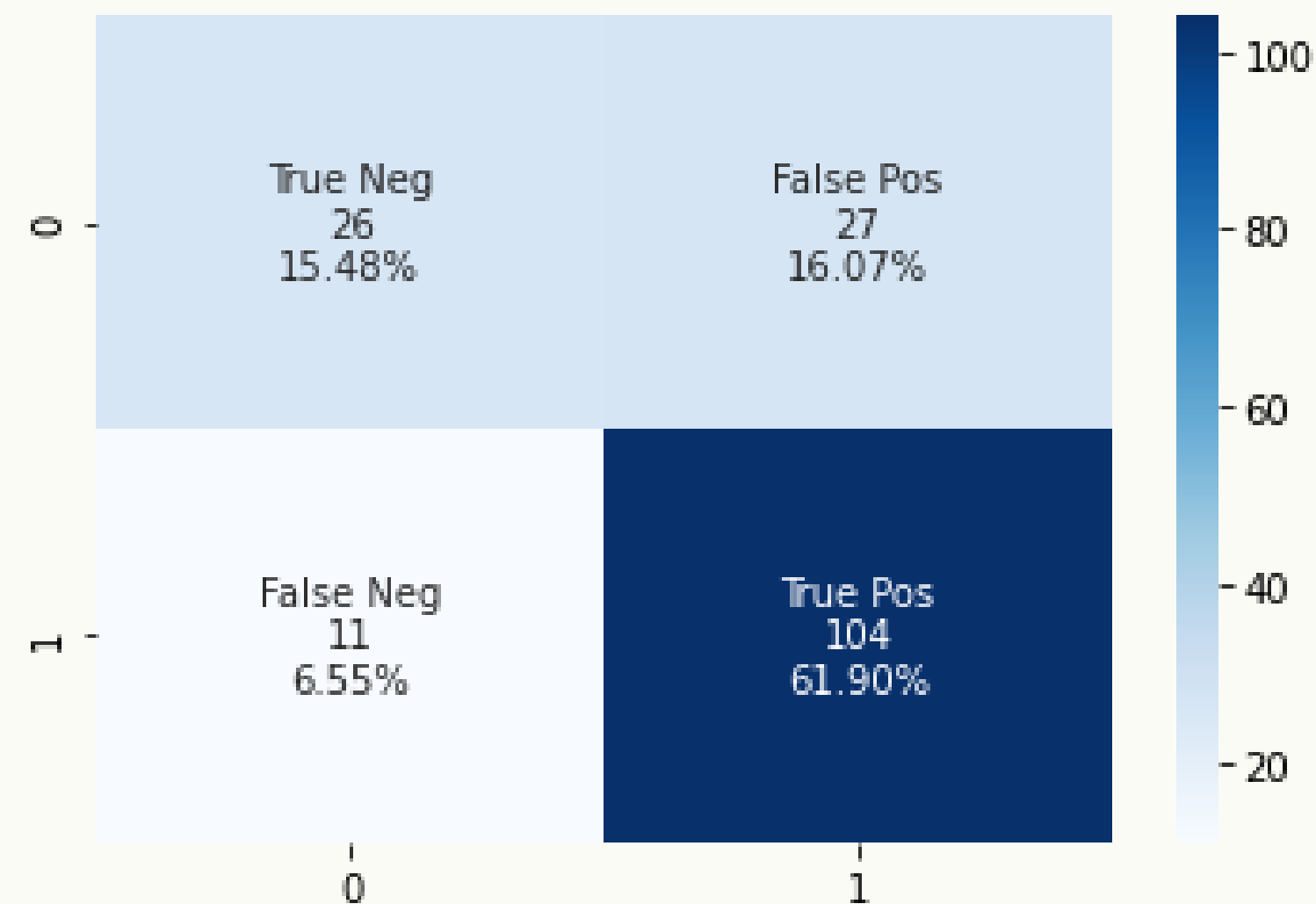


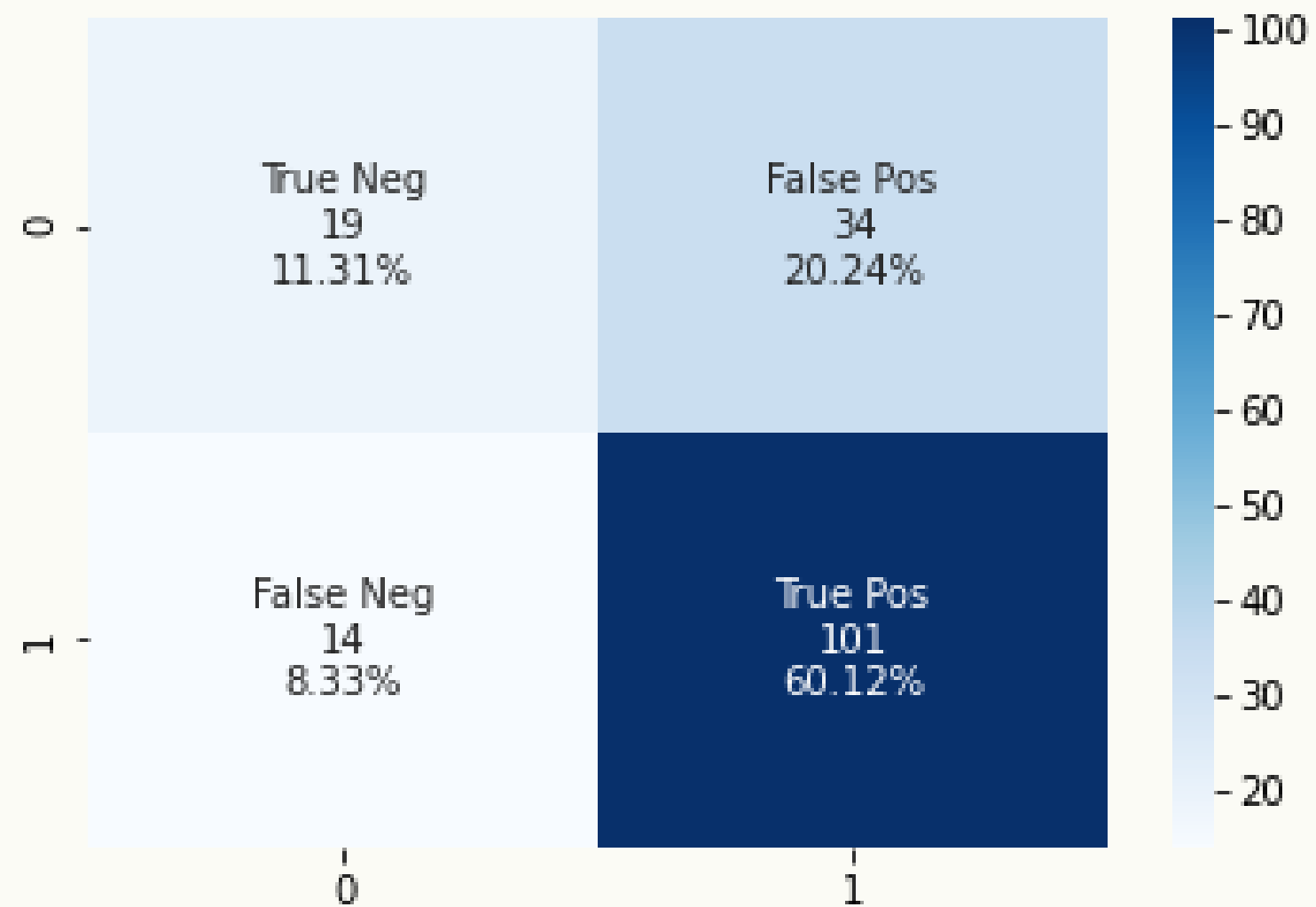
C=1.0, KERNEL=LINEAR

79.17%

C=100.0, KERNEL=LINEAR

77.38%



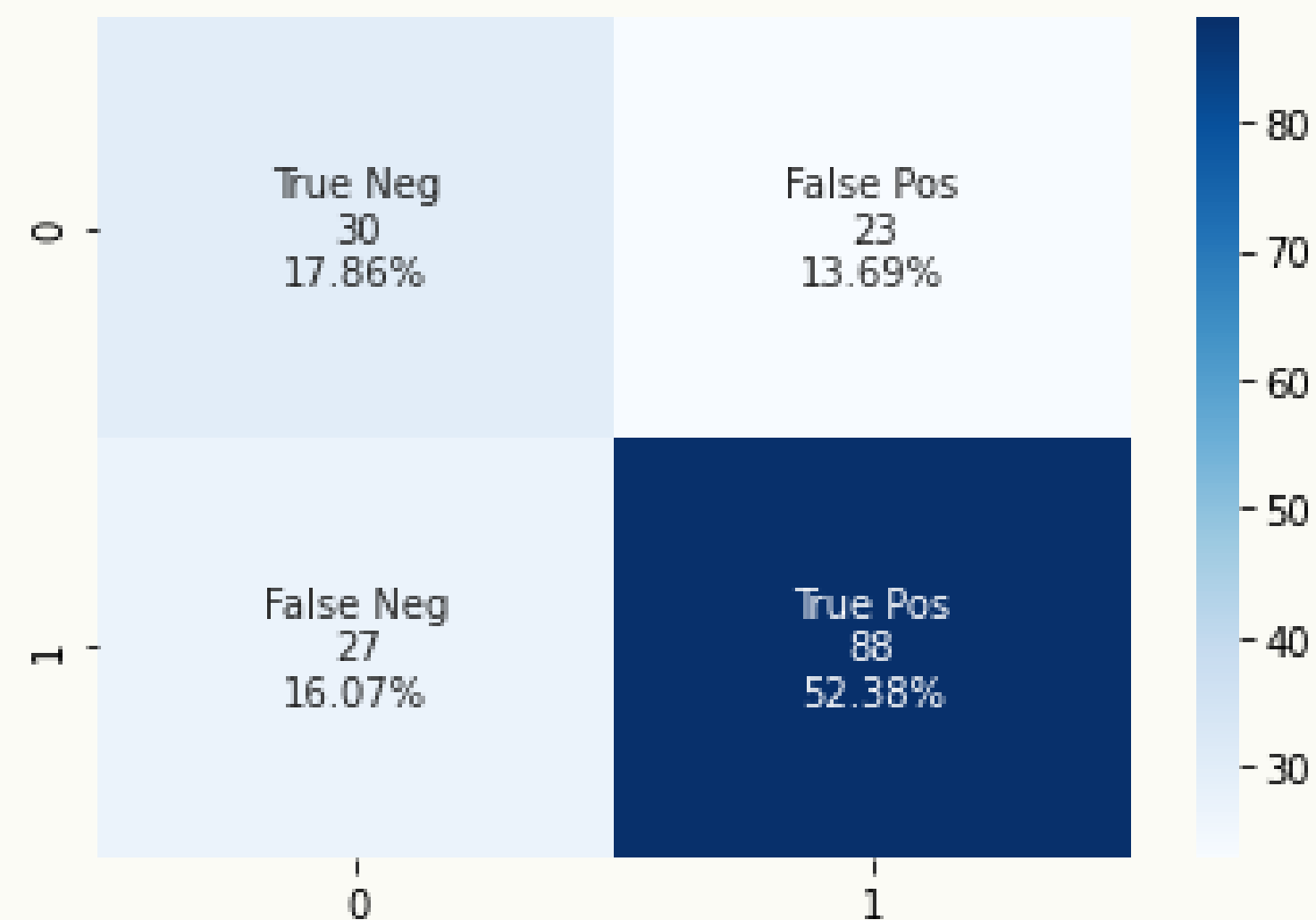


C=1.0, KERNEL=POLYNOMIAL

71.43%

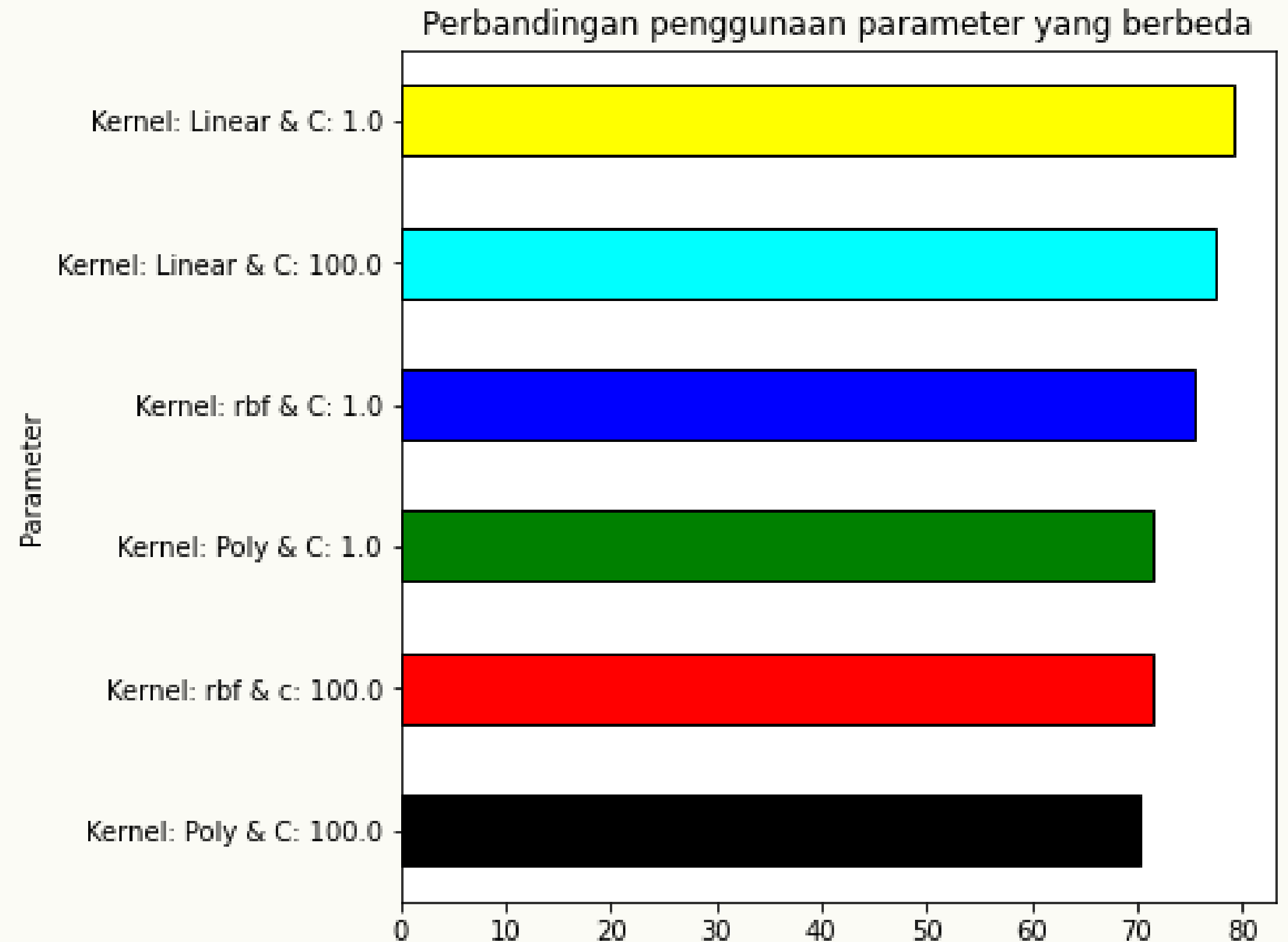
C=100.0, KERNEL=POLYNOMIAL

70.24%



KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan pada data startup maka didapatkan kesimpulan bahwa hasil klasifikasi dengan model klasifikasi Support Vector Machine (SVM) memperoleh nilai akurasi sebesar 79.1% dengan nilai hyperplane Kernel : Linear & C : 1.0



The background features a minimalist design with teal lines forming a large rectangular frame. A solid red rectangle is positioned in the top-left corner, and a solid teal rectangle is in the bottom-right corner.

THANK YOU!