Politeknik Statistika STIS

Implementasi *Cross-Soft Clustering* dan *Classification Modelling* untuk

Pengukuran Kinerja Keuangan

Usaha Mikro, Kecil, dan Menengah (UMKM) Berdasarkan Kategori Laba di Indonesia

Minggu, 15 Juni 2025

Tim StatStorm

Suhendra Widi Prayoga
KS - 222112382

Data bukan hanya tentang apa yang kita miliki, tapi tentang apa yang kita temukan di baliknya.

Muh Farhan
KS - 222112195

Data akan selalu meninggalkan pola berjejak

Dutatama Rosewika Taufiq Hadihardaya
KS - 222111997

Data science bukan hanya tentang membangun model, ini tentang membangun solusi.

Politeknik Statistika STIS

Pendahuluan

Latar Belakang

Peran Strategis UMKM dalam Perekonomian Nasional

- UMKM mencakup **99,9%** unit usaha di Indonesia (**64,2 juta unit** pada 2020).
- Berkontribusi **61,1%** terhadap **PDB** (**Rp8.573,89 triliun**) dan menyerap **97% tenaga kerja** nasional.
- Menjadi **tulang punggung ekonomi** inklusif, khususnya di wilayah terpencil.

Tantangan Struktural dan Digitalisasi UMKM

- **Keterbatasan** modal, legalitas usaha, inovasi, dan strategi pemasaran.
- **Rendahnya literasi digital** dan **keuangan** menghambat daya saing.
- **Belum** adanya **sistem** pemetaan **kebutuhan UMKM** secara presisi dan **berbasis data**.

Kelemahan Penelitian Sebelumnya

- Mayoritas **hanya fokus** pada segmentasi (***clustering***) tanpa analisis prediktif.
- **Minimnya integrasi** antara hasil **klaster** dengan model **prediksi** kinerja keuangan.
- Penelitian seperti oleh **Erkamim et al. (2023) unggul** dalam prediksi, namun **terpisah dari segmentasi**.

Kontribusi Penelitian terhadap SGDs

- Menghasilkan **pemodelan UMKM** yang lebih **komprehensif** dan **presisi**.
- Mendukung **kebijakan** berbasis data, selaras dengan **SDGs 8** dan **9**: pekerjaan layak, pertumbuhan ekonomi, dan inovasi industri.

Tujuan & Manfaat

Tujuan

01 Melakukan *Exploratory Data Analysis* (EDA) terhadap data UMKM untuk memperoleh pemahaman (*insight*) awal.

02 Melakukan analisis klaster dengan pendekatan *cross-cluster* pada data UMKM yang telah dikategorikan ke dalam dua kategori, yaitu struktur dan profil usaha, serta kinerja dan skala usaha.

03 Membangun model klasifikasi status laba (untung atau rugi) menggunakan algoritma *ensemble machine learning* berdasarkan hasil analisis *cross-cluster* yang telah dilakukan.

Manfaat

01 Membantu pelaku usaha memahami posisi mereka dalam klaster, mengidentifikasi faktor penentu kinerja, dan mengambil keputusan strategis berbasis data.

02 Memberikan landasan bagi pemerintah dalam merancang kebijakan yang tepat sasaran, serta membantu lembaga keuangan dan investor menilai kelayakan usaha secara lebih terukur dan objektif.

03 Menyediakan pendekatan metodologis baru yang dapat dimanfaatkan akademisi, serta mendorong pertumbuhan ekonomi lokal dan inklusivitas sistem ekonomi untuk masyarakat luas.

Penelitian Terkait

Santi Styaningsih et. al. (2012)

Melakukan analisis klaster untuk mengevaluasi kinerja UMKM di Indonesia berdasarkan pola pertumbuhan dan strategi yang diterapkan, sehingga diperoleh wawasan tentang segementasi UMKM yang dapat digunakan untuk merancang kebijakan yang lebih efektif.

Sifriyani et. al. (2023)

Menggabungkan teknik *clustering* dan *classification* untuk memprediksi kinerja UMKM. Pendekatan ini memungkinkan identifikasi pola tersembunyi yang dapat digunakan untuk memprediksi keberhasilan atau kegagalan usaha.

Arifin et. al. (2019)

Menerapkan algoritma K-Means untuk mengelompokkan UMKM di Kabuaten Garut berdasarkan atribut seperti kepemilikan NPWP dan izin usaha. Teknik ini berhasil memberikan gambaran kepada pemerintah dalam merumuskan kebijakan untuk pengembangan UMKM.

Politeknik Statistika STIS

Metode Penelitian

Metode Penelitian

Eksplorasi dan Manipulasi Data

- Identifikasi Kategori dan Tipe Data
- Identifikasi Distribusi Data
- Identifikasi *Missing Value*
- Identifikasi Outlier
- Label Encoding
- Imputasi Data
- Standardisasi Data

Soft Clustering

- Pengecekan Multikolinieritas
- Pengujian KMO
- Pemilihan Jumlah Klaster Optimum
- Clustering
- Evaluasi Klaster
- Pemisahan Data Menurut Klaster
- Analisis Profil dan Interpretasi

Classification Modelling

- Tuning Parameter
- Pemodelan Klasifikasi
- Evaluasi Model
- Interpretasi

Politeknik Statistika STIS

Hasil dan Pembahasan

Identifikasi Kategori & Tipe Data

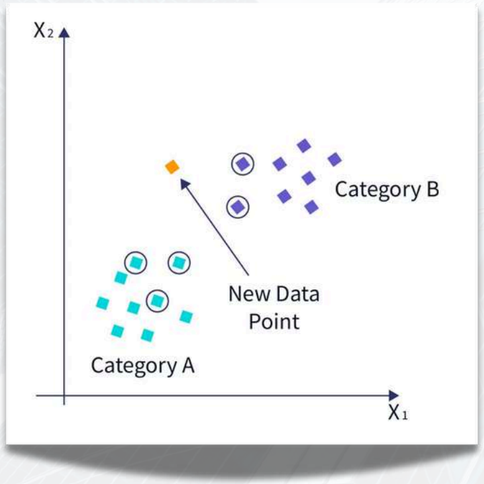
Informasi Fitur yang Digunakan dalam Pemodelan

Atribut	Tipe Data	Persentase Missing Value	Kategori Variabel
jenis_usaha	Kategorik	0,97%	Struktur dan Profil Usaha
tenaga_kerja_perempuan	Numerik	1,15%	Struktur dan Profil Usaha
tenaga_kerja_laki_laki	Numerik	1,00%	Struktur dan Profil Usaha
aset	Numerik	1,05%	Kinerja dan Skala Usaha
omset	Numerik	1,04%	Kinerja dan Skala Usaha
marketplace	Kategorik	1,13%	Struktur dan Profil Usaha
kapasitas_produksi	Numerik	1,21%	Struktur dan Profil Usaha

Atribut	Tipe Data	Persentase Missing Value	Kategori Variabel
status_legalitas	Kategorik	0,96%	Struktur dan Profil Usaha
tahun_berdiri	Numerik	0,94%	Struktur dan Profil Usaha
laba	Numerik	1,16%	None (Target)
biaya_karyawan	Numerik	0,99%	Kinerja dan Skala Usaha
jumlah_pelanggan	Numerik	1,12%	Kinerja dan Skala Usaha
klas_laba	Kategorik	1,16%	None (Target New)

Catatan: Atribut klas_laba diperoleh dari proses *feature engineering* berupa pengkategorian ulang atribut laba dengan ketentuan: untung (laba positif) dan rugi (laba negatif).

Imputasi *Missing Value*, *Label Encoding*, dan Standardisasi



K-NN imputer digunakan dalam proses imputasi *missing value* karena data yang dimiliki berisi **kombinasi fitur numerik dan kategorik.**

Original Data:

Color	Size	Price
Blue	L	100
Green	M	150
Red	S	200
Green	XL	120
Red	M	180

Label Encoding

Label Encoded Data:

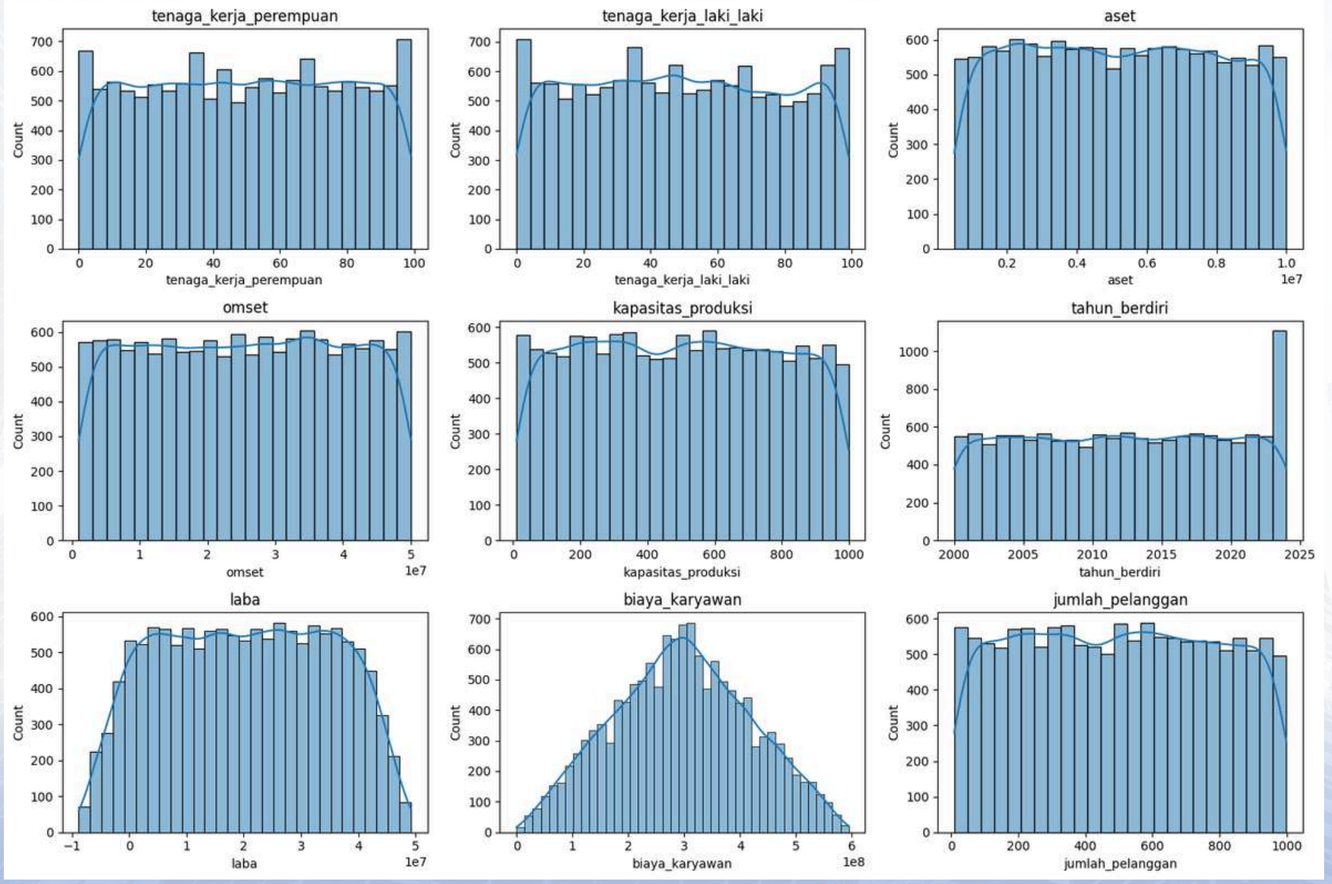
Color	Size	Price
0	0	100
1	1	150
2	2	200
1	3	120
2	1	180

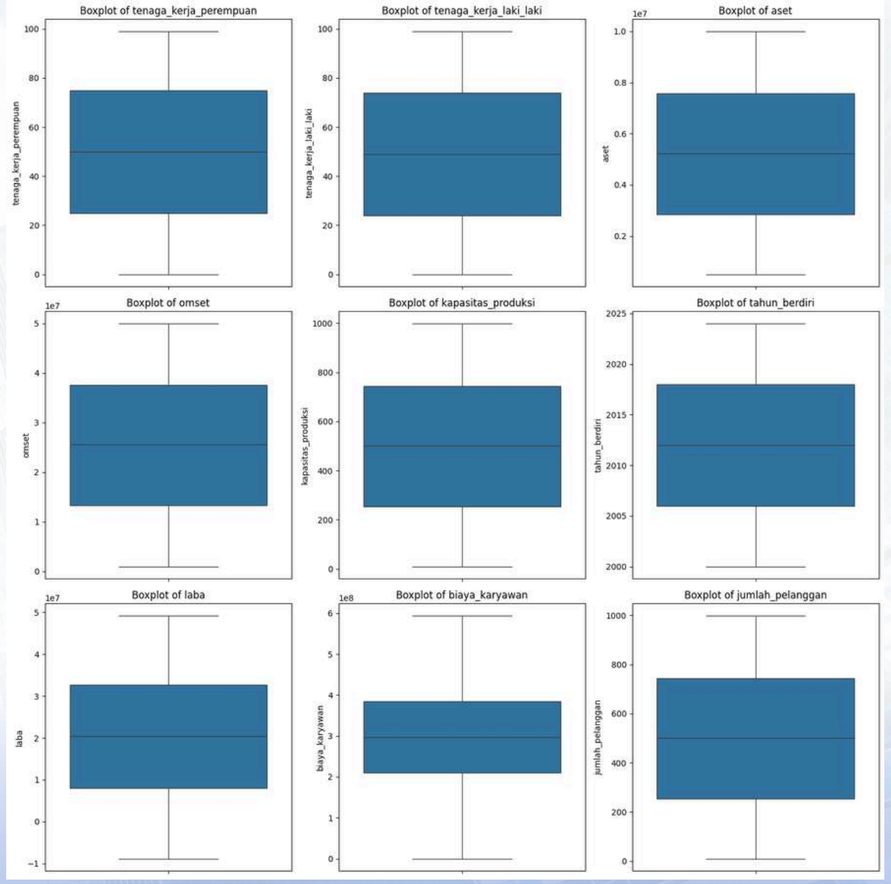
Label encoder digunakan dalam proses *encoding* variabel kategorikal karena data yang dimiliki **tidak mengandung jumlah kategori** yang terlalu banyak dan **sesuai untuk dijadikan input pada model berbasis *tree*** yang digunakan pada penelitian ini.

$$Z = \frac{X - \mu}{\sigma}$$

Standardisasi dilakukan untuk **mengurangi dominasi fitur** yang memiliki nilai tinggi terhadap fitur bernilai lebih rendah. Secara khusus metode standardisasi digunakan karena **dapat diterapkan untuk berbagai jenis distribusi data**.

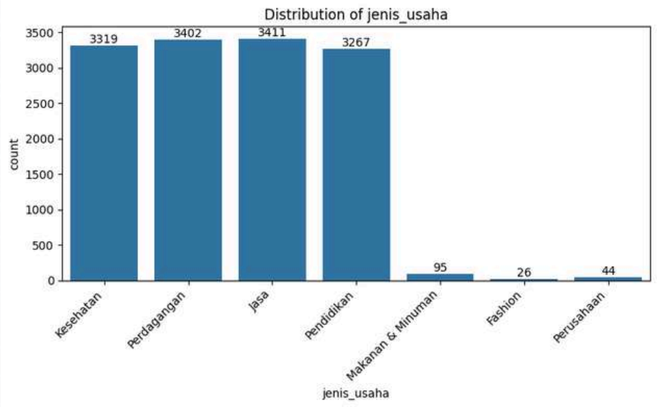
Distribusi Atribut Numerik

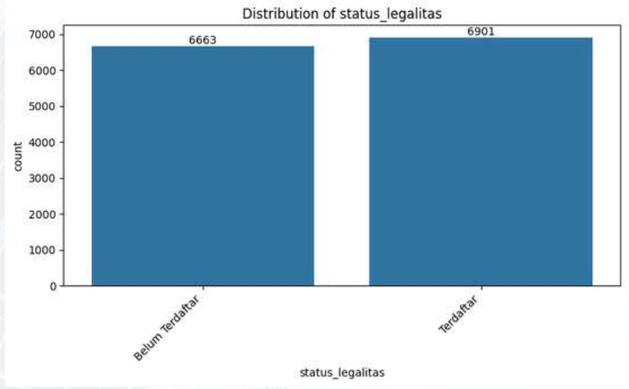


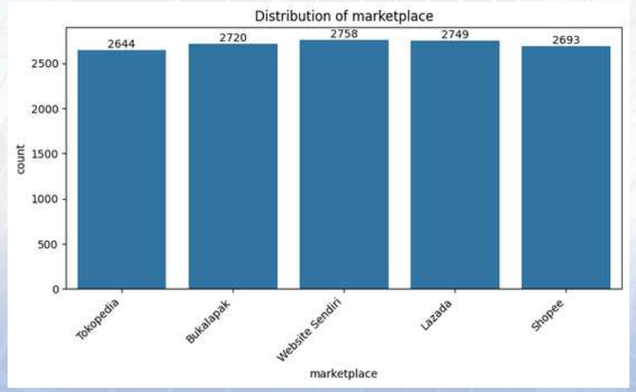


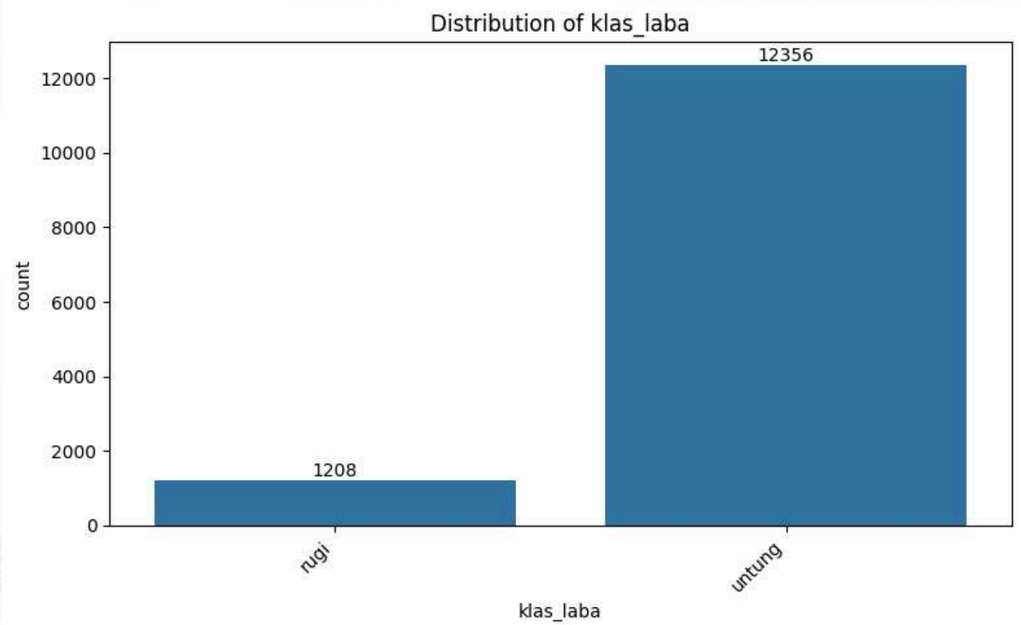
Pada kedua visualisasi di atas, data berdistribusi normal dan terlihat tidak ada *outlier* pada atribut numerik.

Distribusi Atribut Kategorik





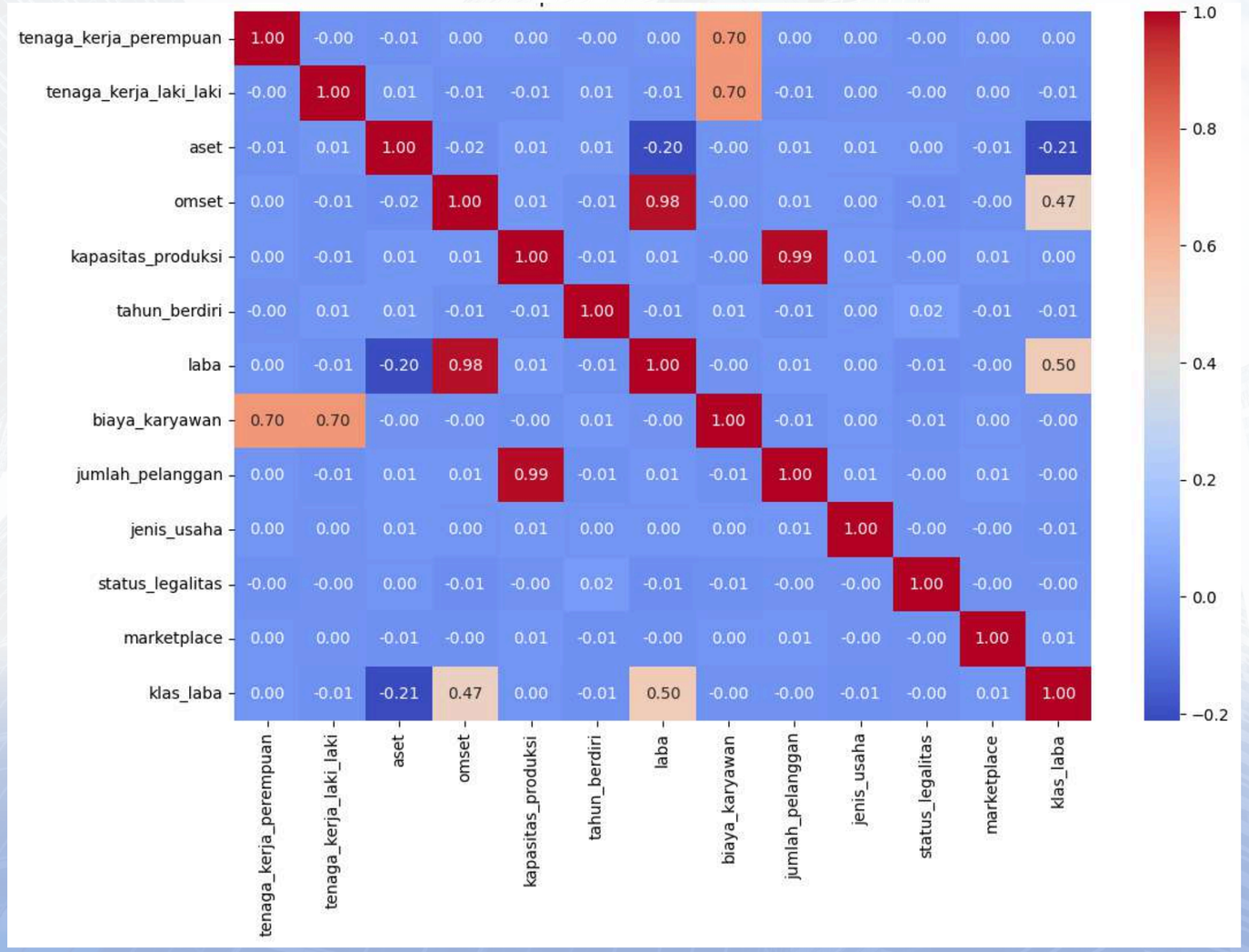




Hanya pada atribut jenis_usaha yang terlihat ketimpangan distribusi label yang sangat ekstrem.

Mayoritas unit usaha dalam data berada pada kondisi “untung”, dengan jumlah mencapai 12.356 UMKM, sedangkan hanya sebagian kecil yang berada pada kondisi “rugi”, yaitu sebanyak 1.208 UMKM.

Heatmap Korelasi



Terlihat korelasi sangat kuat antara omset dan laba (0,98) serta kapasitas produksi dan jumlah pelanggan (0,99), yang logis karena terkait langsung dengan pendapatan. Korelasi positif juga tampak antara biaya karyawan dan jumlah tenaga kerja (0,70), karena semakin banyak karyawan, semakin tinggi biayanya.

Cross-Soft Clustering

Clustering dilakukan untuk **mengelompokkan UMKM** berdasarkan **dua kategori karakteristik** usaha.

Struktur dan Profil Usaha

Fitur-fitur yang menggambarkan informasi tentang sifat dasar usaha dan sumber daya yang dimiliki.

Kinerja dan Skala Usaha

Fitur-fitur yang menggambarkan performa usaha secara finansial dan operasional.

Algoritma *clustering* yang digunakan adalah algoritma yang dapat **menangani data campuran** (kategorik maupun numerik).

K-Prototype

Adaptive K-Means

K-Medoids (Gower)

Agglomerative (Gower)

Pengecekan Multikolinieritas & Pengujian KMO

Struktur dan Profil Usaha

Variabel	VIF
tenaga_kerja_perempuan	1.000035
tenaga_kerja_laki_laki	1.000135
kapasitas_produksi	1.000159
tahun_berdiri	1.000126

Nilai KMO = 0.505

Kinerja dan Skala Usaha

Variabel	VIF
aset	1.000286
omset	1.000297
biaya_karyawan	1.000047
jumlah_pelanggan	1.000143

Nilai KMO = 0.497

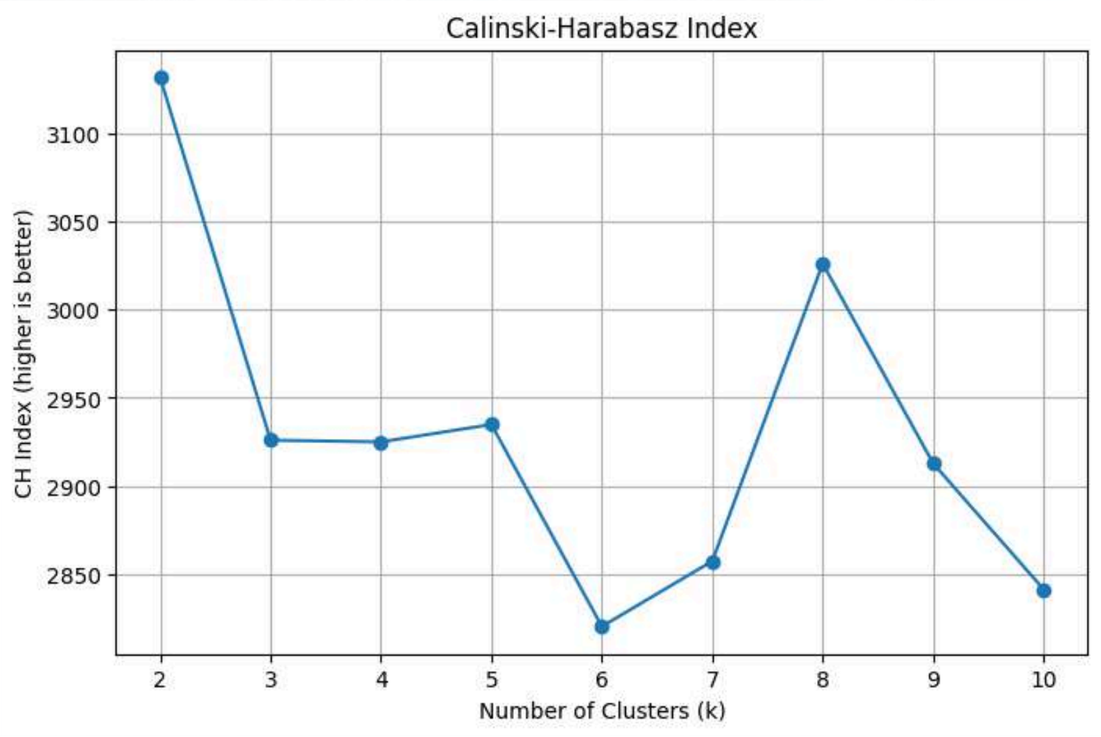
Tidak ada nilai VIF > 5 di kedua kategori. Hal ini menunjukkan bahwa **tidak ada multikolinieritas** di antara variabel-variabel yang akan digunakan dalam pembentukan klaster.

Nilai KMO < 0.6 di kedua kategori mengindikasikan bahwa **tidak terdapat struktur hubungan yang kuat antar variabel** di kedua kategori sehingga **tidak perlu** dilakukan **reduksi variabel**.

Pemilihan Jumlah Klaster Optimum

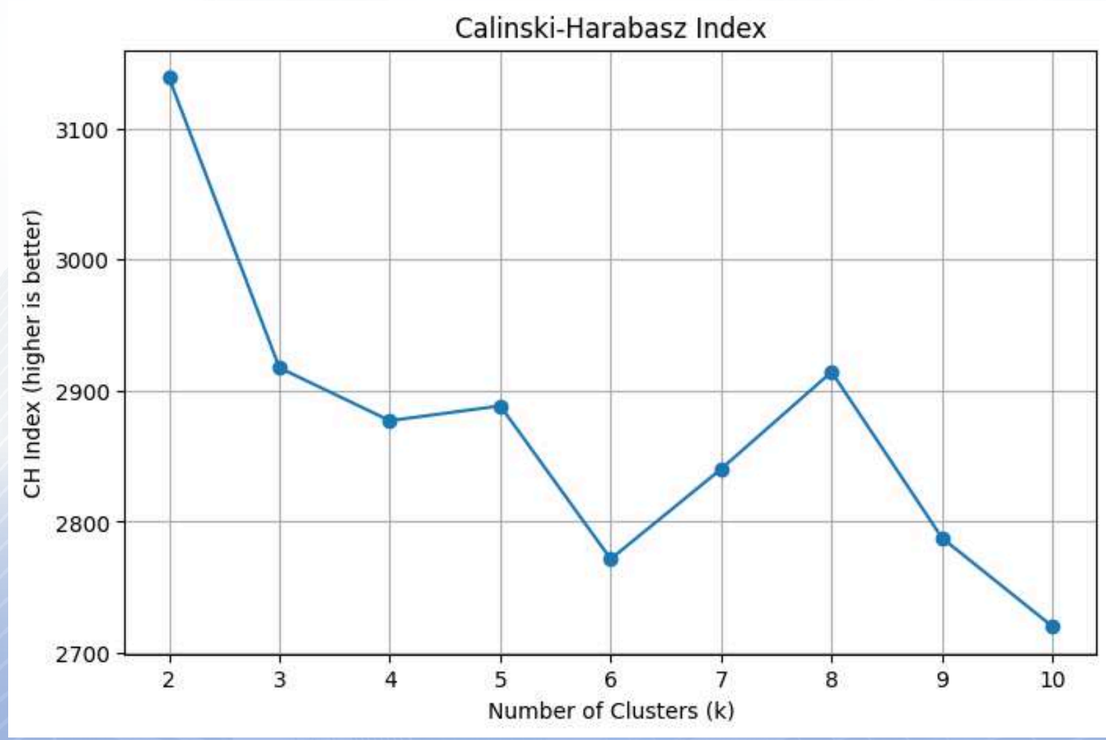
Pemilihan jumlah klaster optimum dilakukan dengan pendekatan **Calinski-Harabasz Index**

Struktur dan Profil Usaha



Jumlah klaster optimum adalah 2 **klaster**.

Kinerja dan Skala Usaha



Jumlah klaster optimum adalah **2 klaster**.

Evaluasi Klaster

Hasil klasterisasi menunjukkan bahwa algoritma ***Adaptive K-Means*** menghasilkan **performa terbaik** di kedua kategori karakteristik.

Struktur dan Profil Usaha

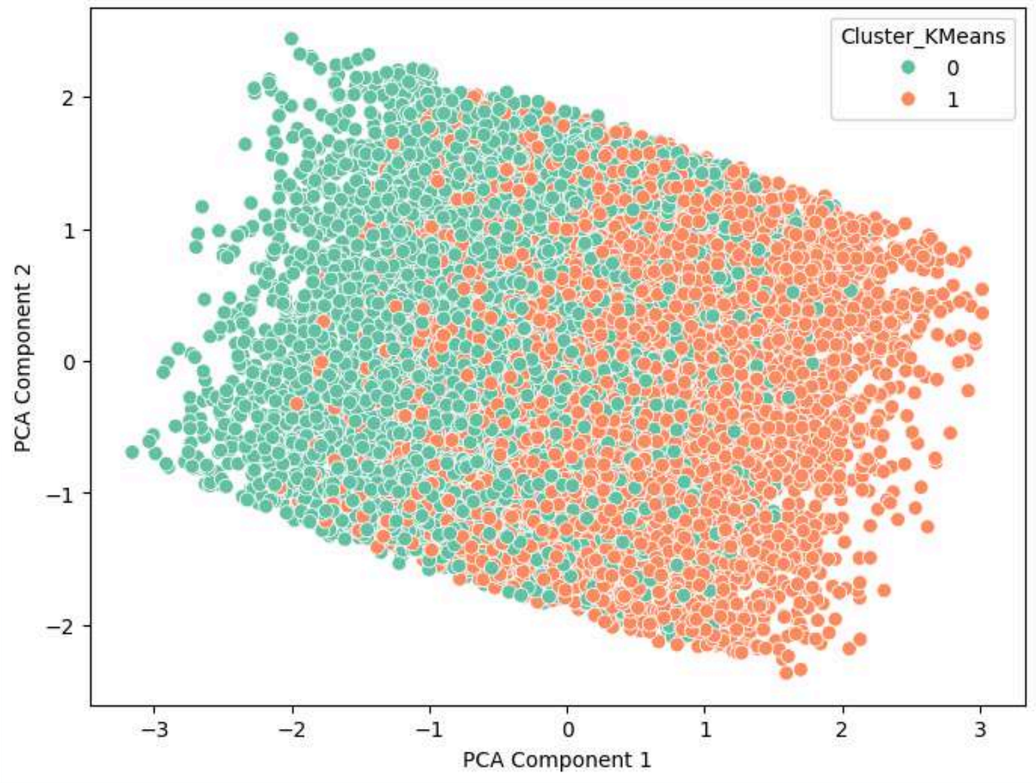
Algoritma	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
K-Prototype	0.108071	5.554048	413.828015
Adaptive K-Means	0.095482	**2.522424**	**2074.293822**
K-Medoids	0.083953	4.847865	560.049517
Agglomerative	**0.250816**	4.633853	612.818043

Kinerja dan Skala Usaha

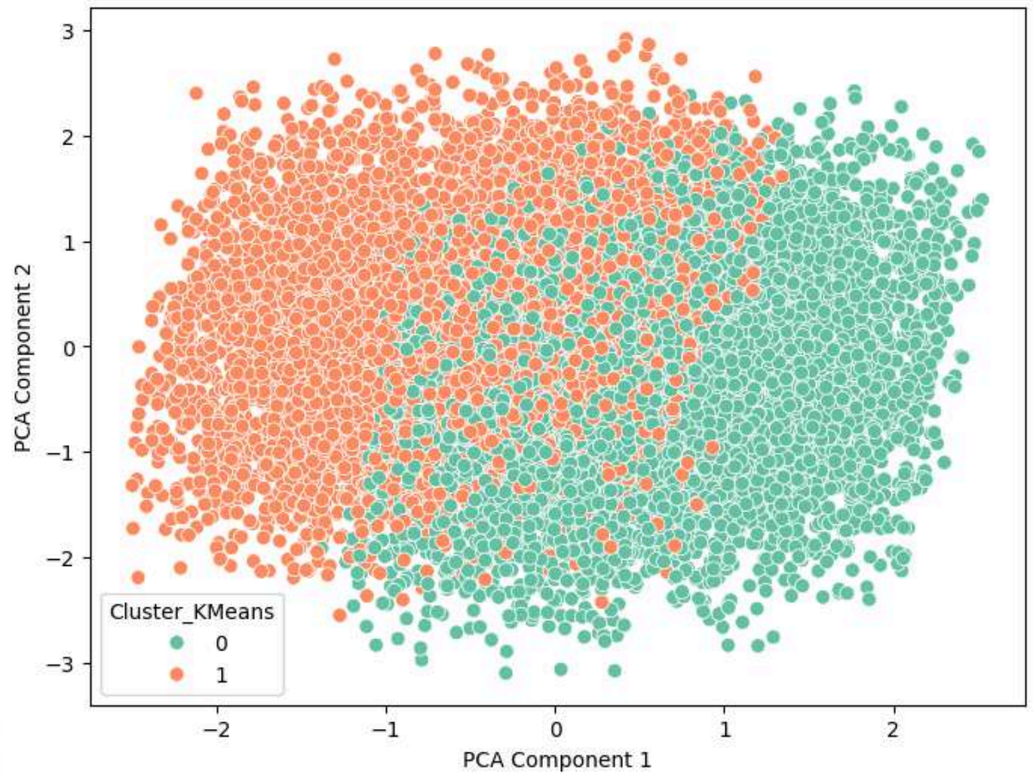
Algoritma	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
Adaptive K-Means	**0.231137**	1.996138	**3139.114491**
K-Medoids	0.227754	2.005573	3110.101286
Agglomerative	0.228071	**1.702222**	14.382511

Visualisasi Klaster

Struktur dan Profil Usaha



Kinerja dan Skala Usaha



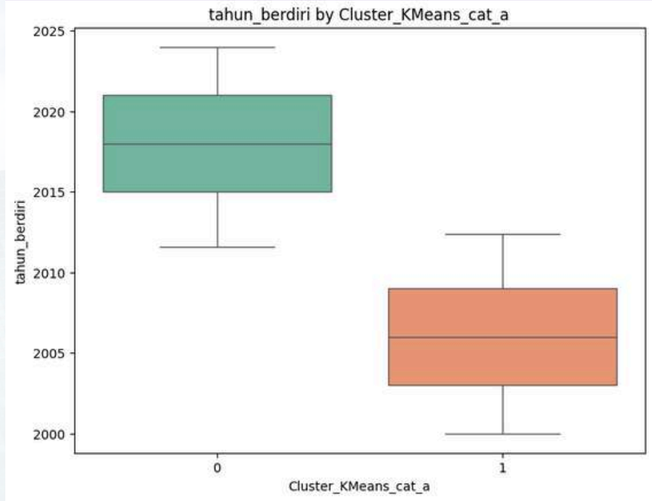
Analisis Profil

Analisis profil dilakukan pada **variabel-variabel** yang memiliki karakteristik yang **berbeda signifikan** di masing-masing klaster.

Struktur dan Profil Usaha

Variabel	Rata-rata		Uji Beda Rata-rata	
	Klaster 0	Klaster 1	t-stat	p-value
Tenaga Kerja Perempuan	49.77	49.65	-0.36	0.72
Tenaga Kerja Laki-Laki	49.58	48.70	2.49	0.01
Kapasitas Produksi	487.78	511.83	-4.46	0.00
Tahun Berdiri	**2018.28**	**2005.76**	**200.94**	**0.00**

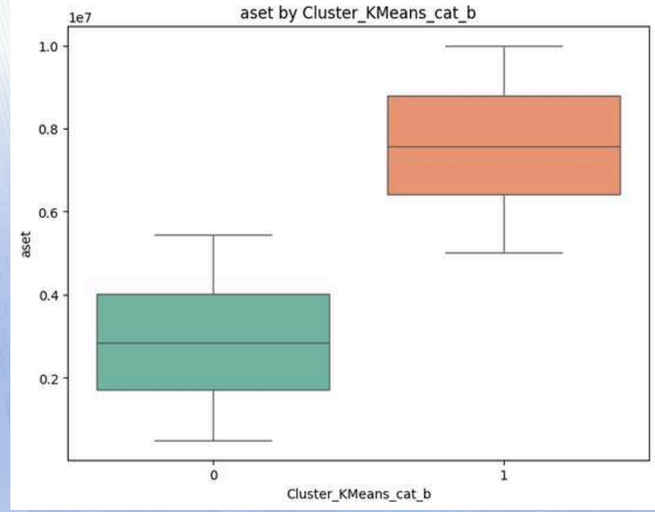
Tahun Berdiri



Kinerja dan Skala Usaha

Variabel	Rata-rata		Uji Beda Rata-rata	
	Klaster 0	Klaster 1	t-stat	p-value
Aset	**2861281.16**	**7594176.00**	**-201.96**	**0.00**
Omset	25789210.27	25321329.6	1.95	0.05
Biaya Karyawan	294942819.3	297837056.7	-1.39	0.17
Jumlah Pelanggan	491.26	508.49	-3.54	0.00

Aset



Hasil Klaster

Kinerja dan Skala Usaha

Struktur dan Profil Usaha

	Klaster 0 (50,10%)	Klaster 1 (49,90%)
Klaster 0 (50,02%)	**Klaster 00:** Usaha dengan **umur muda** dan total **aset rendah**. **(3396 UMKM)**	**Klaster 01:** Usaha dengan **umur muda** dan total **aset tinggi**. **(3389 UMKM)**
Klaster 1 (49,98%)	**Klaster 10:** Usaha dengan **umur tua** dan total **aset rendah**. **(3400 UMKM)**	**Klaster 11:** Usaha dengan **umur tua** dan total **aset tinggi**. **(3379 UMKM)**

Predictive Modelling (Classification)

Penelitian ini juga melakukan pemodelan **klasifikasi** menggunakan ***machine learning*** untuk memprediksi **keberhasilan usaha UMKM** yang ditentukan berdasarkan nilai laba. Dalam hal ini UMKM dengan nilai laba > 0 dikategorikan sebagai UMKM untung sendangkan UMKM dengan nilai laba ≤ 0 dikategorikan sebagai UMKM rugi.

XGBoost — **Random Forest** — **CatBoost** — **LightGBM**

OPTUNA OPTIMIZATION

Jumlah percobaan (*n trials*) sebanyak 50

***Cluster* 00 (LightGBM)**	n_estimators: 200, learning_rate: 0.00842, max_depth: 3, num_leaves:52, subsample: 0.79098, colsample_bytree: 0,70270, reg_alpha: 0.00109, reg_lambda: 0.01814
***Cluster* 01 (CatBoost)**	iterations: 300, learning_rate: 0.00389, depth: 5, subsample: 0.74001, l2_leaf_reg: 1.45730, border_count: 170
***Cluster* 10 (CatBoost)**	iterations: 300, learning_rate: 0.00389, depth: 5, subsample: 0.74001, l2_leaf_reg: 1.45730, border_count: 170
***Cluster* 11 (CatBoost)**	iterations: 300, learning_rate: 0.00389, depth: 5, subsample: 0.74001, l2_leaf_reg: 1.45730, border_count: 170

Evaluasi Model pada Cluster 00

Hasil pemodelan klasifikasi pada *cluster* 00 menunjukkan bahwa model ***Light Gradient Boosting Machine*** **(LightGBM)** menghasilkan **performa terbaik** dari model lainnya. Berikut tabel perbandingan evaluasi model pada penelitian ini:

Metode	AUC	Balanced Acuracy
XGBoost	0.7320	0.7279
CatBoost	0.7272	0.6948
LightGBM	**0.7438**	**0.6661**
Random Forest	0.6663	0.5000

Evaluasi Model pada Cluster 01

Hasil pemodelan klasifikasi pada *cluster* 01 menunjukkan bahwa model **(CatBoost)** menghasilkan **performa terbaik** dari model lainnya. Berikut tabel perbandingan evaluasi model pada penelitian ini:

Metode	AUC	Balanced Acuracy
XGBoost	0.5915	0.6101
CatBoost	**0.6063**	**0.6201**
LightGBM	0.5563	0.5499
Random Forest	0.5508	0.5243

Evaluasi Model pada Cluster 10

Hasil pemodelan klasifikasi pada *cluster* 10 menunjukkan bahwa model **(CatBoost)** menghasilkan **performa terbaik** dari model lainnya. Berikut tabel perbandingan evaluasi model pada penelitian ini:

Metode	AUC	Balanced Acuracy
XGBoost	0.6310	0.5899
CatBoost	**0.6420**	**0.6765**
LightGBM	0.6360	0.5811
Random Forest	0.6329	0.5144

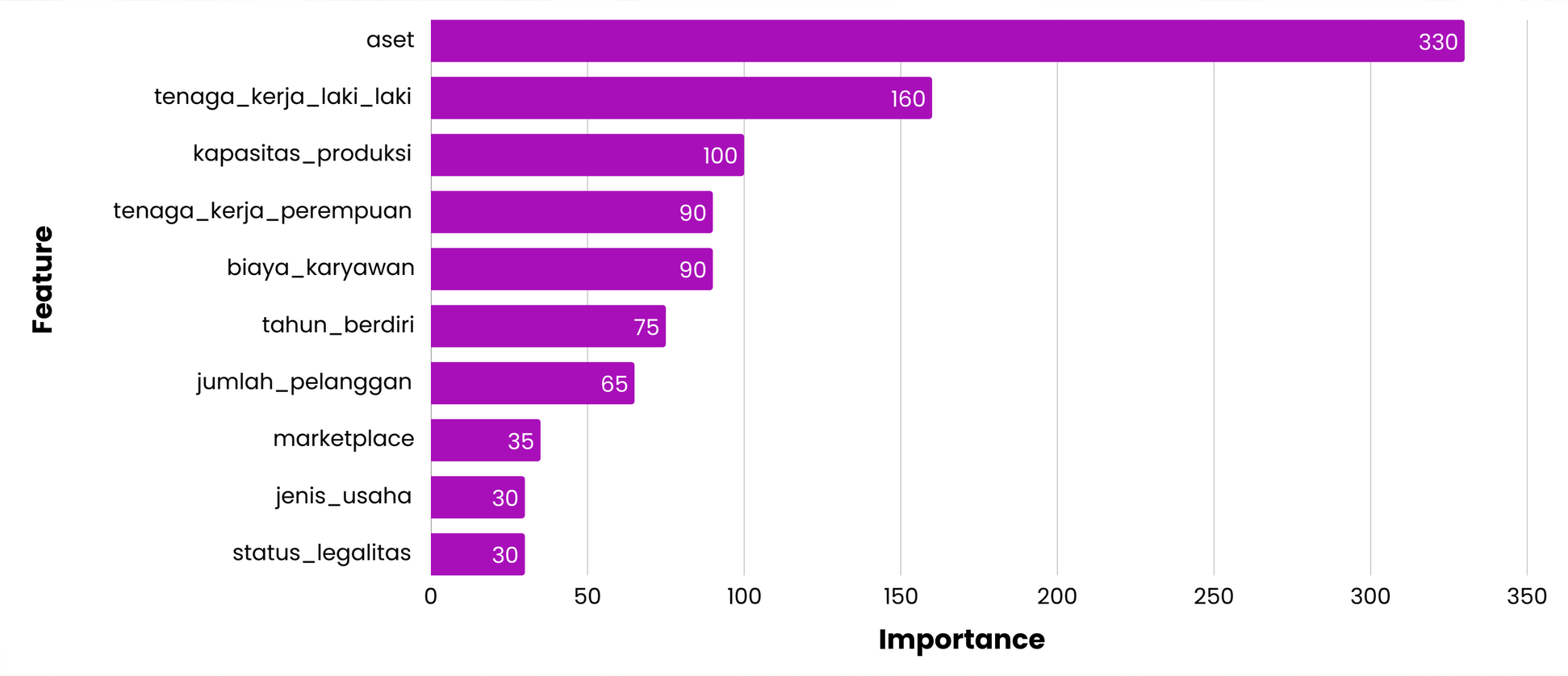
Evaluasi Model pada Cluster 11

Hasil pemodelan klasifikasi pada *cluster* 11 menunjukkan bahwa model **(CatBoost)** menghasilkan **performa terbaik** dari model lainnya. Berikut tabel perbandingan evaluasi model pada penelitian ini:

Metode	AUC	Balanced Acuracy
XGBoost	0.5532	0.5646
CatBoost	**0.5566**	**0.5692**
LightGBM	0.5535	0.5487
Random Forest	0.5108	0.5108

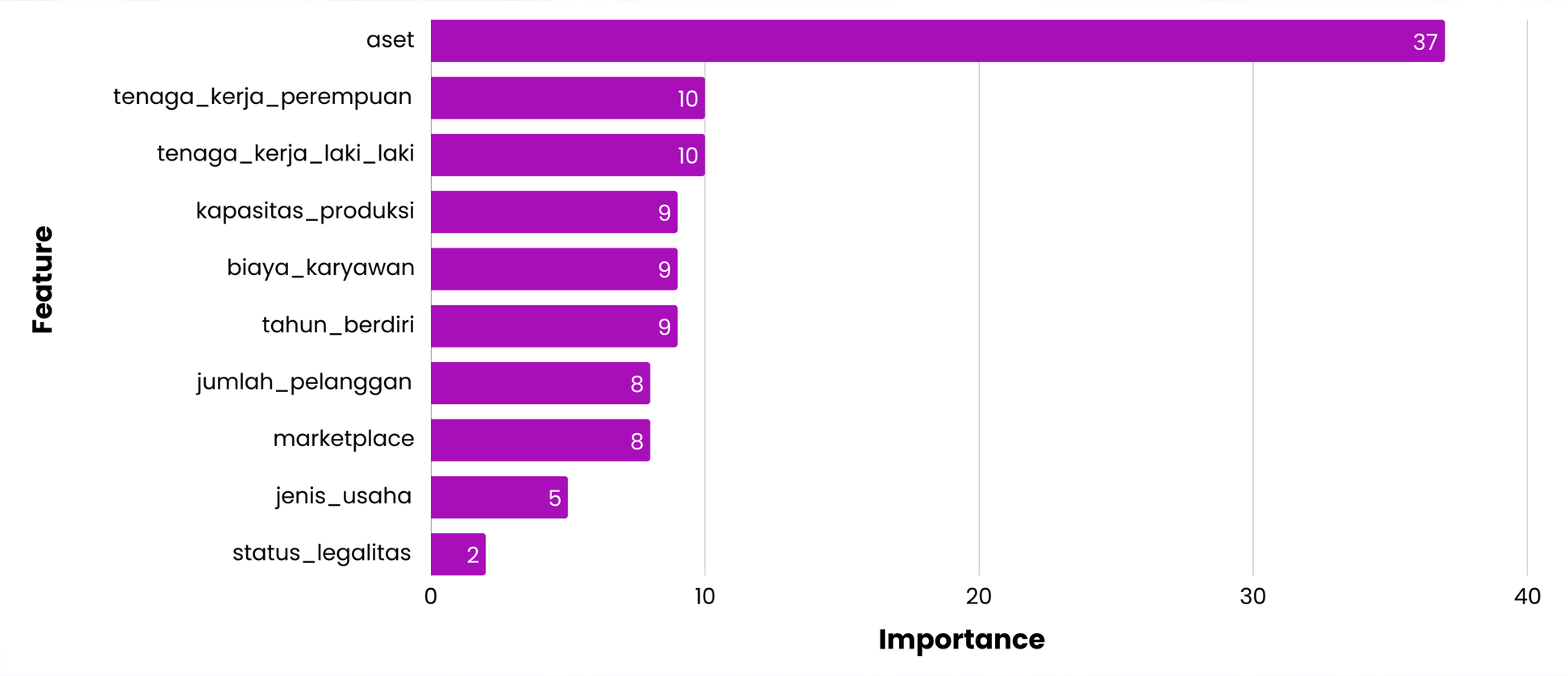
Feature Importance Cluster 00

Berdasarkan hasil model terbaik pada *cluster* 00, yaitu ***Light Gradient Boosting Machine*** **(LightGBM)**, berikut merupakan ***feature importance*** pada model tersebut:



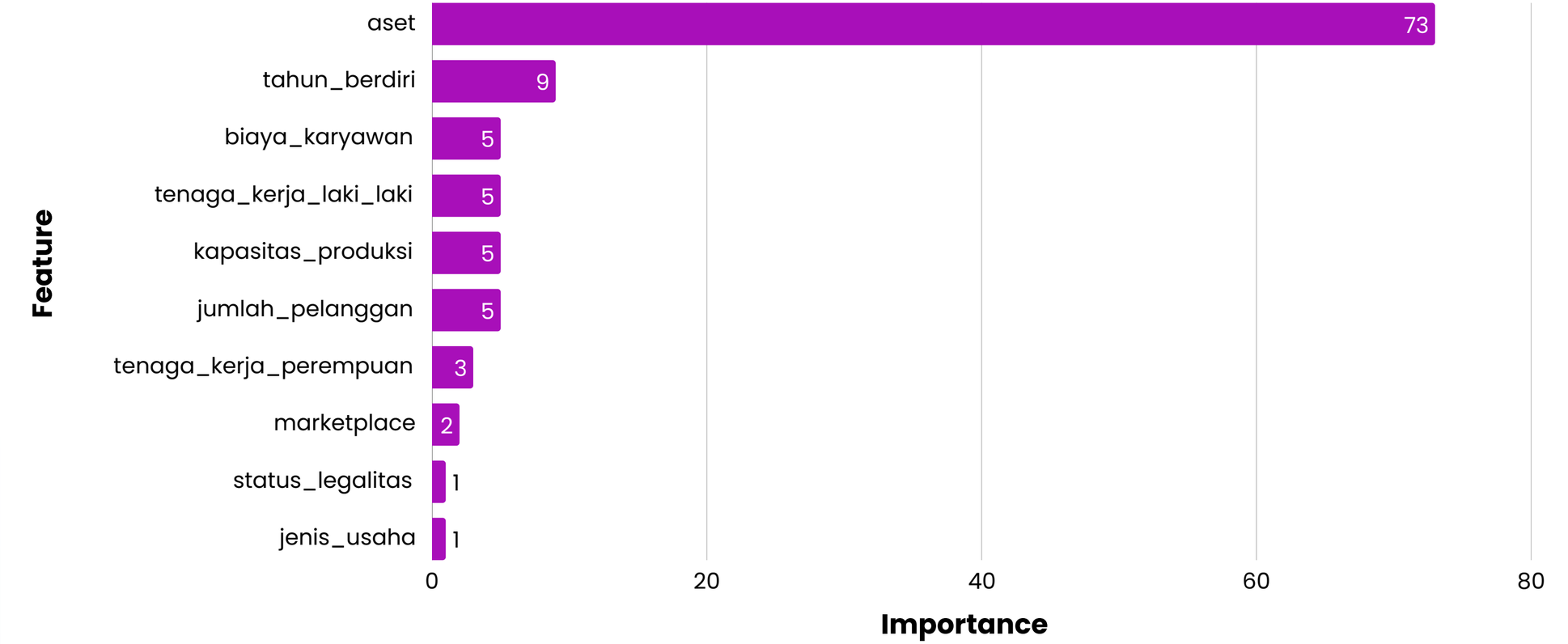
Feature Importance Cluster 01

Berdasarkan hasil model terbaik pada *cluster* 01, yaitu **CatBoost**, berikut merupakan ***feature importance*** pada model tersebut:



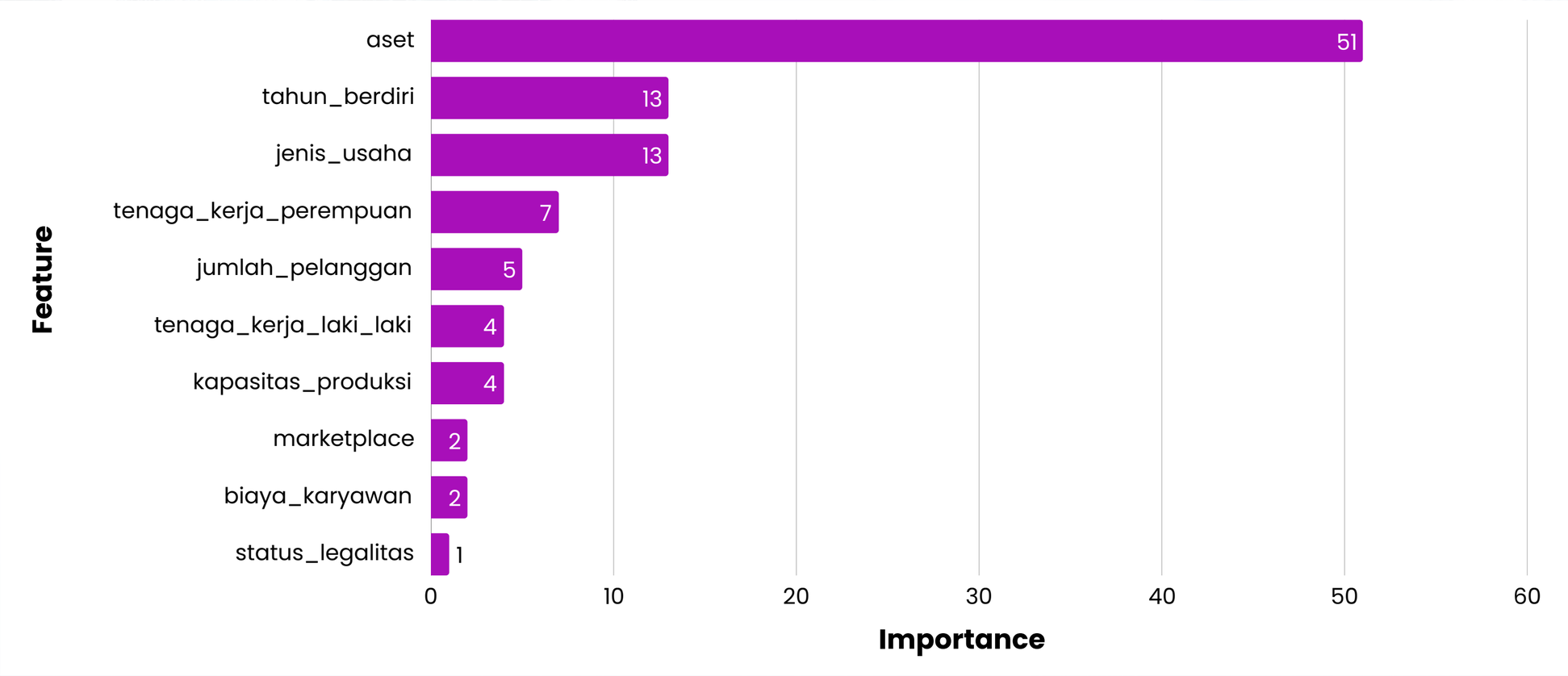
Feature Importance Cluster 10

Berdasarkan hasil model terbaik pada *cluster 10*, yaitu **CatBoost**, berikut merupakan ***feature importance*** pada model tersebut:



Feature Importance Cluster 11

Berdasarkan hasil model terbaik pada *cluster* 11, yaitu **CatBoost**, berikut merupakan ***feature importance*** pada model tersebut:



Politeknik Statistika STIS

Penutup

Kesimpulan

Cluster

Struktur dan Pola Usaha

Kinerja dan Skala Usaha

Dominan UMKM masuk dalam kategori usaha dengan **umur tua dan aset rendah**

Klasifikasi

CatBoost merupakan model terbaik di hampir semua *cross cluster*

Khusus untuk *cross cluster* 00, **LihgtGBM** menjadi model terbaik

Variabel aset secara konsisten menjadi faktor **paling berpengaruh** untuk seluruh *cross cluster*

Rekomendasi

Cluster 00 dan 01

Melakukan **investasi pada aset produktif** yang benar-benar berdampak pada *output* usaha, **meningkatkan keterampilan tenaga kerja** agar lebih produktif, dan **meningkatkan kapasitas produksi** dengan diversifikasi produk atau menambah jam kerja mesin.

Cluster 10 dan 11

Melakukan **investasi pada program pelatihan usaha** yang intensif agar dapat *sustain* dan bertumbuh dalam waktu panjang, serta menerapkan **mekanisme efisiensi pengeluaran SDM yang tidak perlu** namun tetap **mengoptimalkan kompetensi pegawai**.

Politeknik Statistika STIS

Tim StatStorm - Polstat STIS

Terima Kasih

Minggu, 15 Juni 2025