

Identifikasi dan Prediksi Berbasis *Supervised Learning* Proses Pra-*flare* Kelas M dan X Menggunakan Data SHARP

Chandra A. Fernanda - 20322003

Pembimbing: Dr. Dhani Herdiwijaya

PROGRAM STUDI ASTRONOMI
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI BANDUNG
2023

Gambaran Umum

I. Pendahuluan

- a) Latar Belakang
- b) Tujuan dan Batasan
- c) Asumsi dan Rumusan Masalah

II. Hasil Pengolahan Data

- a) Diagram Alir Pengerjaan
- b) Pengurangan Jumlah Fitur dalam Model Pembelajaran Mesin
- c) Hasil Analisis *Principal Component Analysis* (PCA)
- d) Menjalankan Model Prediksi *Support Vector Machine*
- e) Hasil *Support Vector Machine*
- f) Perbandingan dengan Referensi

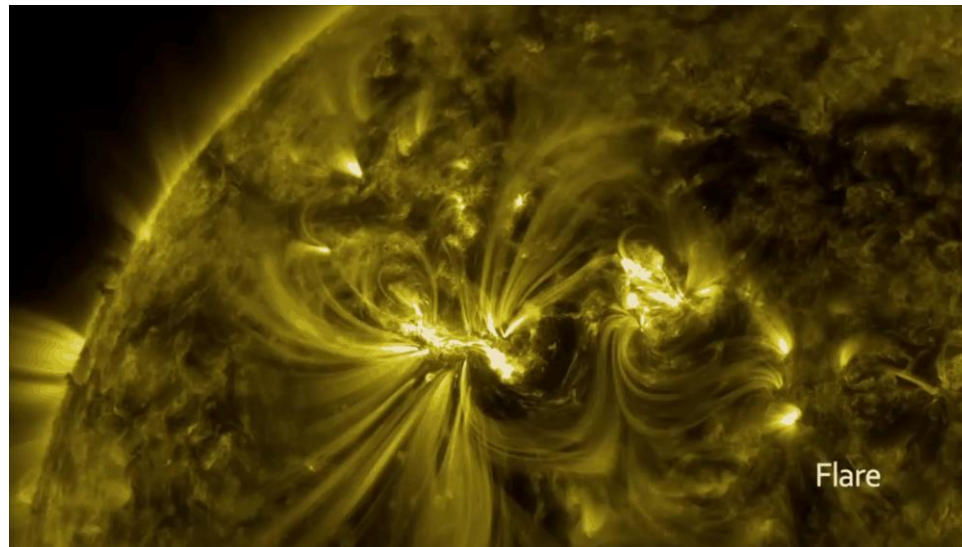
III. Penutup

- a) Simpulan
- b) Saran

Latar Belakang: Suar Surya

- Apa itu suar surya?

Suar surya atau *solar flare* adalah fenomena letusan energi dan gelombang elektromagnetik yang terjadi pada atmosfer Matahari.



Sumber: NASA Goddard / SDO

Latar Belakang: Suar Surya

- Teori pembentukan suar surya?

Suar surya terbentuk ketika terjadi rekoneksi magnetik di atas permukaan Matahari (Hudson & Ryan, 1995)



Sumber: NASA Goddard / Rhessi

Tujuan dan Batasan

- Tujuan
 - Menentukan peran masing-masing parameter magnetik terhadap kemunculan suar surya
 - Memprediksi kemunculan ledakan Matahari kelas X dan M menggunakan *supervised learning*
- Batasan
 - Rentang data (2 Mei 2010 00:00 UTC - 31 Desember 2021 23:48 UTC)
 - Bujur *active region* ($\pm 69^\circ$)
 - Prediksi kelas ledakan Matahari (X & M, non-flaring)

Asumsi dan Rumusan Masalah

- Suar surya didahului oleh proses penumpukan energi yang disebut sebagai proses *pra-flare*
 - Apakah terdapat suatu proses di parameter magnetik tertentu yang selalu mendahului terjadinya suar surya?
- Proses *pra-flare* terlihat sebagai fluktuasi atau peningkatan tiba-tiba pada parameter magnetik
 - Bagaimana kenampakan proses *pra-flare* pada setiap parameter magnetic?
- Bagaimana pengaruh pengurangan jumlah fitur berdasarkan proses *pra-flare* yang tampak pada parameter magnetic terhadap performa model prediksi?

Data

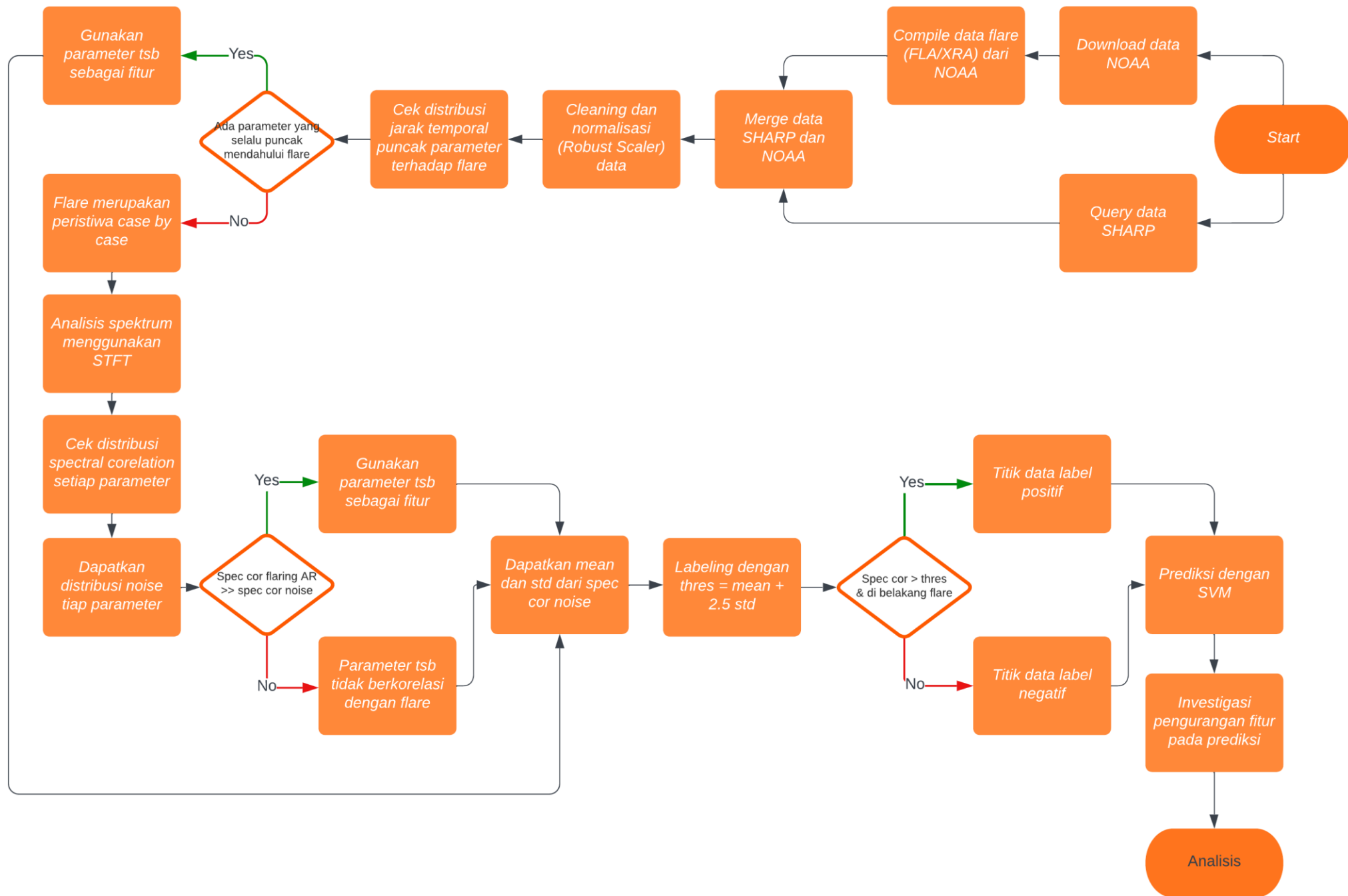
- Data yang digunakan merupakan data parameter magnetik SHARP yang telah dilengkapi dengan kejadian suar surya
- Rentang data: 2 Mei 2010 00:00 UTC - 31 Desember 2021 23:48 UTC (2.651.020 baris, 7912 AR)
- Distribusi kelas suar surya:

C	5464
B	4435
M	535
X	36
A	5

Tabel 1. 18 parameter magnetik SHARP (Bobra dan Couvidat, 2015)

Keyword	Description	Formula	F-Score
TOTUSJH	Total unsigned current helicity	$H_{c_{total}} \propto \sum B_z \cdot J_z $	3560
TOTPOT	Total photospheric magnetic free energy density	$\rho_{tot} \propto \sum (\vec{B}^{Obs} - \vec{B}^{Pot})^2 dA$	2996
TOTUSJZ	Total unsigned vertical current	$J_{z_{total}} = \sum J_z dA$	2733
ABSNJZH	Absolute value of the net current helicity	$H_{c_{abs}} \propto \sum B_z \cdot J_z $	2618
SAVNCPP	Sum of the modulus of the net current per polarity	$J_{z_{sum}} \propto \left \sum_{B_z^+} J_z dA \right + \left \sum_{B_z^-} J_z dA \right $	2448
USFLUX	Total unsigned flux	$\Phi = \sum B_z dA$	2437
AREA_ACR	Area of strong field pixels in the active region	$Area = \sum Pixels$	2047
MEANPOT	Mean photospheric magnetic free energy	$\bar{\rho} \propto \frac{1}{N} \sum (\vec{B}^{Obs} - \vec{B}^{Pot})^2$	1064
R_VALUE	Sum of flux near polarity inversion line	$\Phi = \sum B_{LOS} dA \text{ within R mask}$	1057
SHRGT45	Fraction of Area with Shear > 45°	$Area \text{ with Shear} > 45^\circ / \text{Total Area}$	740.8
MEANSHR	Mean shear angle	$\Gamma = \frac{1}{N} \sum \arccos \left(\frac{\vec{B}^{Obs} \cdot \vec{B}^{Pot}}{ \vec{B}^{Obs} \vec{B}^{Pot} } \right)$	727.9
MEANGAM	Mean angle of field from radial	$\bar{\gamma} = \frac{1}{N} \sum \arctan \left(\frac{B_h}{B_z} \right)$	573.3
MEANGBT	Mean gradient of total field	$ \nabla B_{tot} = \frac{1}{N} \sum \sqrt{\left(\frac{\partial B}{\partial x} \right)^2 + \left(\frac{\partial B}{\partial y} \right)^2}$	192.3
MEANGBZ	Mean gradient of vertical field	$ \nabla B_z = \frac{1}{N} \sum \sqrt{\left(\frac{\partial B_z}{\partial x} \right)^2 + \left(\frac{\partial B_z}{\partial y} \right)^2}$	88.40
MEANGBH	Mean gradient of horizontal field	$ \nabla B_h = \frac{1}{N} \sum \sqrt{\left(\frac{\partial B_h}{\partial x} \right)^2 + \left(\frac{\partial B_h}{\partial y} \right)^2}$	79.40
MEANJZH	Mean current helicity (B_z contribution)	$\overline{H_c} \propto \frac{1}{N} \sum B_z \cdot J_z$	46.73
MEANJZD	Mean vertical current density	$\overline{J_z} \propto \frac{1}{N} \sum \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} \right)$	17.44
MEANALP	Mean characteristic twist parameter, α	$\alpha_{total} \propto \frac{\sum J_z \cdot B_z}{\sum B_z^2}$	10.41

Diagram Alir Pengerjaan



Rumusan Masalah 3

Bagaimana pengaruh pengurangan jumlah fitur berdasarkan proses pre-flare yang tampak pada parameter magnetic terhadap performa model prediksi?

Mengapa Jumlah Fitur Perlu Diperhatikan

- *Curse of dimensionality*
- Kompleksitas fisis dan komputasi yang lebih rendah
- Efisiensi komputasi
- Noise dan informasi yang tidak relevan

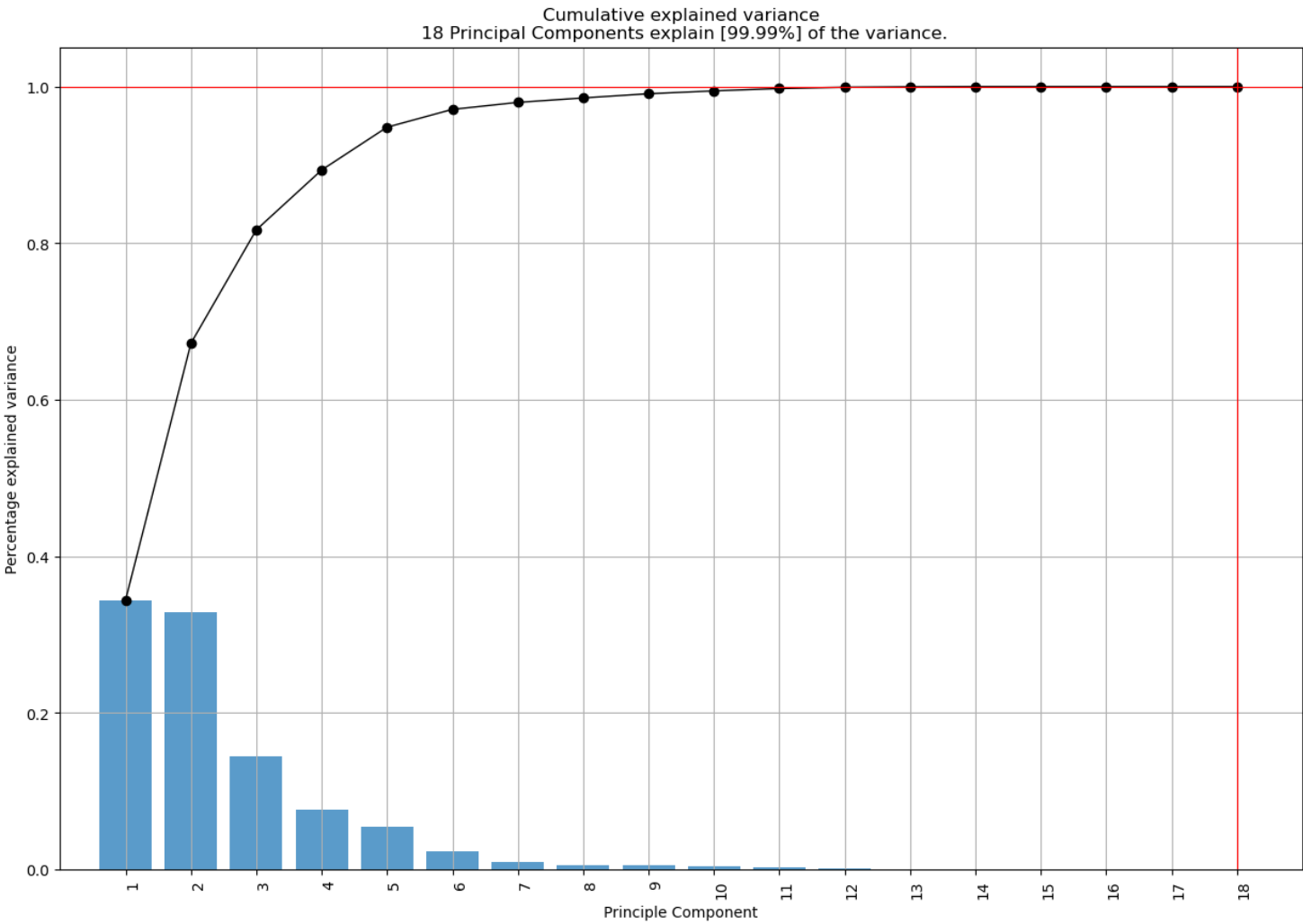
Peringkat Signifikansi Parameter Magnetik terhadap Proses Pra-*flare*

Tabel 2. Peringkat parameter magnetik terhadap korelasinya dengan proses pra-*flare*

Rank	Param	Avg Ratio	Keterangan
1	TOTPOT	77.92398	Total photospheric magnetic energy density
2	ABSNJZH	40.97702	Absolute value of the net current helicity
3	SAVNCPP	20.8962	Sum of the Absolute Value of the Net Currents Per Polarity
4	TOTUSJH	20.16014	Total unsigned current helicity
5	TOTUSJZ	18.59838	Total unsigned vertical current
6	USFLUX	16.96547	Total unsigned flux
7	AREA_ACR	15.36012	De-projected area of active pixels on sphere in micro-hemisphere
8	MEANPOT	9.914903	Mean photospheric excess magnetic energy density
9	SHRGT45	6.914851	Area with shear angle greater than 45 as a percent of total area
10	MEANSHR	3.1668	Mean shear angle for B_total
11	MEANGAM	2.752328	Mean inclination angle, gamma
12	R_VALUE	2.70343	Unsigned Flux R
13	MEANJZH	1.42975	Mean current helicity
14	MEANGBT	1.063462	Mean value of the total field gradient
15	MEANALP	1.002414	Mean twist parameter, alpha
16	MEANGBH	0.955158	Mean value of the horizontal field gradient
17	MEANGBZ	0.95026	Mean value of the vertical field gradient
18	MEANJZD	0.447325	Mean vertical current density

Principal Component Analysis (PCA)

Hasil PCA



Tabel 3. Parameter terbaik pada setiap PC

	PC	feature	loading	type
0	PC1	MEANPOT	-0.364623	best
1	PC2	TOTUSJZ	0.411280	best
2	PC3	MEANGBH	-0.598927	best
3	PC4	AREA_ACR	-0.485993	best
4	PC5	MEANALP	0.994229	best
5	PC6	R_VALUE	-0.840960	best
6	PC7	MEANPOT	0.519831	best
7	PC8	MEANGBH	-0.626952	best
8	PC9	MEANJZH	-0.625122	best
9	PC10	MEANPOT	0.608867	best
10	PC11	USFLUX	-0.757115	best
11	PC12	MEANSHR	0.642448	best
12	PC13	SHRGT45	0.701571	best
13	PC14	MEANGBT	-0.702179	best
14	PC15	MEANJZD	0.760710	best
15	PC16	SAVNCPP	0.564786	best
16	PC17	TOTUSJZ	-0.710970	best
17	PC18	ABSNJZH	0.707118	best
18	PC11	TOTPOT	0.449810	weak
19	PC18	TOTUSJH	-0.707093	weak
20	PC13	MEANGAM	-0.648179	weak
21	PC14	MEANGBZ	0.637066	weak

Tabel 4.
Skor variasi parameter

Feature	Score
MEANGAM	0.182140
MEANSHR	0.175842
SHRGT45	0.174708
MEANGBT	0.170748
MEANPOT	0.169986
MEANGBZ	0.169937
USFLUX	0.169555
TOTPOT	0.164761
AREA_ACR	0.157467
R_VALUE	0.141715
ABSNJZH	0.137829
TOTUSJH	0.137829
MEANJZD	0.137791
MEANJZH	0.137492
SAVNCPP	0.137305
TOTUSJZ	0.137261
MEANGBH	0.110936
MEANALP	0.073455

Mengapa PCA Belum Tentu Menggambarkan Relevansi

- Konsep dasar PCA itu sendiri
- Hasil yang berbeda dengan tabel 1 dan tabel 2

Menjalankan SVM

Pelabelan Data

- Data kemudian ditransformasi dengan STFT sehingga jumlah titik data menjadi 1.328.816 titik data
- Pelabelan dilakukan dengan meninjau distribusi jumlah korelasi spektrum pada 12 parameter magnetik teratas (Tabel 2)
- Diperoleh dataset dengan jumlah label positif sebanyak 53.725 (4,04 %) titik data dan negative sebanyak 1.275.091 (95,95 %)

12 Parameter yang Menunjukkan Proses Pra-*flare*

Tabel 5. 12 Parameter yang menunjukkan korelasi dengan pra-*flare*

Rank	Param	Avg Ratio
1	TOTPOT	77.92398
2	ABSNJZH	40.97702
3	SAVNCPP	20.8962
4	TOTUSJH	20.16014
5	TOTUSJZ	18.59838
6	USFLUX	16.96547
7	AREA_ACR	15.36012
8	MEANPOT	9.914903
9	SHRGT45	6.914851
10	MEANSHR	3.1668
11	MEANGAM	2.752328
12	R_VALUE	2.70343

Tabel 6. 12 parameter teratas pada referensi (Bobra dan Couvidat, 2015)

Keyword	Description	Formula	F-Score
TOTUSJH	Total unsigned current helicity	$H_{c_{total}} \propto \sum B_z \cdot J_z $	3560
TOTPOT	Total photospheric magnetic free energy density	$\rho_{tot} \propto \sum (\vec{B}^{Obs} - \vec{B}^{Pot})^2 dA$	2996
TOTUSJZ	Total unsigned vertical current	$J_{z_{total}} = \sum J_z dA$	2733
ABSNJZH	Absolute value of the net current helicity	$H_{c_{abs}} \propto \sum B_z \cdot J_z $	2618
SAVNCPP	Sum of the modulus of the net current per polarity	$J_{z_{sum}} \propto \left \sum^{B_z^+} J_z dA \right + \left \sum^{B_z^-} J_z dA \right $	2448
USFLUX	Total unsigned flux	$\Phi = \sum B_z dA$	2437
AREA_ACR	Area of strong field pixels in the active region	$Area = \sum Pixels$	2047
MEANPOT	Mean photospheric magnetic free energy	$\bar{\rho} \propto \frac{1}{N} \sum (\vec{B}^{Obs} - \vec{B}^{Pot})^2$	1064
R_VALUE	Sum of flux near polarity inversion line	$\Phi = \sum B_{LoS} dA \text{ within R mask}$	1057
SHRGT45	Fraction of Area with Shear > 45°	$Area \text{ with Shear} > 45^\circ / \text{Total Area}$	740.8
MEANSHR	Mean shear angle	$\bar{\Gamma} = \frac{1}{N} \sum \arccos \left(\frac{\vec{B}^{Obs} \cdot \vec{B}^{Pot}}{ \vec{B}^{Obs} \vec{B}^{Pot} } \right)$	727.9
MEANGAM	Mean angle of field from radial	$\bar{\gamma} = \frac{1}{N} \sum \arctan \left(\frac{B_h}{B_z} \right)$	573.3

Model SVM Sklearn

- Digunakan Pustaka Scikit-learn (Pedregosa, dkk. 2011)
- Liblinear (LinearSVC) vs libsvm (SVC)
- Liblinear:
 - Linier model sehingga hanya mendukung kernel linier
 - Cocok untuk data yang dapat dipisahkan secara linier
 - Waktu komputasi sebanding dengan jumlah data secara linier
- Libsvm:
 - Mendukung model linier dan non-linier
 - Dapat menemukan hubungan fitur-label yang lebih kompleks
 - Waktu komputasi eksponensial terhadap jumlah data
- Suar surya sendiri belum diketahui apakah dapat dipisahkan secara linier atau non-linier. Contoh penelitian yang menggunakan model linier (Leka dan Barnes, 2003) dan non-linier (Bobra dan Couvidat, 2015)

Model SVM sklearn: LinearSVC

```
class sklearn.svm.LinearSVC(penalty='l2', loss='squared_hinge', *, dual='warn', tol=0.0001, C=1.0, multi_class='ovr',  
fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0, random_state=None, max_iter=1000)
```

- C = hasil grid search, class_weight='balanced'
- Train-test split dilakukan dengan Stratified K-fold

```
class sklearn.model_selection.StratifiedKFold(n_splits=5, *, shuffle=False, random_state=None)
```

- n_splits=5
- Rasio test/train = 1/4

Hyperparameter Tuning: Grid Search

- Tuning dilakukan dengan GridSearchCV

```
class sklearn.model_selection.GridSearchCV(estimator, param_grid, *, scoring=None, n_jobs=None, refit=True, cv=None, verbose=0, pre_dispatch='2*n_jobs', error_score=nan, return_train_score=False)
```

- GridSearchCV dijalankan dua kali dengan
 - Param_grid: $C = \text{np.logspace}(-3, -3, 7)$
 - Param_grid: $C = \text{np.logspace}(-2, -1, 7)$
- Diperoleh nilai $C = 0.068$

Metrik Skor

$$akurasi = \frac{TP + TN}{N}$$

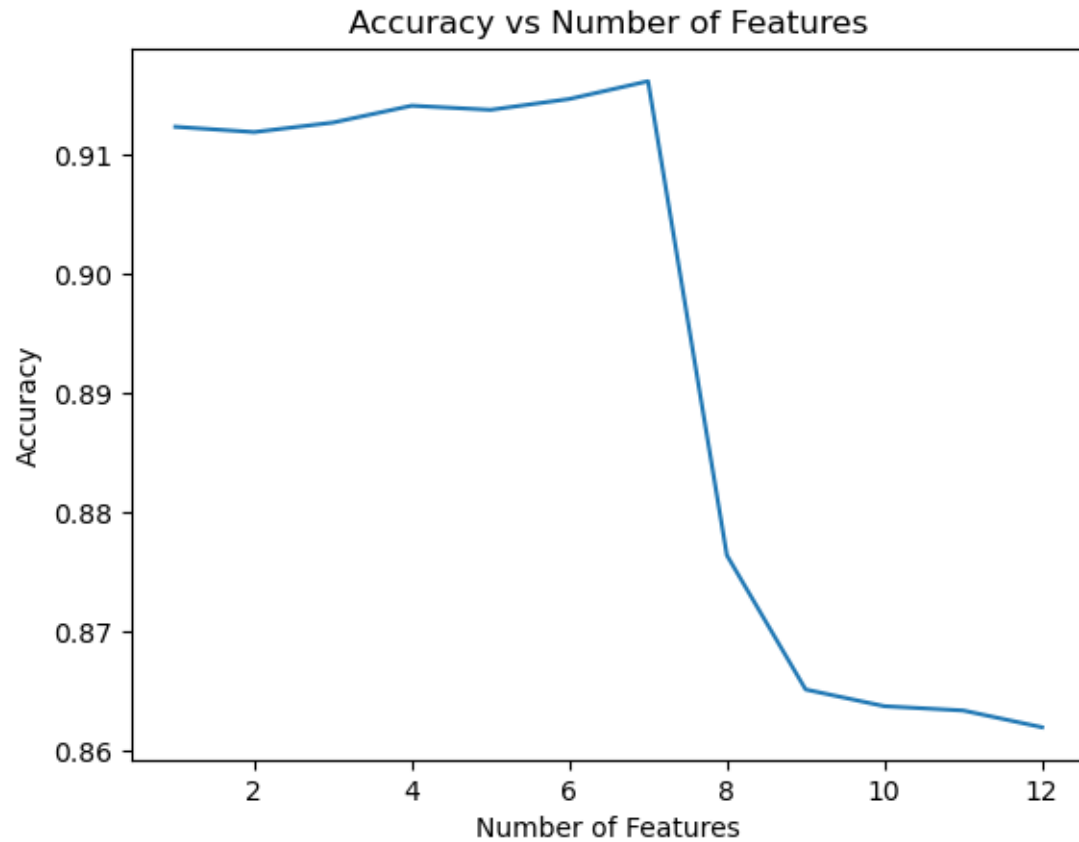
$$presisi = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

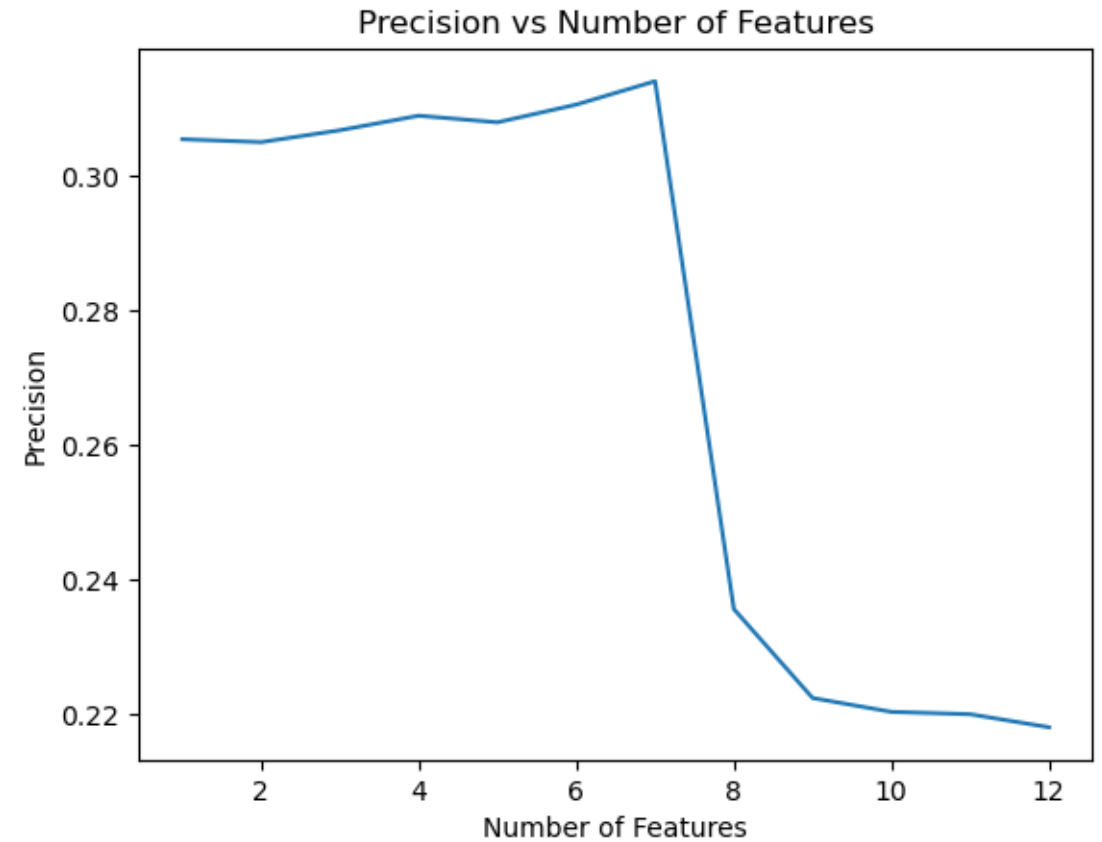
$$f1 = 2 \frac{presisi \cdot recall}{presisi + recall}$$

$$TSS = \frac{TP}{TP + FN} + \frac{FP}{TN + FP}$$

Hasil Kinerja terhadap Jumlah Fitur

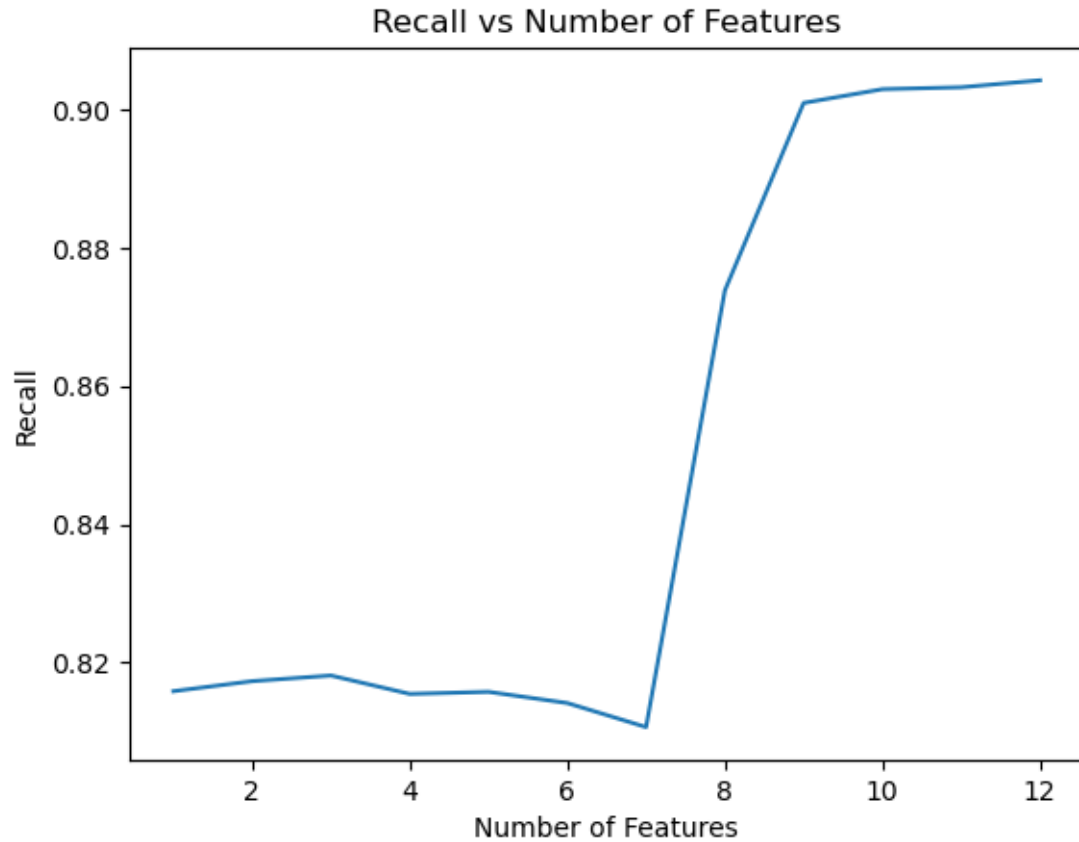


Kenaikan terhadap pengurangan fitur ~5,8%

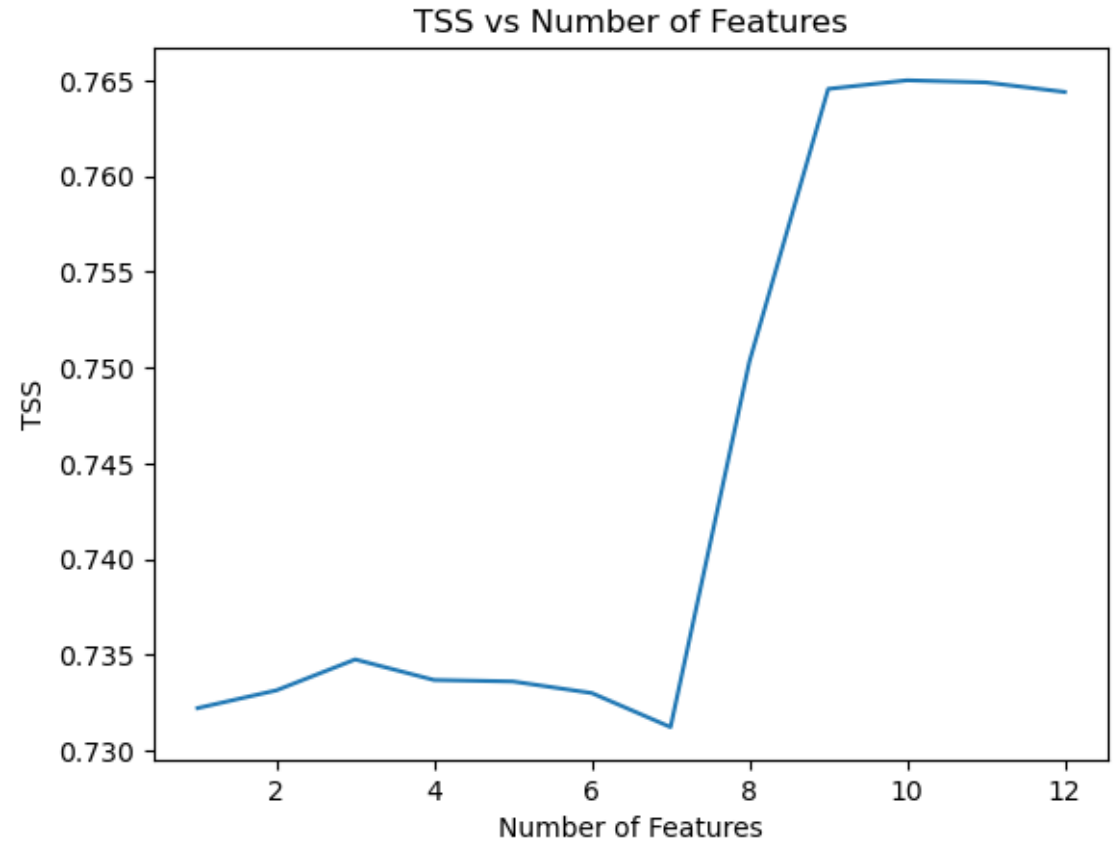


Kenaikan terhadap pengurangan fitur ~40%

Hasil Kinerja terhadap Jumlah Fitur

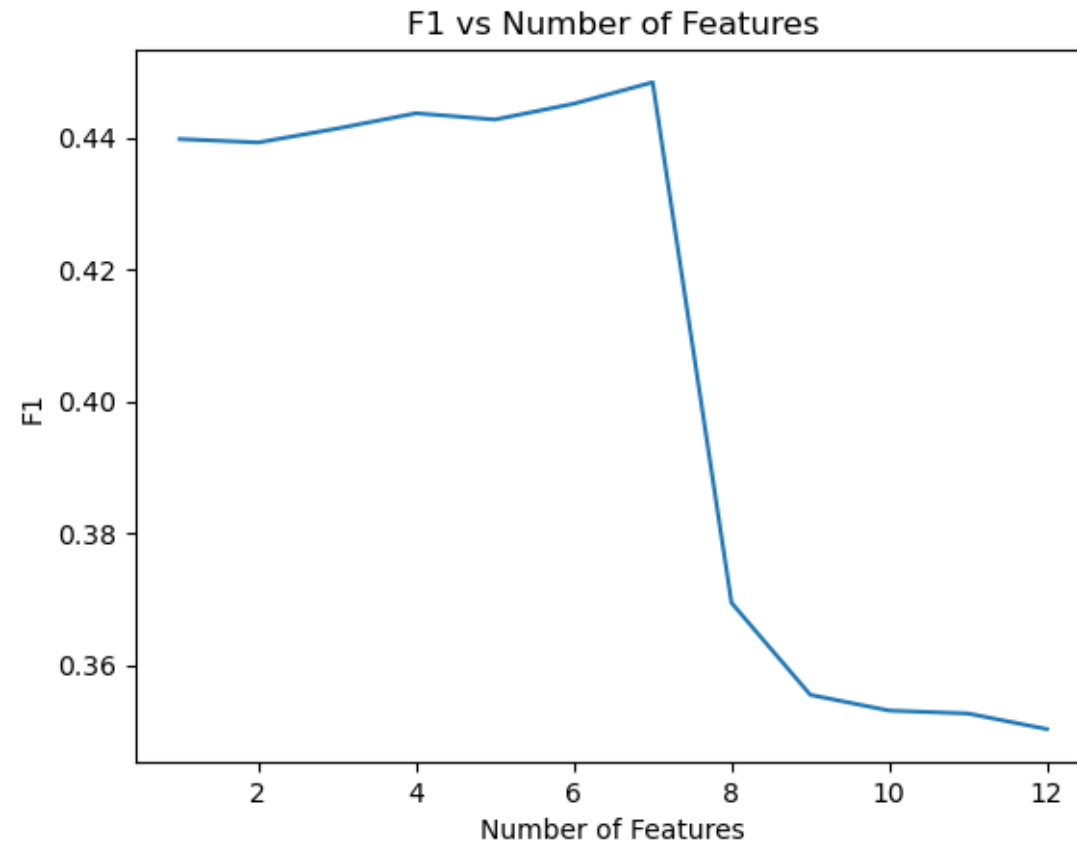


Penurunan terhadap pengurangan fitur ~8,8%



Penurunan terhadap pengurangan fitur ~4,5%

Hasil Kinerja terhadap Jumlah Fitur



Kenaikan terhadap pengurangan fitur ~25,7%

Perbandingan dengan Referensi

Tabel 7. Performa model dengan 7 fitur terhadap referensi.

	SVM 7 fitur	Pelabelan Dinamis (Fernanda, 2022)	Bobra dan Couvidat, 2015
Rasio negatif/positif	23,75	50,3813	16,5
Akurasi	0,91	0,85	0,96
Presisi	0,31	0,12	0,69
Recall	0,82	0,94	0,63
TSS	0,73	0,79	0,61

Simpulan

- Proses pra-flare merupakan proses yang unik pada setiap suar surya.
- Tidak semua parameter magnetik yang ditinjau dalam penelitian ini berkorelasi dengan kemunculan suar surya.
- Penjalanan *support vector machine* dengan pengurangan fitur menunjukkan bahwa jumlah fitur sama dengan 7 merupakan kombinasi fitur yang optimal (tabel 8). Model ini menghasilkan performa metrik pada tabel 9.

Tabel 8. 7 parameter yang digunakan dalam model SVM 7 fitur

Rank	Param	Avg Ratio	Keterangan
1	TOTPOT	77.92398	Total photospheric magnetic energy density
2	ABSNJZH	40.97702	Absolute value of the net current helicity
3	SAVNCPP	20.8962	Sum of the Absolute Value of the Net Currents Per Polarity
4	TOTUSJH	20.16014	Total unsigned current helicity
5	TOTUSJZ	18.59838	Total unsigned vertical current
6	USFLUX	16.96547	Total unsigned flux
7	AREA_ACR	15.36012	De-projected area of active pixels on sphere in micro-hemisphere

Tabel 9. Performa SVM 7 fitur

	SVM 7 fitur
Akurasi	0,91
Presisi	0,31
Recall	0,82
F1	0,44
TSS	0,73

Saran

- Menerapkan model normalisasi *power transform* dalam menormalisasi data
- Mengerucutkan kasus suar surya yang ditinjau
- Penerapan pengelompokkan terhadap data sebelum pelaksanaan PCA
- Memperluas rentang dan hyperparameter yang disesuaikan untuk model prediksi

Referensi

- Hudson, H. dan J. Ryan. 1995. High-Energy Particles in Solar Flares. *Annual Review of Astronomy and Astrophysics*. **33**. hlm. 239-282.
- Bobra, M. G., X. Sun, J. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, K. Leka. 2014. The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs - Space-Weather HMI Active Region Patches. *Solar Physics*. **289**. hlm. 3549-3578.
- Bobra, M. G. dan S. Couvidat. 2015. Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-Learning Algorithm. *The Astrophysical Journal*. **798**. hlm. 135.
- Pedregosa, F., dkk. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. **12**. hlm. 2825-2830.
- Fernanda, C. A. 2022. Prediksi Ledakan Matahari Kelas X dan M Menggunakan Supervised Learning. (Skripsi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung: Bandung).