Analisis Faktor yang Memengaruhi Tingkat Pengangguran Terbuka di Jawa Timur Tahun 2023 dengan Regresi Probit

Dosen Pengampu: Ronny Susetyoko S.Si., M.Si

Disusun Oleh:
Al Rahma Dinda Salsabila (3323600038)

D4 Sains Data Terapan

Pengertian Regresi Probit

Model regresi Probit digunakan untuk menganalisis hubungan antara variabel dependen yang bersifat kategori (kualitatif) dengan variabel-variabel independen yang dapat berupa kualitatif maupun kuantitatif. Berbeda dengan regresi linear, model ini tidak menghasilkan nilai kontinu, melainkan memperkirakan probabilitas suatu peristiwa terjadi. Probit menggunakan fungsi distribusi normal kumulatif (Normal Cumulative Distribution Function / CDF) sebagai link function untuk menghubungkan variabel independen dengan peluang terjadinya kategori tertentu, sehingga hasil estimasi tetap berada dalam rentang probabilitas 0 hingga 1.

Latar Belakang

Tingkat Pengangguran Terbuka (TPT) Jawa Timur pada tahun 2023 tercatat sebesar 4,88%, sedikit lebih rendah dari rata-rata nasional sebesar 5,32% (BPS, 2023). Meski demikian, antar kabupaten/kota di Jawa Timur terdapat variasi yang cukup besar, di mana beberapa daerah menunjukkan TPT jauh di atas rata-rata provinsi, sementara daerah lain relatif rendah. Perbedaan ini diduga dipengaruhi oleh faktor-faktor seperti Tingkat Partisipasi Angkatan Kerja (TPAK), Indeks Pembangunan Manusia (IPM), dan Upah Minimum Kabupaten/Kota (UMK). Oleh karena itu, analisis dengan model Probit biner digunakan untuk mengidentifikasi faktor-faktor yang berpengaruh terhadap kemungkinan suatu daerah memiliki TPT tinggi.

Tujuan

01

01 Mengidentifikasi faktor-faktor yang memengaruhi probabilitas suatu kabupaten/kota di Jawa Timur memiliki TPT tinggi pada tahun 2023.

02 Menerapkan model Probit biner untuk menganalisis hubungan TPAK, IPM, dan UMK dengan status TPT tinggi (1) atau rendah (0).

Manfaat

01

01 Menjadi penerapan nyata dari materi Ekonometrika Terapan melalui penggunaan model Probit dengan data BPS.

02 Memberikan masukan bagi pemerintah daerah mengenai faktor penting yang perlu diperhatikan dalam kebijakan pengurangan pengangguran.

Dataset

Sumber Dataset: BPS Jawa Timur

```
df = pd.read_excel("Dataset TPT Jatim 2023.xlsx")
df
```

	Kabupaten/Kota	TPT	TPAK	IPM	UMK
0	Pacitan	1.83	81.64	70.19	2157270.25
1	Ponorogo	4.66	75.88	72.50	2149709.45
2	Trenggalek	4.52	80.72	71.73	2139426.01
3	Tulungagung	5.65	74.70	74.61	2229358.67
4	Blitar	4.91	73.50	72.49	2215071.18
5	Kediri	5.79	68.74	73.96	2243422.93
6	Malang	5.70	70.66	72.16	3268275.36
7	Lumajang	3.67	68.49	67.87	2200607.20
8	Jember	4.01	72.30	68.64	2555662.91
9	Banyuwangi	4.75	79.04	72.61	2528899.12
10	Bondowoso	4.15	74.39	67.99	2154504.13
11	Situbondo	3.27	75.28	69.16	2137025.85
12	Probolinggo	3.24	69.48	67.79	2753265.95
13	Pasuruan	5.48	71.21	70.29	4515133.19
14	Sidoarjo	8.05	69.62	81.55	4518581.85
15	Mojokerto	4.67	72.51	75.53	4504787.17
16	Jombang	4.66	71.91	74.60	2854095.88
17	Nganjuk	4.68	66.89	73.71	2167007.05
18	Madiun	5.14	72.49	72.97	2154251.34
19	Magetan	4.16	78.48	75.41	2153062.37
20	Ngawi	2.41	69.43	72.47	2158844.59
21	Bojonegoro	4.63	74.29	70.85	2279568.07
22	Tuban	4.40	74.73	70.34	2739224.88
23	Lamongan	5.46	75.08	74.53	2701977.27
24	Gresik	6.82	70.12	77.98	4522030.51
25	Bangkalan	6.18	71.49	65.75	2152450.83
26	Sampang	2.72	73.54	64.13	2114335.27
27	Pamekasan	1.74	77.14	67.96	2133655.03
28	Sumenep	1.71	78.86	68.61	2176819.94
29	Kota Kediri	4.06	71.83	80.44	2318116.63
30	Kota Blitar	5.24	72.26	80.63	2239024.44
31	Kota Malang	6.80	67.58	83.39	3194143.98
32	Kota Probolinggo	4.53	70.61	75.43	2576240.63
33	Kota Pasuruan	5.64	75.65	77.17	3038837.64
34	Kota Mojokerto	4.73	72.50	80.07	2710452.36
35	Kota Madiun	5.85	69.29	82.71	2190216.37
36	Kota Surabaya	6.76	68.73	83.45	4525479.19
37	Kota Batu	4.52	78.99	78.18	3030367.09

Dataset TPT Jawa Timur tahun 2023 mencakup 38 kabupaten/kota dengan variabel TPT, TPAK, IPM, dan UMK. Nilai TPT sangat bervariasi, mulai dari 1,71 persen di Sumenep hingga 8,05 persen di Sidoarjo, sementara variabel lainnya juga menunjukkan perbedaan yang cukup besar antar daerah. Adanya variasi tersebut penting untuk dianalisis lebih lanjut menggunakan model regresi Probit untuk mengetahui faktor-faktor yang memengaruhi TPT di Jawa Timur.

Analisis Deskriptif

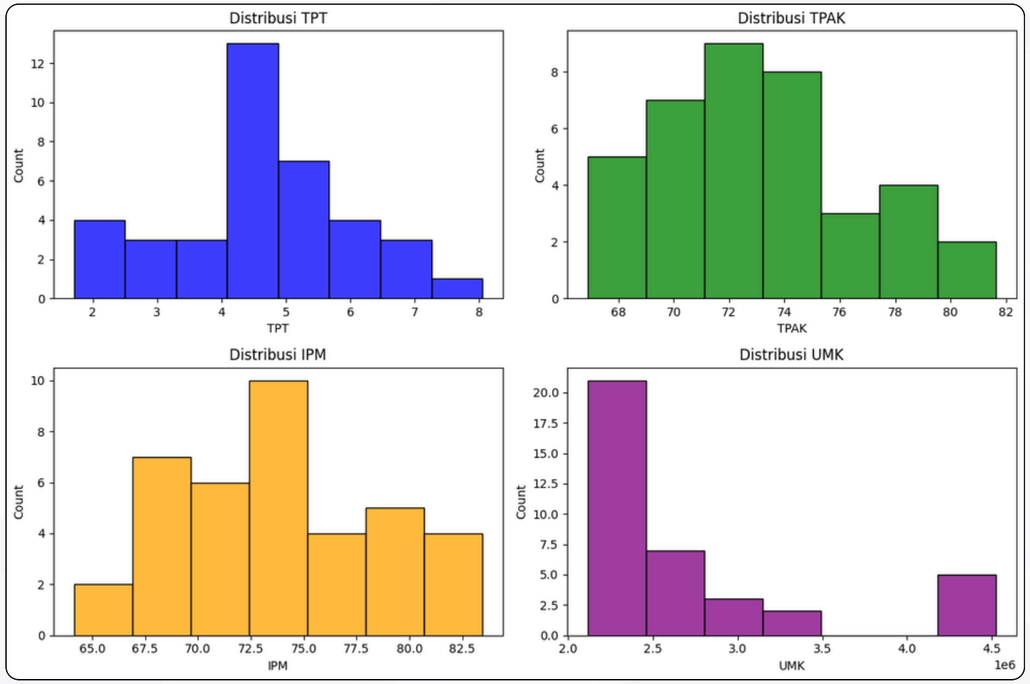
```
df.describe()
```

	TPT	TPAK	IPM	UMK
count	38.000000	38.000000	38.000000	3.800000e+01
mean	4.662895	73.159211	73.680263	2.694768e+06
std	1.428828	3.767150	5.053755	7.891471e+05
min	1.710000	66.890000	64.130000	2.114335e+06
25%	4.082500	70.242500	70.215000	2.157664e+06
50%	4.665000	72.495000	72.790000	2.261496e+06
75%	5.600000	75.230000	76.760000	2.828888e+06
max	8.050000	81.640000	83.450000	4.525479e+06

Hasil analisis deskriptif menunjukkan bahwa rata-rata Tingkat Pengangguran Terbuka (TPT) di Jawa Timur pada tahun 2023 adalah sebesar 4,66 persen, dengan variasi antara 1,71 persen hingga 8,05 persen. Rata-rata Tingkat Partisipasi Angkatan Kerja (TPAK) tercatat 73,16 persen, dengan nilai terendah 66,89 persen dan tertinggi 81,64 persen. Sementara itu, Indeks Pembangunan Manusia (IPM) memiliki rata-rata 73,68, dengan rentang dari 64,13 hingga 83,45. Adapun Upah Minimum Kabupaten/Kota (UMK) rata-rata sebesar Rp2.694.768, dengan nilai minimum Rp2.114.335 dan maksimum Rp4.525.479.

Visualisasi Distribusi Variabel

```
fig, axes = plt.subplots(2, 2, figsize=(12,8))
sns.histplot(df['TPT'], ax=axes[0,0], color="blue"); axes[0,0].set_title("Distribusi TPT")
sns.histplot(df['TPAK'], ax=axes[0,1], color="green"); axes[0,1].set_title("Distribusi TPAK")
sns.histplot(df['IPM'], ax=axes[1,0], color="orange"); axes[1,0].set_title("Distribusi IPM")
sns.histplot(df['UMK'], ax=axes[1,1], color="purple"); axes[1,1].set_title("Distribusi UMK")
plt.tight_layout(); plt.show()
```



Histogram menggambarkan bahwa TPT umumnya berada pada kisaran 4–5 persen. Variabel TPAK sebagian besar terkonsentrasi pada rentang 70–75 persen, sedangkan IPM banyak berada pada kisaran 72,5–75 persen. Untuk variabel UMK, mayoritas daerah berada pada kisaran Rp2–2,5 juta, meskipun terdapat beberapa daerah dengan nilai yang jauh lebih tinggi.

Variabel Target

```
# Menghitung rata-rata TPT provinsi
mean_TPT = df['TPT'].mean()
print("Rata-rata TPT Jatim 2023:", round(mean_TPT,2))

# Membuat variabel target berdasarkan rata-rata
df['TPT_Category'] = (df['TPT'] > mean_TPT).astype(int)

df[['Kabupaten/Kota','TPT','TPT_Category']]
```

```
Rata-rata TPT Jatim 2023: 4.66
```

	Kabupaten/Kota	TPT	TPT_Category
0	Pacitan	1.83	0
1	Ponorogo	4.66	0
2	Trenggalek	4.52	0
3	Tulungagung	5.65	1
4	Blitar	4.91	1
5	Kediri	5.79	1
6	Malang	5.70	1
7	Lumajang	3.67	0
8	Jember	4.01	0
9	Banyuwangi	4.75	1
10	Bondowoso	4.15	0
11	Situbondo	3.27	0
12	Probolinggo	3.24	0
13	Pasuruan	5.48	1
14	Sidoarjo	8.05	1
15	Mojokerto	4.67	1
16	Jombang	4.66	0
17	Nganjuk	4.68	1
18	Madiun	5.14	1
19	Magetan	4.16	0
20	Ngawi	2.41	0
21	Bojonegoro	4.63	0
22	Tuban	4.40	0
23	Lamongan	5.46	1
24	Gresik	6.82	1
25	Bangkalan	6.18	1
26	Sampang	2.72	0
27	Pamekasan	1.74	0
28	Sumenep	1.71	0
29	Kota Kediri	4.06	0
30	Kota Blitar	5.24	1
31	Kota Malang	6.80	1
32	Kota Probolinggo	4.53	0
33	Kota Pasuruan	5.64	1
34	Kota Mojokerto	4.73	1
35	Kota Madiun	5.85	1
36	Kota Surabaya	6.76	1
37	Kota Batu	4.52	0

Rata-rata TPT Jawa Timur tahun 2023 tercatat sebesar 4,66 persen. Berdasarkan angka tersebut, dibentuk variabel target TPT_Category, di mana daerah dengan TPT di atas nilai rata-rata dikategorikan sebagai tinggi (1), sedangkan daerah dengan TPT di bawah atau sama dengan rata-rata dikategorikan sebagai rendah (0). Kategori ini kemudian digunakan sebagai variabel dependen dalam analisis regresi Probit.

Transformasi Variabel

```
df['log_UMK'] = np.log(df['UMK'])
```

Untuk menyederhanakan skala pada variabel UMK sekaligus menormalkan distribusinya, dilakukan transformasi logaritma natural sehingga diperoleh variabel baru, yaitu log_UMK. Variabel ini kemudian digunakan dalam analisis regresi Probit agar hasil estimasi lebih stabil serta lebih mudah untuk diinterpretasikan.

Mendefinisikan Variabel

```
X = df[['TPAK','IPM','log_UMK']]
X = sm.add_constant(X)
y = df['TPT_Category']
```

Pada model ini, variabel independen yang digunakan adalah TPAK, IPM, dan log_UMK, sedangkan variabel dependennya adalah TPT_Category yang menunjukkan apakah suatu daerah termasuk dalam kategori TPT tinggi atau rendah. Selain itu, ditambahkan konstanta pada variabel X agar model Probit dapat menghitung nilai intercept dengan lebih tepat.

Rata-rata Variabel Numerik Berdasarkan Kategori Target

TPT_Category	TPT	TPAK	IPM	UMK	log_UMK
0	3.625789	74.631053	71.278421	2.356937e+06	14.665964
1	5.700000	71.687368	76.082105	3.032600e+06	14.879607

Berdasarkan perbandingan rata-rata, daerah dengan kategori TPT tinggi (1) memiliki rata-rata TPT sebesar 5,70 persen, dengan nilai TPAK yang lebih rendah (71,69 persen), IPM yang lebih tinggi (76,08), serta UMK dan log_UMK yang juga lebih besar dibandingkan dengan daerah yang masuk kategori TPT rendah (0).

Estimasi Model Probit

```
model = Probit(y, X).fit()
print(model.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.483999
         Iterations 6
                          Probit Regression Results
==============================================================================
Dep. Variable:           TPT_Category   No. Observations:                   38
Model:                         Probit   Df Residuals:                       34
Method:                           MLE   Df Model:                            3
Date:                Fri, 26 Sep 2025   Pseudo R-squ.:                  0.3017
Time:                        21:27:41   Log-Likelihood:                -18.392
converged:                       True   LL-Null:                       -26.340
Covariance Type:            nonrobust   LLR p-value:                  0.001191
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -27.3190     20.006     -1.366      0.172     -66.530      11.892
TPAK          -0.1138      0.070     -1.626      0.104      -0.251       0.023
IPM            0.1103      0.057      1.928      0.054      -0.002       0.222
log_UMK        1.8678      1.330      1.404      0.160      -0.739       4.474
==============================================================================
```

Hasil estimasi regresi Probit menunjukkan bahwa model yang digunakan signifikan secara keseluruhan, dengan nilai LLR p-value sebesar 0,001 dan Pseudo R^2 sebesar 0,30. Secara parsial, variabel IPM berpengaruh positif terhadap peluang suatu daerah masuk kategori TPT tinggi dengan tingkat signifikansi mendekati 5 persen. Sementara itu, variabel TPAK cenderung berpengaruh negatif dan log_UMK berpengaruh positif, namun keduanya belum menunjukkan pengaruh yang signifikan secara statistik.

Marginal Effects

```
mfx = model.get_margeff(at='overall', method='dydx')
print(mfx.summary())
```

```
       Probit Marginal Effects
=====================================
Dep. Variable:           TPT_Category
Method:                          dydx
At:                           overall
==============================================================================
                dy/dx    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
TPAK          -0.0307      0.017     -1.835      0.067      -0.063       0.002
IPM            0.0297      0.013      2.258      0.024       0.004       0.056
log_UMK        0.5035      0.328      1.535      0.125      -0.139       1.146
==============================================================================
```

Berdasarkan hasil marginal effects, variabel IPM memiliki pengaruh positif dan signifikan terhadap kemungkinan suatu daerah masuk kategori TPT tinggi. Peningkatan satu unit IPM akan meningkatkan probabilitas TPT sekitar 2,97%. Sementara itu, variabel TPAK cenderung menurunkan probabilitas TPT sebesar 3,07%, dan log_UMK berpotensi meningkatkan probabilitas TPT hingga 50,35%, namun keduanya hanya bersifat indikatif karena tidak signifikan secara statistik. Dengan demikian, dapat disimpulkan bahwa IPM merupakan variabel yang paling berpengaruh, sedangkan TPAK dan log_UMK hanya menunjukkan indikasi pengaruh tanpa bukti statistik yang kuat.

Goodness-of-fit: Pseudo R^2 (McFadden)

```
llf = model.llf       # log-likelihood model
llnull = model.llnull # log-likelihood model kosong
pseudo_r2 = 1 - (llf/llnull)
print("Pseudo R-squared (McFadden):", round(pseudo_r2,3))
```

```
Pseudo R-squared (McFadden): 0.302
```

Nilai Pseudo R^2 McFadden sebesar 0,302 menunjukkan bahwa model Probit mampu menjelaskan sekitar 30% variasi kategori TPT. Untuk model diskrit seperti Probit, angka ini sudah termasuk cukup baik karena Pseudo R^2 biasanya lebih rendah dibandingkan R^2 pada regresi linear, di mana kisaran 0,2–0,4 dianggap cukup kuat.

Validasi Model

```
df['pred_prob'] = model.predict(X)                    # probabilitas prediksi
df['pred_class'] = (df['pred_prob'] > 0.5).astype(int) # klasifikasi biner (threshold 0.5)

# Confusion Matrix
cm = confusion_matrix(y, df['pred_class'])
print("Confusion Matrix:\n", cm)

# AUC Score
auc = roc_auc_score(y, df['pred_prob'])
print("\nAUC:", round(auc,3))

# Classification Report
print("\nClassification Report:\n", classification_report(y, df['pred_class']))
```

Pada tahap ini, hasil prediksi model dalam bentuk probabilitas dikonversi menjadi klasifikasi biner dengan menggunakan threshold 0,5. Hal ini berarti bahwa jika probabilitas suatu daerah lebih besar dari 0,5 maka dikategorikan sebagai TPT tinggi, sedangkan jika sama dengan atau kurang dari 0,5 dikategorikan sebagai TPT rendah. Konversi ini dilakukan agar prediksi model dapat dibandingkan langsung dengan data aktual. Selanjutnya, performa model dievaluasi melalui confusion matrix, AUC score, dan classification report untuk mengetahui ketepatan klasifikasi serta menilai akurasi, precision, dan recall yang dihasilkan.

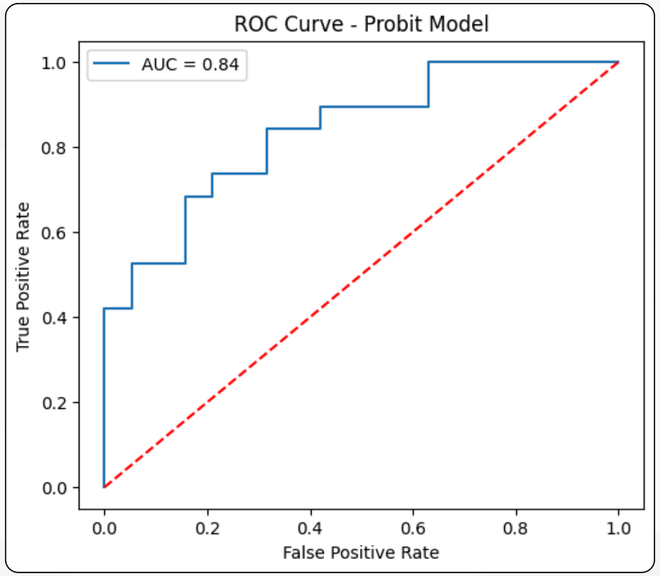
Validasi Model

```
Confusion Matrix:
 [[14  5]
 [ 5 14]]

AUC: 0.837

Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.74      0.74        19
           1       0.74      0.74      0.74        19

    accuracy                           0.74        38
   macro avg       0.74      0.74      0.74        38
weighted avg       0.74      0.74      0.74        38
```

- Confusion matrix menunjukkan model mampu mengklasifikasikan 14 dari 19 observasi dengan benar pada masing-masing kelas, meskipun masih terdapat 5 kesalahan prediksi.
- Nilai AUC sebesar 0,837 menandakan bahwa model cukup baik dalam membedakan kategori TPT tinggi (1) dan rendah (0).
- Berdasarkan classification report, akurasi keseluruhan model adalah 74%, dengan precision, recall, dan f1-score masing-masing sebesar 0,74.
- Hasil ini mengindikasikan bahwa performa model relatif seimbang dalam mengklasifikasikan kedua kategori TPT.

ROC Curve

```
fpr, tpr, thresholds = roc_curve(y, df['pred_prob'])
plt.figure(figsize=(6,5))
plt.plot(fpr, tpr, label=f"AUC = {auc:.2f}")
plt.plot([0,1],[0,1],'--',color='red')
plt.xlabel("False Positive Rate"); plt.ylabel("True Positive Rate")
plt.title("ROC Curve - Probit Model"); plt.legend(); plt.show()
```



Berdasarkan nilai AUC sebesar 0,84, dapat disimpulkan bahwa model Probit memiliki kemampuan klasifikasi yang cukup baik. Hal ini menunjukkan bahwa model efektif dalam membedakan antara kategori TPT tinggi dan TPT rendah pada data yang digunakan.

Kesimpulan

- Model Probit terbukti layak digunakan, terlihat dari hasil uji yang signifikan (LLR p-value = 0,001), nilai Pseudo R^2 = 0,302, dan AUC = 0,84, yang menunjukkan kemampuan model dalam klasifikasi sudah cukup baik.
- Variabel IPM menjadi faktor paling berpengaruh dengan arah positif. Artinya, semakin tinggi kualitas pendidikan, kesehatan, dan standar hidup (cerminan IPM), justru semakin besar peluang suatu daerah mengalami TPT tinggi. Hal ini dapat terjadi karena peningkatan kualitas SDM tidak selalu sejalan dengan ketersediaan lapangan kerja. Akibatnya, tenaga kerja berpendidikan tinggi sering sulit terserap, sehingga angka pengangguran terbuka justru meningkat.
- Variabel TPAK cenderung menurunkan peluang TPT tinggi, sedangkan UMK berarah positif, namun keduanya tidak signifikan secara statistik.
- Secara keseluruhan, IPM merupakan variabel yang paling berpengaruh, sementara TPAK dan UMK hanya menunjukkan pengaruh yang lebih lemah.

Link Akses Dataset & Source Code

Source Code & Dataset

Terima Kasih

D4 Sains Data Terapan