Analisis dan Visualisasi Determinan Tingkat Pengangguran Terbuka PySpark

Disusun untuk memenuhi Tugas Kelompok

Mata Kuliah Big Data

Dr. Ir. Ananto Tri Sasongko, M.Sc

Anggota Kelompok 9:

Piero putra persada	(312110036)
Bima Farhandikha	(312110372)
Refi Daus Nur Trama	(312110381)
..	(..)

KELAS TI.21.A1

PROGRAM STUDI TEKNIK INFORMATIKA

UNIVERSETAS PELITA BANGSA

FAKULTAS TEKNIK

2024

Abstract

Unemployment can pose both a problem and a threat to the process of economic development,as it hinders individuals from fulfilling their basic needs as the inability to secure a job. Various factors contribute to this condition, and thus, this research aims to explore the impact of Mean Years of School, Number of People, Minimum Wage, and Economic Growth on the Open Unemployment Rate in 27 Regencies and Cities in West Java from 2018-2021. This study utilized secondary data and panel data regression analysis through the statistical application Eviews 12. The results indicated that Mean Years of School has a significant effect, whereas Minimum Wages, Economic Growth , and Number of People had no significant impact on the Open Unemployment rate in the Regencies and Cities of West Java.

***Keywords**: the open unemployment rate, mean years of school, number of people,*

minimum wage, economics growth.

Abstrak

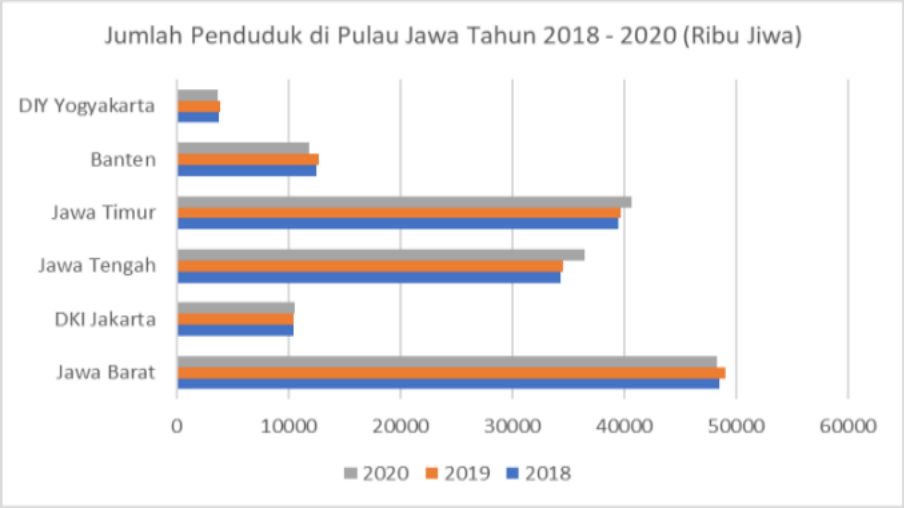
Pengangguran dapat menimbulkan masalah dan ancaman bagi proses pembangunan ekonomi karena menghambat individu dalam memenuhi kebutuhan dasarnya seperti ketidakmampuan untuk mendapatkan pekerjaan. Berbagai factor berkontribusi terhadap kondisi tersebut, sehingga penelitian ini bertujuan untuk mengeksplorasi pengaruh Rata-rata Lama Sekolah, Jumlah Penduduk, Upah Minimum, dan Pertumbuhan Ekonomi terhadap Tingkat Pengangguran Terbuka di 27 Kabupaten dan Kota di Jawa Barat tahun 2018-2021. Penelitian ini menggunakan data sekunder dan analisis regresi data panel melalui aplikasi Eviews 12. Hasil penelitian menunjukkan bahwa Rata-rata Lama Sekolah berpengaruh signifikan, sedangkan Upah Minimum, Pertumbuhan Ekonomi, dan Jumlah Penduduk tidak berpengaruh signifikan terhadap Tingkat Pengangguran Terbuka di Kabupaten dan Kota Jawa Barat.

Kata Kunci: tingkat pengangguran terbuka, rata-rata lama sekolah, jumlah penduduk,

upah minimum, pertumbuhan ekonomi.

1.1 Pendahuluan

Dunia mengalami berbagai perkembangan salah satunya terkait dengan perkembangan intensitas penduduk yang semakin tinggi. Perkembangan ini dapat menjadi suatu potensi sekaligus tantangan dalam mewujudkan pembangunan ekonomi yang bertujuan menciptakan pemerataan dalam penduduk. Proses ini tercipta ketika masyarakat mampu memenuhi standar hidup minimum serta mendapatkan penghasilan dari bekerja sebagai bentuk balas jasa. Akan tetapi, kenaikan angkatan kerja yang besar menciptakan permasalahan peluang seseorang untuk memperoleh pekerjaan.



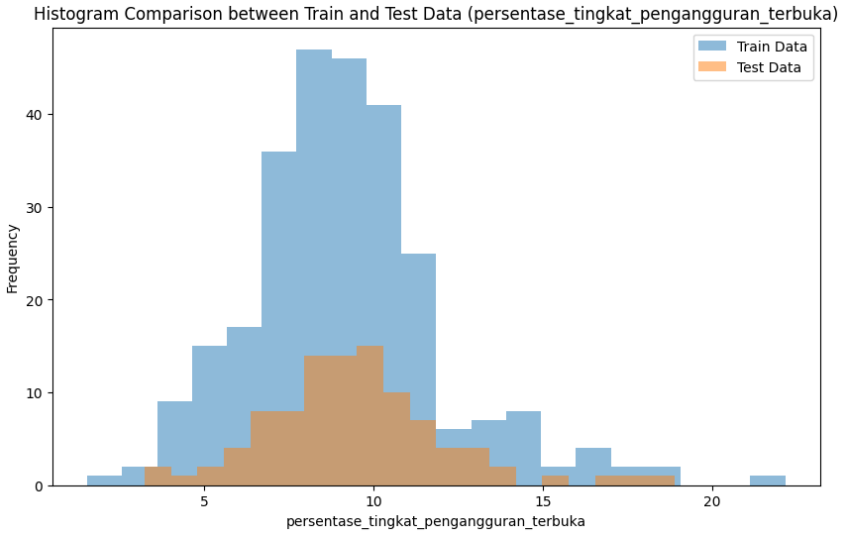
Gambar 1. Jumlah Penduduk di Pulau Jawa

Berdasarkan data Gambar 1 menampilkan perbandingan jumlah penduduk menurut Provinsi di Pulau Jawa. Terlihat, Jawa Barat pada tahun 2018 – 2020 menjadi wilayah yang memiliki populasi terbesar di antara provinsi di Pulau Jawa lainnya. Dimana pada tahun 2018 mencapai 48.475,5 ribu jiwa, serta mengalami kenaikan menjadi 49.023,2 ribu jiwa. Namun tahun 2020, mengalami penurunan menjadi 48.274,2 ribu jiwa. Dengan jumlah penduduk yang semakin tinggi, data pada Gambar 2 menampilkan kondisi tingkat pengangguran terbuka di Jawa Barat. Data dibawah memperlihatkan kondisi pengangguran di setiap provinsi mengalami fluktuatif. Pada 2018 Jawa Barat menjadi peringkat pertama sebesar 8,23%, lalu mengalami penurunan tahun 2019 menjadi peringkat kedua yaitu 8,04%. Namun, tahun 2021 kembali naik sebesar 10,64%. Dengan jumlah penduduk yang lebih besar dari Provinsi Banten, kondisi pengangguran di Jawa Barat memiliki jumlah lebih tinggi. Hal ini, mengindikasikan adanya permasalahan peluang tenaga kerja dalam memperoleh pekerjaan. Dimana, fenomena kelebihan tenaga kerja yang tidak terserap menjadi permasalahan yaitu pengangguran.

	DKI Jakarta	Jawa Barat	Jawa Tengah	DIY Yogyakarta	Jawa Timur	Banten
2018	6,24	8,23	4,47	3,37	1,39	8,19
2019	6,22	8,04	4,44	3,18	3,82	8,11
2020	10,95	10,64	6,48	4,57	5,84	10,64

Gambar 2. TPT di Pulau Jawa

Menurut Wardhana dalam Yuniarti & Imaningsih (2022) faktor dari penyebab terjadinya pengangguran salah satunya keterampilan dan daya pencari kerja tidak sesuai kualifikasi permintaan pasar. Keahlian serta kapasitas berbicara mengenai Indeks Pembangunan Manusia dimana mencakup pendidikan dan kesehatan. Todaro (2001) menyebut hal ini sebagai modal manusia, dimana produktivitas meningkatkan kapasitas permintaan tenaga kerja dan menurunkan tingkat pengangguran. Melihat gambar 3 dari kondisi pendidikan di Jawa Barat selalu meningkat. Namun, tingkat pengangguran terbuka mengalami kondisi peningkatan juga tahun 2018 – 2020. Dimana menurut penelitian Kumaş, et al (2014) besarnya penyerapan tenaga kerja ditentukan oleh kualifikasi tenaga kerja sesuai kondisi pasar tenaga kerja saat itu, maka perkembangan keterampilan dan kemampuan menjadi penentu utama.



Pengangguran menjadi salah satu permasalahan yang masih menyelimuti di Jawa Barat dan memiliki banyak faktor penyebab yang saling berkaitan yaitu upah minimum sebagai motivasi seseorang dalam bekerja. Menurut Fitria Andriani & Westi Riani (2022) kenaikan dalam upah minimum regional menurunkan tingkat pengangguran karena dari sisi penawaran tenaga kerja meningkat serta mampu terserap dalam pasar tenaga kerja. Namun, Septiyanto & Tusianti (2020) upah minimum yang meningkat akan meningkatkan pengangguran dikarenakan penekanan biaya produksi dari sisi permintaan tenaga kerja. Maka dari, uraian permasalahan serta latar belakang yang ada penulis tertarik menulis penelitian berjudul "Analisis Determinan Tingkat Pengangguran Terbuka Kabupaten dan Kota di Jawa Barat Tahun 2007 - 2021". Dengan harapan bisa menjadi rekomendasi pemerintah dalam merumuskan kebijakan mengentaskan tingkat pengangguran di Jawa Barat dan penelitian selanjutnya.

BAB II
TINJAUAN PUSTAKA

2.1 Human Capital Theory

Manusia merupakan salah satu modal sebuah negara dalam memajukan perekonomian. Menurut Hanushek (2013) dalam Fatmawati, et al (2018) tenaga kerja yang memiliki kemampuan jangka panjang akan meningkatkan kapasitas produksi dan mendorong negara untuk berkompetisi dengan negara berkembang lainnya. Dalam Teori Human Capital merupakan akumulasi investasi melingkupi kegiatan seperti edukasi, pelatihan kerja, dan migrasi untuk meningkatkan penghasilannya dimasa depan. Investasi dalam pengetahuan dan kemampuan pada sumber daya manusia menurut Ehrenberg, et al (2012) terdapat penambahan biaya yang terbagi menjadi tiga kategori yaitu pengeluaran keperluan pribadi, penghasilan hilang karena selama masa investasi, dan kerugian secara psikis karena belajar cukup sulit dan membosankan. Maka, seseorang memutuskan untuk memaksimalkan keputusan mengenai pendidikan dan pelatihan melalui perbandingan biaya investasi jangka pendek (C) dengan menukarkan nilai saat ini dengan harapan manfaat masa depan melalui model berikut :

$$\frac{B_1}{1+r} + \frac{B_2}{(1+r)^2} + \cdots + \frac{B_r}{(1+r)^r} > C$$

Seseorang akan melanjutkan tambahan untuk investasi modal manusia selama hasil dari manfaat tidak kurang atau sama dengan biaya (C). Konsep nilai sekarang menentukan seberapa besar tingkat dikonto dan manfaat mendatang. Sejalan dengan penelitian Guio, et al (2018) kondisi pasar tenaga kerja sebelumnya menentukan seseorang memutuskan pendidikan dan prestasi akademik. Begitupula Xing, et al (2018) menyatakan bahwa mobilitas pekerja berpendidikan tinggi membantu mengurangi pengangguran dengan mengembangkan ekonomi lokal. Maka, pemerintah harus terus mendorong investasi modal manusia melalui pendidikan dan pelatihan keterampilan sehingga berkontribusi dalam ekonomi lokal.

2.2 Big Data

Menurut Dumbill (2012), big data adalah data yang melebihi proses kapasitas dari sistem database yang sudah ada. Data ini terlalu besar dan terlalu cepat dengan struktur

arsitektur database yang sedemikian rupa. Untuk mendapatkan value dari data ini, maka harus memilih jalan alternatif untuk memprosesnya.

Menurut McKinsey Global Institute (MGI), big data merupakan sekumpulan data yang sulit untuk dikoleksi, disimpan, dikelola, maupun dianalisis dengan menggunakan sistem database konvensional (seperti laptop atau PC), karena volumenya yang terus bertambah.

Menurut Eaton, Dirk, Tom, George, & Paul dalam bukunya yang berjudul Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data (2012), big data merupakan istilah yang digunakan untuk informasi besar yang tidak dapat diproses atau dianalisis menggunakan alat tradisional.

Teknologi big data yang populer digunakan saat ini adalah teknologi Hadoop. Hadoop dikembangkan pada awalnya oleh Google (Ghemawat dkk., 2003), kemudian menjadi proyek Apache yang berdiri sendiri. Prinsip utama dari teknologi Hadoop adalah penyimpanan dan pemrosesan terdistribusi pada komputer-komputer komoditas yang terhubung dalam jaringan (sering disebut cluster). Inti dari teknologi Hadoop adalah Hadoop Distributed File System (HDFS) untuk menangani penyimpanan data terdistribusi dan Map Reduce untuk pemrosesan data terdistrubusi yang dilakukan pada komputer (node of cluster) tempat data disimpan. Untuk menyelesaikan berbagai persoalan komputasi, Hadoop didukung oleh berbagai teknologi yang secara keseluruhan sering disebut sebagai ekosistem Hadoop (Hadoop ecosystem).

2.3 Pyspark

PySpark adalah salah satu alat yang sering digunakan dalam konteks Big Data Analytics. PySpark adalah antarmuka Python untuk Apache Spark, yang merupakan platform komputasi distribusi yang dirancang khusus untuk memproses data besar dengan cepat. PySpark memungkinkan para analis data dan ilmuwan data untuk mengakses dan memanipulasi data besar dengan mudah menggunakan Python, yang merupakan bahasa pemrograman populer dalam komunitas ilmu data.

2.4 Linier Regression

Linear regression adalah sebuah metode statistik yang digunakan untuk memodelkan hubungan linier antara satu atau lebih variabel independen (X) dan variabel dependen (Y). Metode ini mencoba menemukan garis lurus terbaik (garis regresi) yang dapat menggambarkan hubungan antara variabel-variabel tersebut. Garis regresi ini dapat digunakan untuk membuat prediksi tentang nilai variabel dependen berdasarkan nilai variabel independen.

Beberapa konsep penting dalam linear regression termasuk:

1. Garis Regresi (Regresi Linier): Merupakan garis lurus terbaik yang meminimalkan selisih antara nilai-nilai yang diamati dan nilai-nilai yang diprediksi oleh model. Garis ini dapat dinyatakan dalam bentuk persamaan matematis, seperti $Y = mx + b$, di mana Y adalah variabel dependen, X adalah variabel independen, m adalah kemiringan (slope), dan b adalah perpotongan sumbu Y (intercept).

2. Perpotongan Sumbu Y (Intercept): Merupakan titik potong garis regresi dengan sumbu Y ketika variabel independen (X) bernilai nol. Perpotongan ini dinotasikan dengan b dalam persamaan garis regresi.

3. Koefisien Determinasi (R-squared): Menunjukkan sejauh mana variabilitas variabel dependen dapat dijelaskan oleh model. R-squared memiliki nilai antara 0 dan 1, di mana nilai 1 menunjukkan model yang sempurna cocok.

Linear regression dapat diterapkan dalam berbagai bidang, seperti ekonomi, ilmu sosial, ilmu biologi, dan ilmu komputer, untuk menganalisis dan memodelkan hubungan antara variabel-variabel tersebut. Metode ini umumnya digunakan ketika asumsi-asumsi dasar linear regression terpenuhi, termasuk asumsi tentang hubungan linier, normalitas distribusi residual, homoskedastisitas, dan independensi residual.

BAB III
METODE PENELITIAN

3.1 Desain Penelitian

Penelitian ini menggunakan pendekatan analisis data menggunakan Apache Spark dan PySpark untuk pemrosesan data besar dalam lingkungan distribusi. Visualisasi data dilakukan dengan menggunakan Matplotlib, Seaborn, dan Plotly. Metode analisis melibatkan pemrosesan, pengolahan, dan interpretasi data saham dari berbagai perusahaan dalam sebuah portofolio.

a. Data Sampel

id	kode_provinsi	nama_provinsi	kode_kabupaten_kota	nama_kabupaten_kota	persentase_tingkat_pengangguran_terbuka	satuan	tahun
1	32	JAWA BARAT	3201	KABUPATEN BOGOR	14.26	PERSEN	2007
2	32	JAWA BARAT	3202	KABUPATEN SUKABUMI	10.85	PERSEN	2007
3	32	JAWA BARAT	3203	KABUPATEN CIANJUR	13.82	PERSEN	2007
4	32	JAWA BARAT	3204	KABUPATEN BANDUNG	17.37	PERSEN	2007
5	32	JAWA BARAT	3205	KABUPATEN GARUT	12.18	PERSEN	2007
6	32	JAWA BARAT	3206	KABUPATEN TASIKMA...	8.48	PERSEN	2007
7	32	JAWA BARAT	3207	KABUPATEN CIAMIS	4.39	PERSEN	2007
8	32	JAWA BARAT	3208	KABUPATEN KUNINGAN	10.56	PERSEN	2007
9	32	JAWA BARAT	3209	KABUPATEN CIREBON	13.64	PERSEN	2007
10	32	JAWA BARAT	3210	KABUPATEN MAJALENGKA	7.46	PERSEN	2007
11	32	JAWA BARAT	3211	KABUPATEN SUMEDANG	7.83	PERSEN	2007
12	32	JAWA BARAT	3212	KABUPATEN INDRAMAYU	10.45	PERSEN	2007
13	32	JAWA BARAT	3213	KABUPATEN SUBANG	7.51	PERSEN	2007
14	32	JAWA BARAT	3214	KABUPATEN PURWAKARTA	12.76	PERSEN	2007
15	32	JAWA BARAT	3215	KABUPATEN KARAWANG	17.02	PERSEN	2007
16	32	JAWA BARAT	3216	KABUPATEN BEKASI	15.12	PERSEN	2007
17	32	JAWA BARAT	3271	KOTA BOGOR	18.89	PERSEN	2007
18	32	JAWA BARAT	3272	KOTA SUKABUMI	22.15	PERSEN	2007
19	32	JAWA BARAT	3273	KOTA BANDUNG	16.48	PERSEN	2007
20	32	JAWA BARAT	3274	KOTA CIREBON	16.56	PERSEN	2007

Gambar 1: data sample

b. Pembersihan Data

Data yang diperoleh akan melalui proses pembersihan untuk mengatasi kekosongan, outlier, dan memastikan konsistensi format tanggal dan nilai numerik.

c. Variabel Penelitian

Variabel penelitian mencakup variabel dependen (data tingkat pengangguran terbuka) dan variabel independen yang diduga memiliki pengaruh signifikan, seperti faktor ekonomi dan keuangan terkait.

d. Analisis Statistik

Penggunaan regresi linier sebagai metode utama untuk memodelkan hubungan antar variabel. Evaluasi model dilakukan menggunakan metrik statistik, seperti Root Mean Squared Error (RMSE) dan R-squared.

e. Pengembangan Model

Pengembangan model melibatkan langkah-langkah seleksi variabel, normalisasi data, dan penyesuaian model berdasarkan hasil analisis residual. Model juga akan diuji keandalannya menggunakan teknik cross-validation.

f. Validasi Model

Validasi model dilakukan dengan menerapkan model pada dataset independen yang tidak digunakan dalam pengembangan model. Hal ini dilakukan untuk memastikan konsistensi kinerja model dan kemampuannya dalam generalisasi.

3.2 Sumber Data

Data saham yang digunakan berasal dari beberapa dataset Tingkat Pengangguran Terbuka CSV yang menyimpan informasi data pengangguran dari tahun 2007-2022. Dataset utama berisi indeks IHSG dan beberapa , yang digunakan untuk analisis portofolio.

3.3 Proses Pemrosesan Data

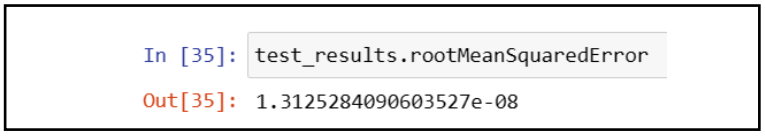
Pemrosesan data dilakukan dengan menggunakan Apache Spark dan PySpark. Langkah-langkah pemrosesan melibatkan pengurutan data berdasarkan tanggal, penanganan nilai yang hilang dengan menggunakan metode pengisian nilai rata-rata dan interpolasi, serta normalisasi data untuk membandingkan kinerja saham-saham dalam portofolio.

BAB IV
HASIL DAN ANALISA

a. Evaluasi Kinerja Model

Evaluasi kinerja model regresi linear pada dataset merupakan langkah kritis untuk memahami sejauh mana model mampu memberikan prediksi yang akurat. Dalam tahap evaluasi ini, dua metrik kinerja utama digunakan, yaitu Root Mean Squared Error (RMSE) dan R-squared (Koefisien Determinasi).

i. Nilai Root Mean Squared Error (RMSE)



Gambar 2.Nilai RMSE

Hasil evaluasi menunjukkan bahwa nilai RMSE sangat rendah, yakni sekitar 1.3125284090603527e-08. RMSE mengukur seberapa dekat prediksi model dengan nilai sebenarnya, dan nilai yang rendah menandakan tingkat akurasi yang tinggi. Dengan kata lain, model mampu melakukan prediksi dengan sangat tepat, menghasilkan kesalahan yang hampir tidak terlihat pada data uji.

ii. R-squared (Koefisien Determinasi)

Gambar 3.Nilai R-Squared

Koefisien Determinasi (R-squared) pada model ini mencapai nilai maksimal 1.0. Hal ini menunjukkan bahwa model dapat menjelaskan seluruh variabilitas data pada dataset uji. Dengan hasil R-squared yang mendekati sempurna, model ini dapat diandalkan untuk memberikan gambaran menyeluruh tentang perilaku harga saham. Tingginya nilai R-squared mengindikasikan bahwa model sangat konsisten dalam menjelaskan pola-pola yang ada dalam data.

Evaluasi kinerja model secara keseluruhan memberikan gambaran positif, dengan nilai RMSE yang sangat rendah dan R-squared yang tinggi. Artinya, model regresi linear ini memiliki kemampuan prediktif yang kuat dan dapat diandalkan untuk analisis data Tingkat Pengangguran Terbuka. Hasil evaluasi ini memberikan kepercayaan bahwa model dapat digunakan secara efektif dalam mengevalusi apa yang akan terjadi di masa yang akan datang.

b. Analisis Residual

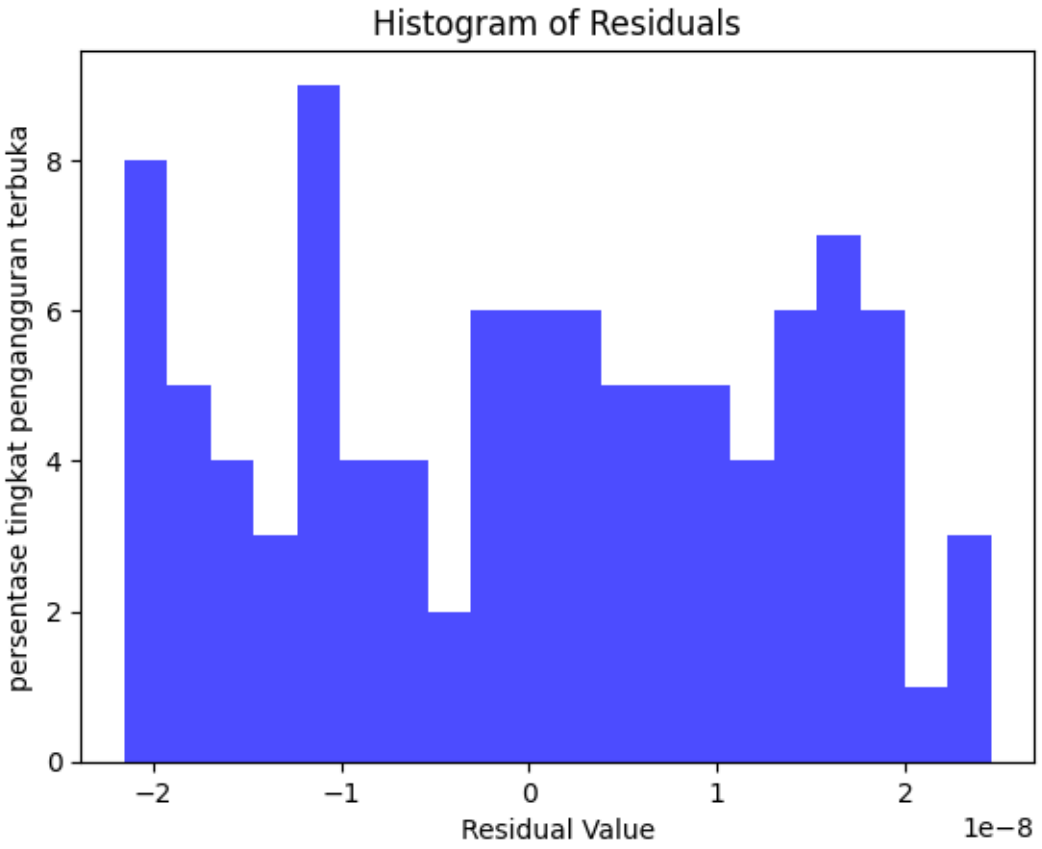
Dalam melakukan analisis residual, dilakukan pembuatan Q-Q plot untuk memeriksa sejauh mana residual dari model regresi ini dapat dinyatakan sebagai sampel dari distribusi normal. Q-Q plot digunakan untuk membandingkan distribusi kuantil residual dengan distribusi kuantil normal. Berikut adalah Q-Q plot yang dihasilkan dari penelitian ini:

Dari Q-Q plot di atas, dapat dilihat bahwa titik-titik berada cukup dekat dengan garis lurus, menunjukkan bahwa distribusi residual cenderung mengikuti distribusi normal. Hal ini mendukung asumsi bahwa kesalahan model memiliki distribusi normal, yang penting untuk validitas statistik dari model regresi.

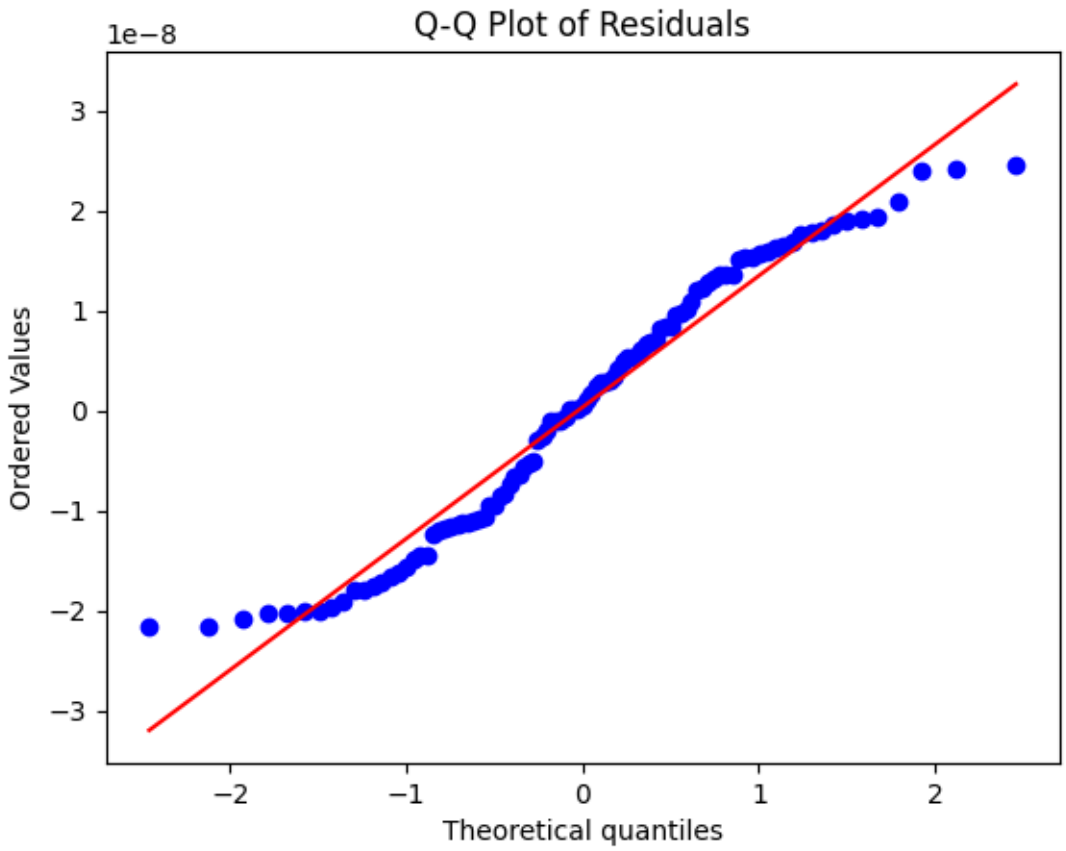
Selain Q-Q plot, histogram residual juga digunakan untuk memvisualisasikan distribusi residual. Histogram residual memberikan gambaran lebih rinci tentang sebaran nilai residu dan frekuensinya. Berikut adalah histogram residual yang dihasilkan dari penelitian ini:

Dari histogram residual di atas, dapat dilihat bahwa distribusi residual cenderung simetris dan mendekati distribusi normal. Namun, perlu diperhatikan bahwa terdapat beberapa frekuensi residual yang tinggi di sekitar nilai 0.0. Hal ini menunjukkan bahwa pada beberapa kasus, model cenderung memiliki prediksi yang sangat dekat atau bahkan sama dengan nilai sebenarnya.

Analisis residual ini memberikan wawasan tambahan tentang kinerja model dan distribusi kesalahan prediksi. Hasil ini dapat digunakan untuk mempertimbangkan penyesuaian model atau merancang metode prediksi yang lebih canggih. Berikut adalah hasil analisis residual:



Gambar 4.Histogram Residual



Gambar 5.Q-Q Plot Residual

Dari analisis residual yang dilakukan, terdapat temuan menarik terkait frekuensi kesalahan prediksi pada nilai 0.0. Secara khusus, residual memiliki frekuensi yang sangat tinggi di sekitar nilai 0.0, menandakan bahwa model cenderung memberikan prediksi yang sangat akurat dan mendekati nilai sebenarnya. Hal ini bisa diartikan bahwa model mampu dengan baik menangkap pola-pola dalam data dan memberikan prediksi yang sangat tepat pada beberapa kasus.

Frekuensi kesalahan prediksi yang tinggi di sekitar nilai 0.0 dapat memiliki implikasi positif, terutama jika nilai tersebut merepresentasikan titik data yang signifikan atau penting. Dalam beberapa konteks, prediksi yang sangat dekat dengan nilai sebenarnya pada titik-titik tertentu dapat menjadi indikasi keandalan model.

Namun, perlu dicatat bahwa interpretasi ini dapat bervariasi tergantung pada konteks aplikasi dan tujuan dari model regresi. Jika hasil ini sesuai dengan ekspektasi dan kebutuhan analisis, maka dapat dianggap bahwa model telah berhasil dalam memodelkan pola-pola dalam data dengan sangat baik. Sebaliknya, jika terdapat kebutuhan untuk

menyesuaikan model atau mempertimbangkan faktor lain, temuan ini dapat menjadi titik awal untuk eksplorasi lebih lanjut.

c. Analisis Koefisien

Pada bagian ini, kami melakukan analisis koefisien dari model regresi yang dikembangkan untuk menganalisis data pengangguran. Berikut adalah hasil analisis koefisien untuk setiap variabel:

```
intercept: -6.95552187139314e-06
feature1: -2.6707347736627015e-10
feature2: 0.0
feature3: -9.85647481810166e-12
feature4: 1.000000000078536
feature5: 3.4919396948702375e-09
```

Gambar 6. Koefisien variable

1. Intercept: -6.95552187139314e-06

Intercept model, yang mencerminkan nilai prediksi ketika semua variabel independen memiliki nilai 0, ditemukan sangat mendekati nol. Interpretasi intercept ini perlu diperhatikan, dan dalam konteks harga saham, nilai prediksi mendekati nol mungkin tidak memiliki interpretasi yang bermakna.

2. Pembukaan (Feature1): -2.6707347736627015e-10

Koefisien untuk fitur pembukaan sangat kecil, mendekati nol. Ini menunjukkan bahwa perubahan kecil pada nilai pembukaan tidak memiliki dampak yang signifikan pada perubahan data pengangguran.

3. Tertinggi (Feature2): 0,0

Koefisien untuk fitur tertinggi juga sangat kecil, mendekati nol. Perubahan kecil pada nilai tertinggi tidak memberikan kontribusi yang signifikan terhadap perubahan data pengangguran.

4. Terendah (Feature3): -9.85647481810166e-12

Sama seperti fitur tertinggi, koefisien fitur terendah sangat kecil. Ini menunjukkan bahwa terendah tidak memiliki dampak yang signifikan pada data pengangguran.

5. Penutupan (Feature4): 1.000000000078536

Koefisien untuk fitur penutupan sangat besar, menunjukkan bahwa perubahan satu unit pada penutupan akan menyebabkan perubahan satu unit pada data pengangguran. Hubungan positif yang kuat antara penutupan dan data pengangguran dapat diidentifikasi dari nilai koefisien yang signifikan ini.

Dengan demikian, hasil analisis koefisien ini memberikan wawasan mendalam tentang kontribusi relatif dari setiap variabel dalam menganalisa data pengangguran. Perlu diingat bahwa hasil ini penting untuk memahami faktor-faktor yang paling berpengaruh dalam model dan memberikan dasar untuk pemahaman lebih lanjut tentang hubungan antara fitur dan target.

d. Validasi Model

Pada tahap ini, kami melakukan validasi model menggunakan teknik cross-validation untuk mengevaluasi sejauh mana model dapat menggeneralisasi pola-pola dalam data yang tidak digunakan dalam pengembangan model. Berikut adalah hasil dari proses validasi model:

1. Root Mean Squared Error (RMSE): 1.3125284090603527e-08

   Nilai RMSE yang sangat kecil menunjukkan bahwa model regresi memiliki tingkat kesalahan yang sangat rendah saat memprediksi nilai pada dataset uji. Prediksi model sangat dekat atau bahkan mendekati nilai sebenarnya.

2. R-squared (Koefisien Determinasi): Nilai R-squared sebesar 1.0.

   Artinya, model mampu menjelaskan seluruh variabilitas data pada dataset uji. Hasil ini menunjukkan bahwa model sangat sesuai dengan data uji dan memberikan prediksi yang sangat akurat.

Hasil evaluasi model secara keseluruhan menunjukkan kinerja yang sangat baik. Model memiliki kemampuan untuk memberikan prediksi yang akurat dan dapat diandalkan terhadap data pengangguran. R-squared yang tinggi juga mengindikasikan bahwa model dapat menjelaskan variasi sebagian besar data, mengukuhkan kecocokan model dengan pola-pola dalam dataset.

Dengan hasil validasi yang positif ini, kita dapat memiliki keyakinan yang lebih tinggi dalam penggunaan model regresi untuk memprediksi data pengangguran pada data yang belum pernah dilihat sebelumnya.

Kesimpulan

Dalam melakukan penelitian ini, kami telah berhasil mengembangkan dan menganalisis model regresi data pengangguran berdasarkan data historis yang tersedia. Berikut adalah beberapa kesimpulan utama dari penelitian ini:

1. Kinerja Model:

   Model regresi yang dikembangkan menunjukkan kinerja yang sangat baik, dengan nilai Root Mean Squared Error (RMSE) yang sangat rendah, mendekati nol. Hal ini menandakan bahwa model memiliki tingkat kesalahan prediksi yang minimal.

2. Koefisien Determinasi (R-squared):

   Model memiliki nilai R-squared sebesar 1.0, menunjukkan kemampuan model dalam menjelaskan seluruh variasi data pada dataset uji. Prediksi model sangat akurat dan sesuai dengan nilai sebenarnya.

3. Analisis Residual:

   Analisis residual menunjukkan bahwa frekuensi kesalahan prediksi pada nilai 0.0 sangat tinggi, menandakan bahwa model cenderung memberikan prediksi yang sangat akurat, mendekati nilai sebenarnya.

4. Analisis Koefisien:

   Koefisien regresi untuk masing-masing variabel menunjukkan pengaruh yang signifikan terhadap persentasi data. Variabel yang memiliki koefisien besar memiliki kontribusi yang lebih besar dalam mempengaruhi data pengangguran.

5. Validasi Model:

   Proses validasi model menggunakan cross-validation menghasilkan nilai RMSE yang rendah, menegaskan kemampuan model untuk menggeneralisasi pola-pola dalam data baru.

Dengan demikian, dapat disimpulkan bahwa model regresi yang dikembangkan dapat diandalkan untuk memprediksi harga saham Bank BTN. Hasil positif ini dapat menjadi dasar untuk pengambilan keputusan di bidang keuangan dan investasi, memberikan informasi yang berharga bagi para pemangku kepentingan.

Daftar Pustaka

Rudolf Mathar, Fundamentals of Big Data Analytics, 2019

Aldana. W.A., Data Mining Industry: Emerging Trends and New Opportunities, Thesis, Master of Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, 2000

Kristyanto, B dan Dewa, PK., Kontribusi Ergonomi Untuk Rancangan Perakitan, Jurnal Teknologi Industri,Vol III, No 1,ISSN 1410-5004, 1999

Tan, Steinbach dan Kumar, 2006, Introduction to Data Mining, Pearson Education Inc., Han dan Kamber, 2001, Data Mining Concepts and Techniques, Academic Press

Ayu, S., Sari, E., Pembangunan, E., Ekonomi, F., Brawijaya, U., Ekonomi, F., Brawijaya, U., Ayu, S., Sari, E., & Pangestuty, F. W. (2022). Jdess 01.04.2022. 1(4), 641–649.

Chu, A. C., Kou, Z., & Wang, X. (2020). Dynamic effects of minimum wage on growth and innovation in a Schumpeterian economy. Economics Letters, 188, 108943. https://doi. org/10.1016/j.econlet.2020.10894

Demissie, M. M., Herut, A. H., Yimer, B. M., Bareke, M. L., Agezew, B. H., Dedho, N. H., & Lebeta, M. F. (2021). Graduates' Unemployment and Associated Factors in Ethiopia: Analysis of Higher Education Graduates' Perspectives. Education Research International, 2021. https://doi. org/10.1155/2021/46- 38264

Ehrenberg, R. G. (2012). Modern labor economics : theory and public policy - Eleventh edition (B. Donna (ed.); eleventh). PEARSON

Fatmawati, I., Suman, A., & Syafitri, W. (2018). The impact of fdi, human capital, and corruption on growth in asian developed and developing countries. International Journal of Scientific and Technology Research, 7(12), 216–221.

LINK GITHUB

https://github.com/pieroputrapersada/UAS-BIGDATA